Comparative Analysis of Stereo Matching Pipelines: From Classical Optimization to Deep Learning

Wenye Xiong, Xinyue Wang, Meihan Zheng
ES143 Computer Vision, Harvard University

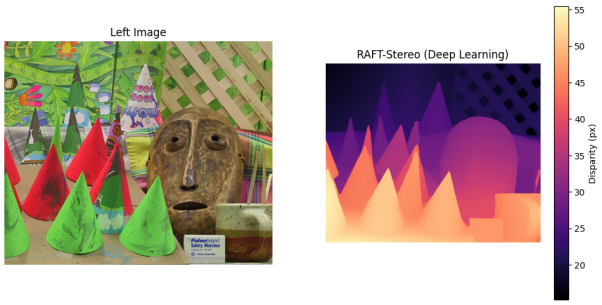
Abstract

Stereo matching remains a fundamental problem in computer vision, serving as the computational analog to human binocular depth perception. This project implements and evaluates a complete stereo vision pipeline, ranging from hardware calibration to dense disparity estimation. We present a comparative analysis of three distinct approaches: (1) local methods using Winner-Take-All (WTA) strategies with Sum of Absolute Differences (SAD), (2) global optimization via Dynamic Programming, and (3) state-of-the-art deep learning architectures (RAFT-Stereo) contrasted with classical semi-global block matching (SGBM). We utilize a custom-calibrated dual-smartphone rig to capture real-world datasets. Our results demonstrate the trade-offs between computational efficiency and accuracy, highlighting the superior robustness of deep learning methods in textureless regions and the efficacy of global optimization in handling occlusion.

1. Introduction

Depth estimation is critical for applications ranging from autonomous driving to augmented reality. Stereo matching recovers depth by identifying corresponding pixels in a rectified stereo pair, exploiting the inverse relationship between depth Z and disparity d: $Z = \frac{f \cdot B}{d}$, where f is the focal length and B is the baseline.

In this work, we explore the evolution of stereo algorithms[3]. We begin by establishing a rigorous geometric foundation through camera calibration and epipolar rectification. We then implement custom dense matching algorithms from scratch, including local block matching and scanline-based dynamic programming. Finally, we benchmark these against industry-standard baselines: OpenCV's Semi-Global Block Matching (SGBM) and the RAFT-Stereo neural network[1, 4].



(a) Example of Stereo Matching

Figure 1. An example of stereo matching output. The input left view is processed to generate the disparity map, where brighter pixels correspond to objects closer to the camera (larger disparity).

2. Methods

2.1. Data Collection

Two different data acquisition methods were used in the project

2.1.1. Dual-camera Capture

Two smartphone cameras were mounted on a tripod with a fixed relative angle so that both recorded the same scene from different viewpoints. For this setup, two categories of data were collected. First, 10 calibration images containing an AprilTag board placed at various distances and orientations were captured simultaneously by both cameras to enable intrinsic and stereo calibration. Second, 3 stereo scene pairs of the target environment were recorded for later reconstruction evaluation.

2.1.2. Mirror-based One Camera Capture

In this configuration, a single smartphone camera was used together with a planar mirror positioned adjacent to the camera lens. The mirror was fixed such that the direct view and the reflected view each occupied approximately half of the camera image, effectively forming two virtual cameras. Using this setup, 15 calibration images containing the AprilTag board under diverse poses and 3 scene images were collected.

2.2. Data Preprocessing

2.2.1. Dual-camera Capture

Because the two smartphones used in this setup have different native resolutions, the captured images must be standardized before calibration. All images were downsampled to a uniform resolution of 3024×4032 (width, height). In addition, since the raw data were stored in the HEIC format, all images were converted to PNG to ensure compatibility with the calibration and image processing pipeline.

2.2.2. Mirror-based One Camera Capture

The mirror configuration produces a single image of size 4032×3024 containing both the direct view and the reflected view. Due to the optical geometry of the mirror, the central region of the image (approximately 50%–55% of the width) exhibits overlapping content and blur, and therefore cannot be used. This region was discarded.

The remaining left 50% of the image was extracted as the direct-view image corresponding to the virtual left camera. The right 45% of the image was extracted as the mirror-view image. Because mirror imaging reverses orientation, the extracted right portion was horizontally flipped before being used as the virtual right camera image.

2.3. Single-Camera Calibration

2.3.1. Calibration

Each camera view—whether obtained from a physical camera in the dual-camera setup or from the direct/mirror view in the mirror-based configuration—was calibrated independently. All calibration images were first converted to grayscale, and AprilTag detection was performed to extract 2D corner coordinates associated with the known 3D coordinates of the AprilTag board.

These 2D–3D correspondences were then passed to OpenCV's `calibrateCamera` function to estimate the intrinsic matrix and lens distortion parameters. To ensure numerical stability, the principal point was fixed at the center of the image, tangential distortion was disabled, and higher-order radial distortion coefficients were held at zero.

For the mirror-based configuration, although the final calibration is applied to cropped image halves (the direct-view left half and the mirror-view right half), the calibration procedure uses the original full-resolution image size of 4032×3024 when invoking `calibrateCamera`. This ensures that the principal point, pixel coordinates, and resulting intrinsic parameters correspond to the true geometry of the single physical camera before the views are separated.

2.3.2. Evaluation

Calibration accuracy was evaluated using per-image reprojection error. For each calibration frame, the estimated intrinsics and extrinsics were used to reproject the 3D AprilTag corners into the image plane. The root-mean-square error (RMSE) between the detected 2D corner positions and the reprojected points was computed as the reprojection error.

2.4. Stereo Calibration

2.4.1. Calibration

After obtaining the intrinsic parameters for each camera view, stereo calibration was performed to estimate the relative pose between the two views. This procedure is applied uniformly to both the dual-camera setup and the mirror-based configuration, as both produce two synchronized perspectives of the same scene.

For each pair of calibration images, AprilTag IDs were detected and compared across the two views. Only image pairs containing a sufficient number of common tags were retained. For each retained pair, corresponding 2D observations were established by matching identical tag IDs and sorting them in a consistent order. The matched 2D points, together with the AprilTag board's 3D coordinates, formed the input to OpenCV's `stereoCalibrate` function.

During stereo calibration, the intrinsic parameters of both views were held fixed, and only the extrinsic parameters were optimized. The optimization returned the relative rotation matrix $\mathbf{R}$, translation vector $\mathbf{T}$, essential matrix $\mathbf{E}$, and fundamental matrix $\mathbf{F}$. These parameters describe the geometric relationship between the two views and are required for generating rectified stereo pairs and computing disparity.

2.4.2. Evaluation

To evaluate the quality of the estimated stereo geometry, epipolar consistency was assessed using the recovered fundamental matrix $\mathbf{F}$. For each matched feature pair, the corresponding epipolar line in the opposite view was computed, and the perpendicular distance from the detected point to its epipolar line was measured. This point-to-line distance represents the epipolar error.

2.5. Stereo Rectification

2.5.1. Rectification

Following stereo calibration, stereo rectification was performed to transform the two camera views into a geometry where corresponding points lie on the same horizontal scanline. This step is essential for disparity estimation and requires only the intrinsic parameters and the extrinsic pose $(\mathbf{R}, \mathbf{T})$ obtained from stereo calibration.

Using OpenCV's `stereoRectify` function, rectification transforms were computed for both views. The procedure returns the rectification rotation matrices $\mathbf{R}_1$ and $\mathbf{R}_2$, the corresponding projection matrices $\mathbf{P}_1$ and $\mathbf{P}_2$, and the disparity-to-depth mapping matrix $\mathbf{Q}$. These matrices define the new, rectified coordinate systems and ensure that the two views become epipolar-aligned.

Pixel-level remapping functions for each view were then generated using `initUndistortRectifyMap`. These remapping functions incorporate the intrinsics, distortion coefficients, and rectification transforms, and allow each rectified image to be produced via a single call to `remap`. The resulting rectified image pairs exhibit horizontally aligned structures and consistent epipolar geometry.

2.5.2. Evaluation

Rectification quality was assessed visually. Rectified left and right images were stacked horizontally with evenly spaced horizontal reference lines overlaid. Proper rectification results in these reference lines intersecting corresponding scene features at identical heights in both images.

2.6. Disparity Range Estimation

After stereo rectification, the projection matrices take the canonical form

$$P_1 = \begin{bmatrix} f & 0 & c_x & 0 \\ 0 & f & c_y & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \qquad P_2 = \begin{bmatrix} f & 0 & c_x & -fB \\ 0 & f & c_y & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}. \tag{1}$$

From P_1, the focal length is obtained directly as

$$f = P_1(1,1), \tag{2}$$

The baseline B appears in the translation term of P_2. Because the rectified form satisfies

$$P_2(1,4) = -fB, \tag{3}$$

the baseline is recovered by

$$B = -\frac{P_2(1,4)}{P_2(1,1)}. \tag{4}$$

Given f and B, the disparity corresponding to depth Z follows

$$d(Z) = \frac{fB}{Z}. \tag{5}$$

2.7. Problem Formulation

Given a rectified stereo pair (I_L, I_R) of size $H \times W$, the goal is to estimate a dense disparity map

$$D : \Omega \to \mathbb{R},$$

where Ω denotes pixel locations in the left image. Since the images are rectified, a pixel (x, y) in the left image matches a pixel $(x - d, y)$ in the right image for some $d \in [d_{\min}, d_{\max}]$. [3]

For each candidate disparity d, we define a matching cost $C(p, d)$, and the disparity is obtained by a WTA rule:

$$D(p) = \arg\min_d C(p, d).$$

The three methods we evaluate differ in how $C(p, d)$ is constructed and aggregated.

2.8. Local Dense Matching Methods

Local stereo methods compute $C(p, d)$ independently at each pixel, followed by a WTA selection. We implement two variants.

2.8.1. WTA with SAD

For each pixel p we use a square window W_p of radius r. The cost for disparity d is

$$C_{\mathrm{SAD}}(p, d) = \sum_{(u,v)\in W_p} |I_L(u, v) - I_R(u - d, v)|.$$

We compute this efficiently via a box filter over the absolute difference image. Border pixels where the window is incomplete are marked invalid.

2.8.2. WTA with Census Transform and Guided Filtering

To improve robustness to illumination differences, we use the Census transform:

$$\mathrm{Census}(I, p) = (\mathbf{1}(I(q_1) < I(p)), \dots),$$

encoded as a 64-bit signature. The matching cost is the Hamming distance:

$$C_{\mathrm{Census}}(p, d) = \mathrm{Hamming}(I_L^{\mathrm{cen}}(p), I_R^{\mathrm{cen}}(p - (d, 0))).$$

Because raw Census costs are noisy, we apply guided filtering with I_L as guidance, producing aggregated cost slices $\tilde{C}(p, d)$:

$$\tilde{C}(p, d) = \bar{a}(p) I_L(p) + \bar{b}(p).$$

The final disparity is obtained by WTA over $\tilde{C}(p, d)$.

2.9. Scanline Dynamic Programming for Global Matching

Local WTA methods estimate disparities independently and therefore lack smoothness along epipolar lines. To introduce global regularization, we adopt a scanline dynamic programming (DP) formulation.

For each scanline y, we optimize a 1D disparity sequence $\mathbf{d} = (d_1, \dots, d_W)$ by minimizing

$$E(\mathbf{d}) = \sum_{x=1}^{W} C(x, y, d_x) + \sum_{x=2}^{W} V(d_x, d_{x-1}),$$

where $C(x, y, d)$ is the SAD-based data term and the smoothness cost is

$$V(d_x, d_{x-1}) = \begin{cases} 0, & d_x = d_{x-1}, \\ P_1, & |d_x - d_{x-1}| = 1, \\ P_2, & \text{otherwise.} \end{cases}$$

Let $L_x^{\rightarrow}(d)$ denote the minimum cumulative energy of reaching pixel x with disparity d. It satisfies the recurrence

$$L_x^{\rightarrow}(d) = C(x,y,d) + \min\Big(L_{x-1}^{\rightarrow}(d),\ L_{x-1}^{\rightarrow}(d \pm 1) + P_1,\ \min_{d' \neq d} L_{x-1}^{\rightarrow}(d') + P_2\Big).$$

A symmetric backward pass gives $L_x^{\leftarrow}(d)$, and the final aggregated cost is

$$L(x,y,d) = L_x^{\rightarrow}(d) + L_x^{\leftarrow}(d).$$

The DP estimate is obtained by WTA over the aggregated cost:

$$D_{\mathrm{DP}}(x,y) = \arg\min_d L(x,y,d).$$

This regularized inference improves consistency in textureless or ambiguous regions.

2.10. Semi-Global Block Matching (SGBM)

While Scanline DP introduces smoothness along epipolar lines, it fails to enforce vertical consistency, often resulting in horizontal "streaking" artifacts. To address this, we evaluate the Semi-Global Block Matching (SGBM) algorithm [2], which approximates a global 2D energy minimization by aggregating costs along multiple 1D paths.

For a pixel p and disparity d, SGBM computes an aggregated cost $S(p,d)$ by summing the path costs $L_{\mathbf{r}}$ from multiple directions $\mathbf{r}$ (typically 8 or 16 directions):

$$S(p,d) = \sum_{\mathbf{r}} L_{\mathbf{r}}(p,d).$$

The path cost $L_{\mathbf{r}}(p,d)$ is calculated recursively. Unlike the standard DP, which strictly enforces the Viterbi path, SGBM adds a penalty term while subtracting the minimum path cost of the previous pixel to prevent numerical overflow:

$$\begin{aligned} L_{\mathbf{r}}(p,d) = C(p,d) + \min\Big(&L_{\mathbf{r}}(p-\mathbf{r},d),\\ &L_{\mathbf{r}}(p-\mathbf{r},d\pm1)+P_1,\\ &\min_k L_{\mathbf{r}}(p-\mathbf{r},k)+P_2\Big)\\ &-\min_k L_{\mathbf{r}}(p-\mathbf{r},k), \end{aligned}$$

where P_1 penalizes small disparity changes (slanted surfaces) and P_2 penalizes large disparity jumps (depth discontinuities), with $P_2 > P_1$. The final disparity is determined by a WTA strategy on the summed cost $S(p,d)$, followed by sub-pixel interpolation and left-right consistency checks.

2.11. Deep Learning Approach: RAFT-Stereo

Traditional methods like SGBM rely on hand-crafted matching costs (SAD/Census) and smoothness priors (P_1, P_2), which often fail in textureless regions or non-Lambertian surfaces. To overcome these limitations, we implement RAFT-Stereo [4], a state-of-the-art deep learning architecture based on Recurrent All-Pairs Field Transforms.

RAFT-Stereo operates on a rectified stereo pair (I_L, I_R) and estimates disparity through an iterative refinement process resembling an optimization algorithm. The architecture consists of three main components:

1. **Feature Extraction:** A shared convolutional neural network extracts pixel-wise features $f_L, f_R \in \mathbb{R}^{H/4 \times W/4 \times D}$ from the input images. Context features are also extracted from I_L to guide the refinement.
2. **Correlation Pyramid:** Instead of a single cost volume, RAFT-Stereo constructs a multi-scale correlation pyramid by computing the dot product between feature vectors:
$$\mathcal{C}(u,v,d) = f_L(u,v) \cdot f_R(u-d,v).$$
This captures similarity at different spatial resolutions, handling both large displacements and fine details.
3. **Recurrent Update Operator:** A Gated Recurrent Unit (GRU) iteratively updates the disparity field. At each iteration k, the GRU takes the current disparity estimate $\mathbf{d}_k$, the retrieved correlation features, and the context features to predict a residual update $\Delta\mathbf{d}$:
$$\mathbf{d}_{k+1} = \mathbf{d}_k + \Delta\mathbf{d}.$$

The network is trained in a supervised manner using the L_1 distance between the predicted and ground-truth disparity over a sequence of iterations. This learned approach implicitly encodes global context and smoothness priors, allowing it to "hallucinate" valid disparities even in occluded or textureless regions where SGBM would output invalid values.

3. Results

3.1. Single-Camera Calibration Results

The accuracy of the intrinsic calibration was evaluated by comparing the detected AprilTag corner locations with the corresponding reprojected points. Figures 2–3 show representative calibration images for both the mirror-based single-camera setup and the dual-camera setup, where detected points are shown in blue and reprojected points in red.

For the mirror-based configuration, the mean reprojection errors for the two virtual views were

$$\mathrm{RMSE}_1 = 1.617 \text{ px}, \qquad \mathrm{RMSE}_2 = 2.597 \text{ px}.$$

The second view exhibits larger error due to the reflection and the cropping–flipping operations applied during preprocessing, but the error remains well within acceptable limits for stereo processing.

For the dual-camera configuration, the two physical cameras achieve lower reprojection errors:

$$\mathrm{RMSE}_1 = 0.991 \text{ px}, \qquad \mathrm{RMSE}_2 = 1.176 \text{ px}.$$

These lower values reflect the absence of mirror distortion and the greater imaging stability of two independent physical sensors.

Overall, all reprojection errors remain below 3 px, indicating that the estimated intrinsic parameters are geometrically consistent and suitable for subsequent stereo calibration and rectification.

3.2. Stereo Calibration Results

Stereo calibration was evaluated for both the mirror-based configuration and the dual-camera setup. The mirror-based system yielded a mean stereo reprojection error of

$$\mathrm{RMSE}_{\text{mirror}} = 6.5847 \text{ px},$$

whereas the dual-camera system achieved a lower error of

$$\mathrm{RMSE}_{\text{dual}} = 1.47 \text{ px}.$$

The higher error from the mirror-based configuration is expected due to the additional reflection and cropping transformations applied during preprocessing.

To further validate the accuracy of the estimated extrinsics, we visualized the epipolar geometry for both setups. As shown in Fig. 4, nearly all epipolar lines pass through their corresponding detected object points in both configurations. This shows that, despite the numerical differences in reprojection error, the fundamental matrices of both systems accurately describe the stereo geometry.

We also compared the recovered camera poses. Figure 5 illustrates the estimated camera orientations and positions for the mirror-based and dual-camera setups. The dual-camera calibration yields two physical camera poses that align well with the real-world arrangement of the devices during data capture. In contrast, the mirror-based calibration produces two symmetric virtual-camera poses, consistent with the reflective geometry imposed by the planar mirror.

3.3. Rectification Results

After stereo calibration, rectification was performed for both the mirror-based and dual-camera setups. The rectified image pairs for each configuration are shown in Fig. 6. In both cases, the corresponding points of interest in the left and right views lie along nearly the same horizontal scanlines. This alignment indicates that the rectification transformation successfully enforced the epipolar constraint and brought the two images to a geometry suitable for the estimation of the disparities.

3.4. Disparity Estimation Results

Using the rectified projection matrices, the focal length f and baseline B were extracted to compute the feasible disparity range for each configuration. For the mirror-based setup, the resulting disparity interval was

$$d_{\text{mirror}} \in [4,\ 32],$$

which reflects the small effective baseline induced by the planar mirror geometry. In contrast, the dual-camera configuration produced a substantially larger disparity interval,

$$d_{\text{two}} \in [146,\ 592],$$

due to the significantly wider physical baseline between the two real cameras.

3.5. Qualitative Comparison

Figure 7 shows qualitative disparity estimation results for three different scenes using our three stereo matching algorithms: (1) local WTA with SAD, (2) local WTA with Census and guided filtering, and (3) scanline dynamic programming (DP).

As shown in Fig. 7, the three evaluated stereo algorithms produce noticeably different disparity characteristics across all scenes. Across all scenes, the SAD baseline is able to recover the coarse structure of the scene, but the results are heavily contaminated by noise and block artifacts. The Census-based method reduces much of this noise, yet its output often appears more blocky, and object boundaries become less distinct. The DP method recovers the overall scene geometry more clearly and further suppresses noise, but its 1D optimization introduces horizontal smoothing, which leads to noticeable blurring along epipolar lines.

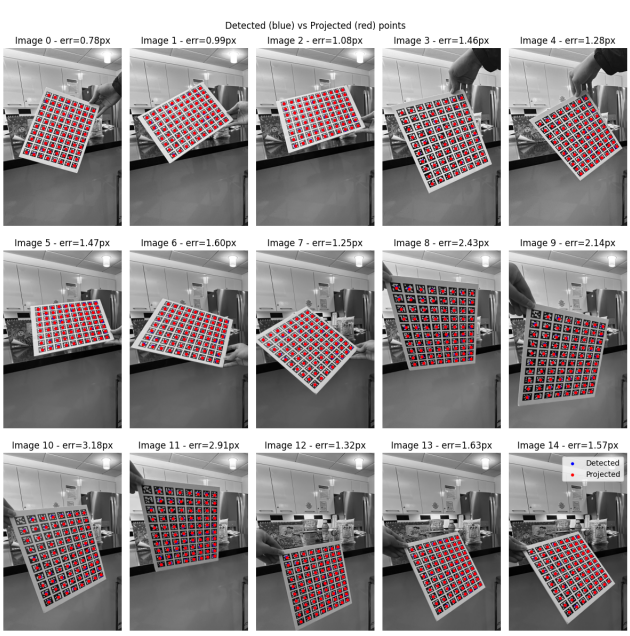
Overall, the comparison illustrates that each method has characteristic trade-offs: SAD preserves coarse shapes but is highly noisy, Census reduces noise but loses contour sharpness, and DP provides smoother surfaces at the cost of horizontal blur.

3.6. Third-Party Algorithm Comparison

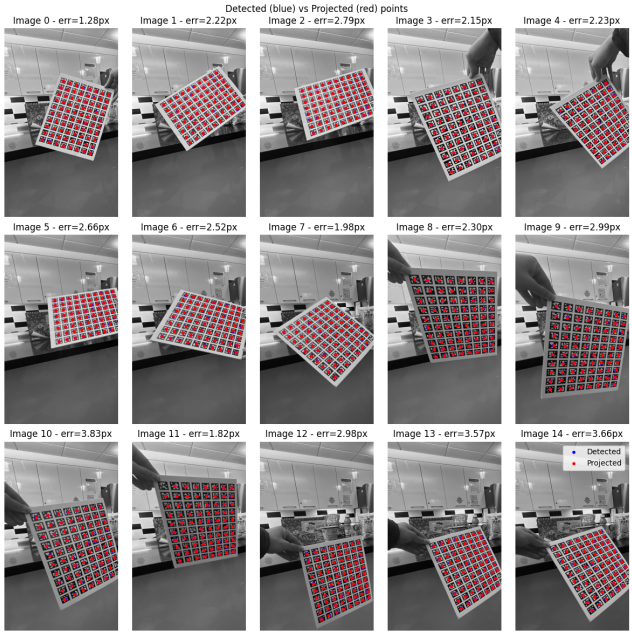
Following the evaluation of our custom implementations, we benchmarked two third-party solutions representing the classical and deep learning paradigms: OpenCV's Semi-Global Block Matching (SGBM) and RAFT-Stereo. Figure 9 presents the results across the same three scenes.

Semi-Global Block Matching (SGBM): As a classical baseline, SGBM aggregates matching costs along multiple 1D paths to approximate a 2D global optimization. The results exhibit a significant improvement in structural coherence compared to the local SAD and Census methods. However, two limitations remain prominent:

1. **Invalid Disparity Band and Geometric Implications:** A wide vertical band of invalid pixels (black) is visible

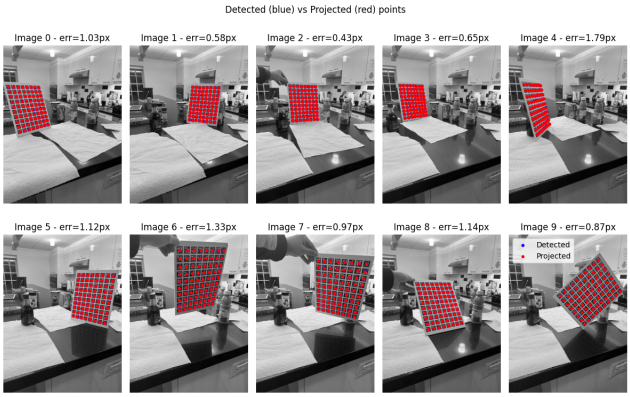


(a) Virtual view 1 (direct view)

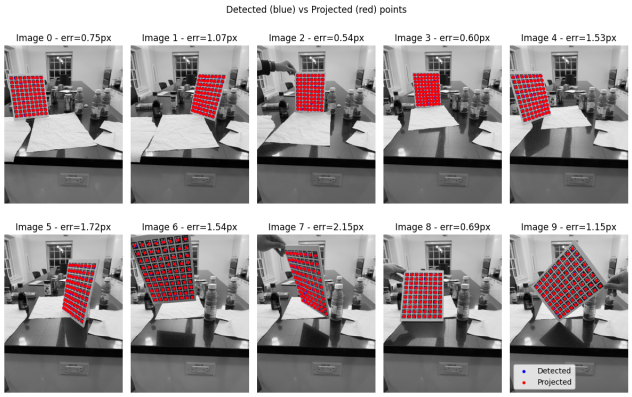


(b) Virtual view 2 (mirror view)

Figure 2. Mirror-based single-camera calibration. Detected AprilTag corners (blue) and reprojected points (red).



(a) Camera 1



(b) Camera 2

Figure 3. Dual-camera calibration results. Detected AprilTag corners (blue) and reprojected points (red).

on the left margin. While typically attributed to the limited Field of View (FOV) overlap in large-baseline setups, we hypothesize this artifact is exacerbated by our specific capture geometry. Our setup employed a significant *convergence angle* (toed-in configuration), causing the intersection of the optical axes to occur *in front* of the scene geometry. This places the objects of interest behind the zero-disparity plane (horopter), effectively shifting the disparity distribution into the negative range ($d < 0$). To align the epipolar lines for such a setup, the rectification process introduces substantial principal point shifts and perspective warping, thereby expanding the invalid region on the image periphery where correspondences are geometrically undefined.

2. **Textureless Ambiguity:** In weak-texture regions (e.g., the white cabinet surfaces in Scene 2), SGBM still struggles to resolve unique matches, resulting in scattered "speckle" noise or voids where the uniqueness ratio check fails.

RAFT-Stereo: The deep learning approach, RAFT-Stereo, yields the highest visual quality among all evaluated methods. The resulting disparity maps are fully dense and exhibit sub-pixel smoothness, effectively eliminating the discretization "staircasing" effects seen in SGBM. Crit-

ically, the network leverages learned global context to infer smooth surfaces even in the textureless and reflective regions where classical block matching fails. While computationally heavier, RAFT-Stereo demonstrates superior robustness to the ill-posed conditions inherent in indoor environments.

4. Limitations

While our experiments successfully demonstrated the progression from basic local matching to advanced deep learning techniques, several limitations in both the hardware configuration and algorithmic performance were observed.

4.1. Geometric and Hardware Constraints

The most significant limitation in the single-camera mirror setup was the restricted baseline. As calculated in the disparity estimation results, the effective baseline for the virtual stereo pair was small, resulting in a narrow disparity range ($d \in [4, 32]$). Since depth resolution is directly proportional to the baseline length, this setup inherently suffered from poor quantization of depth, making it unsuitable for high-precision measurements at distance. Additionally, the mirror setup introduced higher reprojection errors (RMSE ≈ 6.58 px), likely due to imperfections in the mirror surface planarity and distortions introduced during the cropping and flipping preprocessing steps.

4.2. Rectification Artifacts in Toed-in Configurations

For the dual-camera setup, we employed a convergent (toed-in) physical arrangement to maximize the overlap of the field of view. However, as noted in the SGBM analysis, this geometry places the optical axes' intersection in front of the scene. The subsequent rectification process required to align the epipolar lines induced significant perspective distortion. This resulted in a large "invalid" region on the left margin of the disparity maps where no correspondence could be computed. This suggests that while toed-in configurations increase binocular overlap, they impose a penalty on the effective image width available for stereo matching after rectification.

4.3. Algorithmic Robustness in Textureless Regions

A recurring failure mode across all classical algorithms (SAD, Census, and DP) and semi-global methods (SGBM) was the inability to handle textureless regions, such as the white cabinet surfaces in our test scenes. These methods rely on local intensity variations to compute matching costs; in the absence of texture, the cost volume becomes ambiguous, leading to the "speckle" noise and voids observed in our results. While RAFT-Stereo successfully resolved these regions by leveraging learned global priors, the reliance on heavy GPU compute makes it less applicable for resource-constrained, real-time embedded systems compared to the efficient SGBM.

5. Conclusion

In this work, we presented a comprehensive evaluation of stereo vision pipelines, comparing a low-cost single-camera mirror setup against a standard dual-camera rig, and benchmarking algorithms ranging from local window-based methods to state-of-the-art deep learning models.

Our results demonstrate a clear hierarchy in hardware and software performance. On the hardware side, the dual-camera system significantly outperformed the mirror-based virtual stereo approach. Although the mirror setup offers the advantage of intrinsic hardware synchronization and lower cost, its utility is severely limited by the small effective baseline and higher calibration errors. The dual-camera setup, despite requiring more complex extrinsic calibration and synchronization, provided a stable geometry with high depth resolution ($d \in [146, 592]$).

Algorithmically, our comparisons highlight the trade-off between computational efficiency and reconstruction density. Local methods like SAD and Census proved insufficient for indoor environments due to high noise sensitivity. Scanline Dynamic Programming offered an improvement in surface smoothness but introduced characteristic streaking artifacts. The industry-standard SGBM provided the best balance of speed and structural accuracy among non-learning methods, though it remained vulnerable to occlusion and textureless surfaces. Finally, RAFT-Stereo demonstrated that deep learning approaches can effectively solve the ill-posed problems of stereo matching—smoothness, occlusion, and textureless regions—achieving fully dense, sub-pixel accurate disparity maps that far exceed classical baselines.

Future work could focus on optimizing the rectification process for convergent camera setups to minimize data loss, or exploring lightweight stereo networks that bridge the gap between the efficiency of SGBM and the accuracy of RAFT.

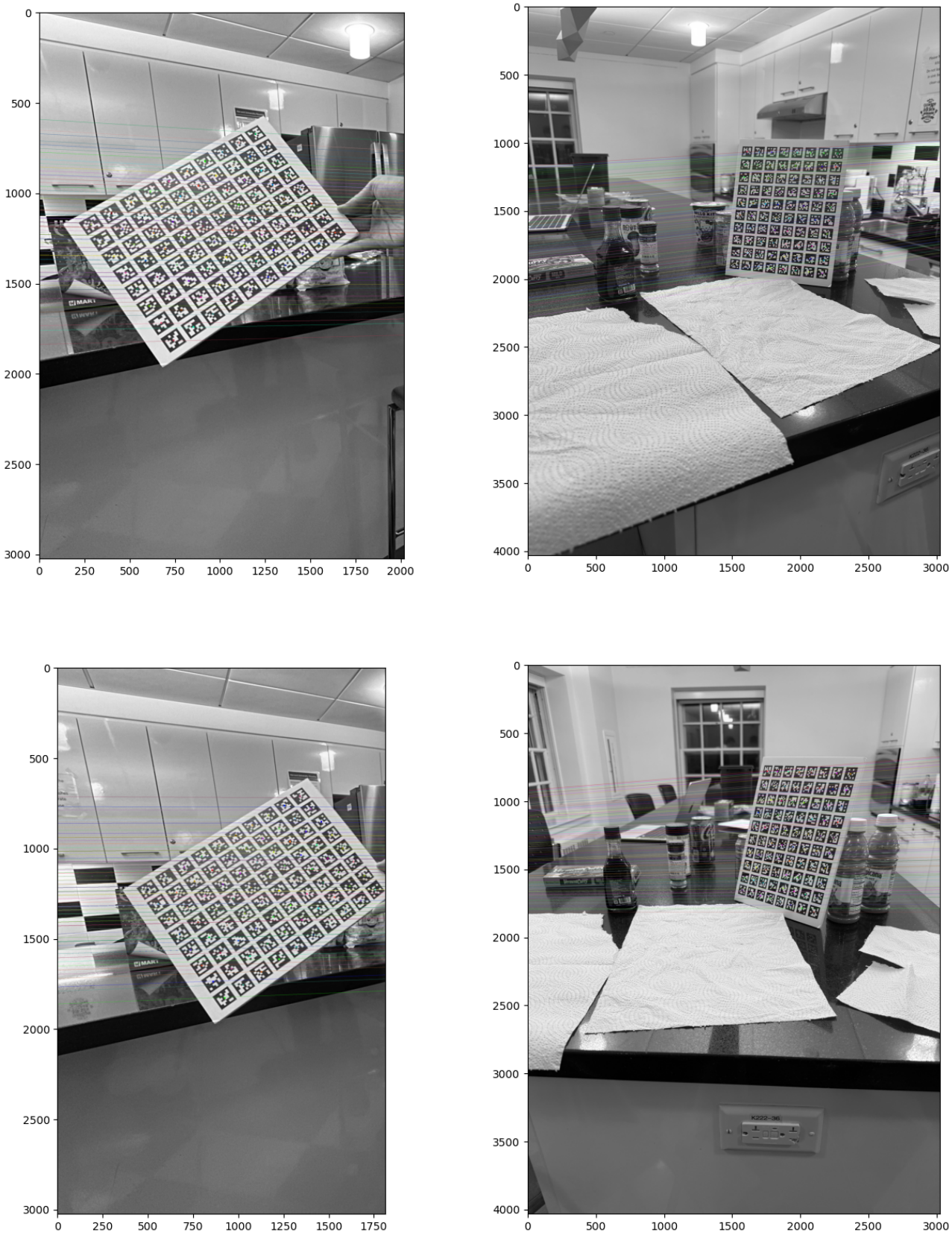
References

[1] Gary Bradski. The opencv library. *Dr. Dobb's Journal of Software Tools*, 2000. 1

[2] Heiko Hirschmuller. Stereo processing by semiglobal matching and mutual information. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(2):328–341, 2008. 4

[3] Richard Szeliski. *Computer Vision: Algorithms and Applications*. Springer Science & Business Media, 2010. 1, 3

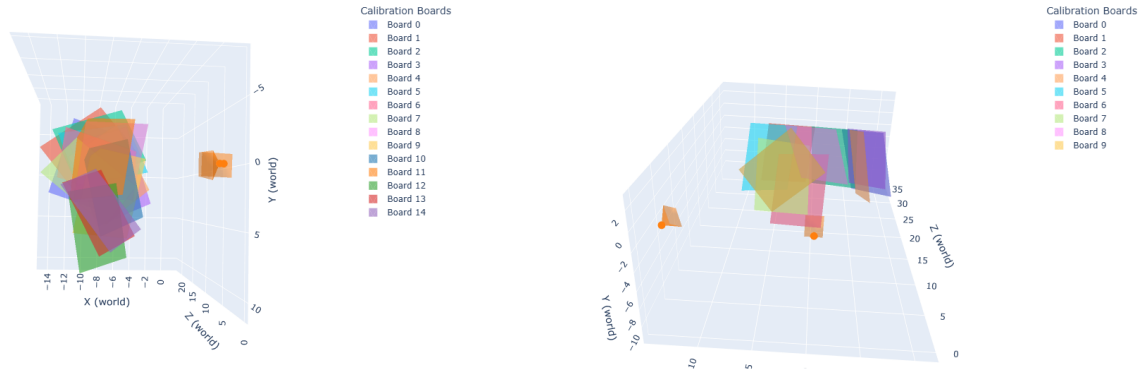
[4] Zachary Teed and Jia Deng. Raft-stereo: Multilevel recurrent field transforms for stereo matching. In *International Conference on 3D Vision (3DV)*. IEEE, 2021. 1, 4



(a) Mirror-based stereo calibration

(b) Dual-camera stereo calibration

Figure 4. Comparison of epipolar geometry between the mirror-based setup (left) and the dual-camera setup (right).



(a) Mirror-based virtual camera poses

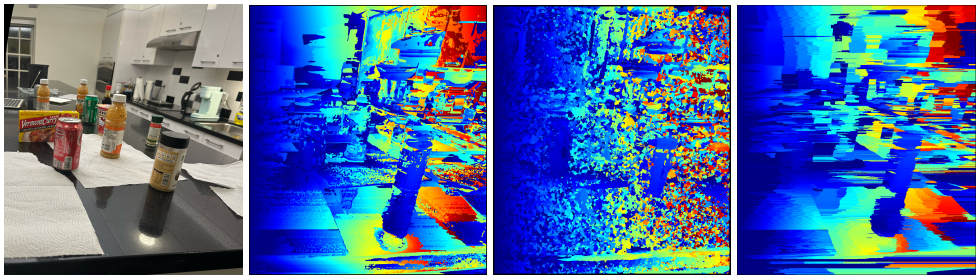
(b) Dual-camera physical camera poses

Figure 5. Estimated camera poses for the mirror-based configuration (left) and the dual-camera configuration (right).

(a) Mirror-based rectified stereo pair

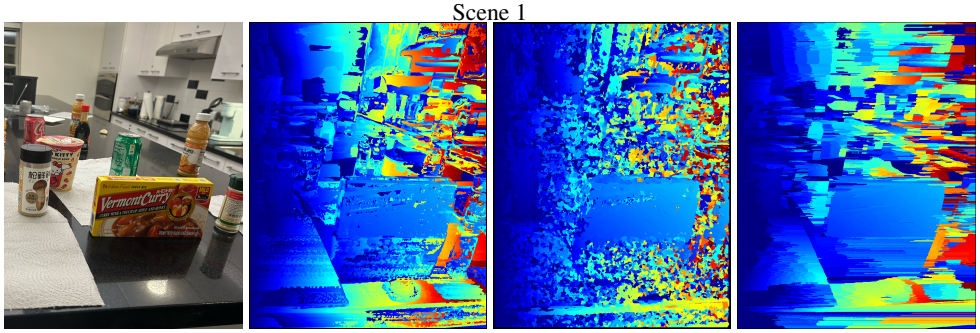
(b) Dual-camera rectified stereo pair

Figure 6. Comparison of rectified stereo pairs.



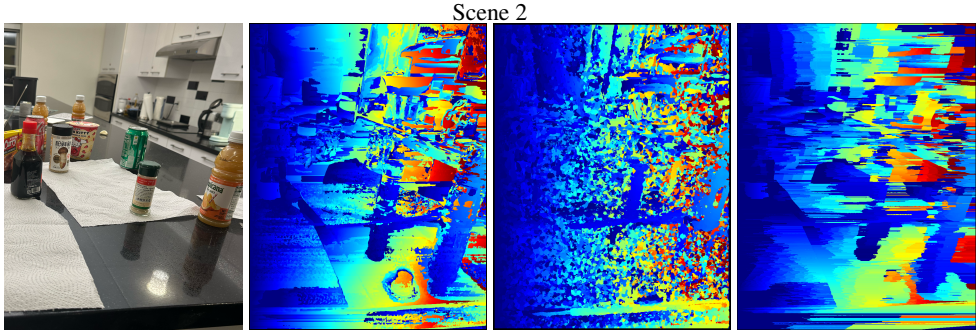
(a) GT (b) SAD (c) Census + Guided Filter (d) Scanline DP

Scene 1



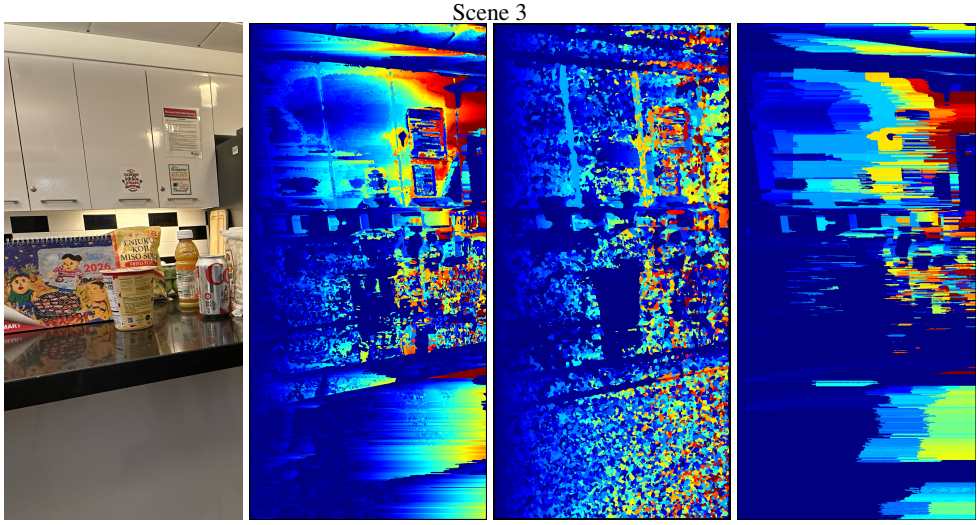
(e) GT (f) SAD (g) Census + Guided Filter (h) Scanline DP

Scene 2



(i) GT (j) SAD (k) Census + Guided Filter (l) Scanline DP

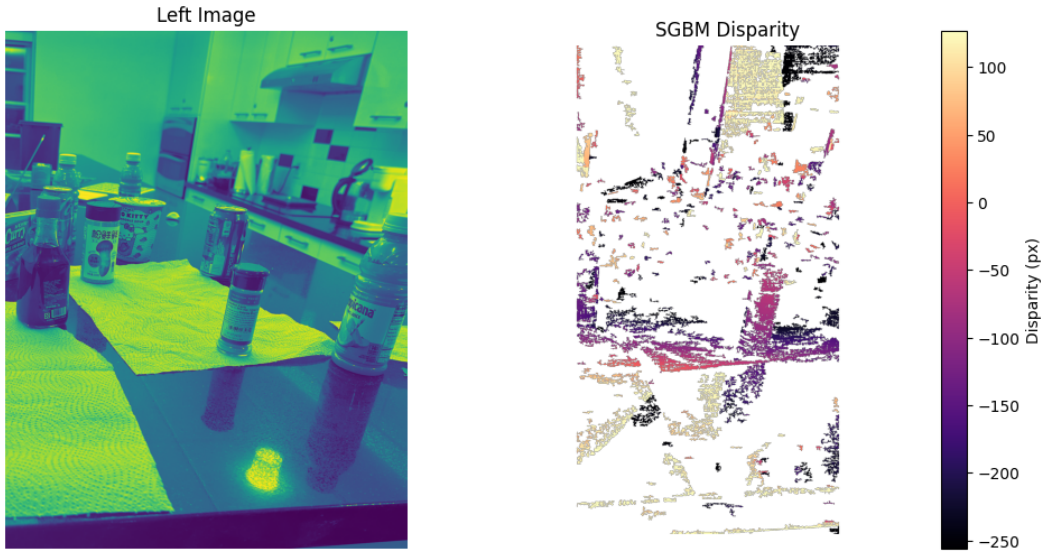
Scene 3



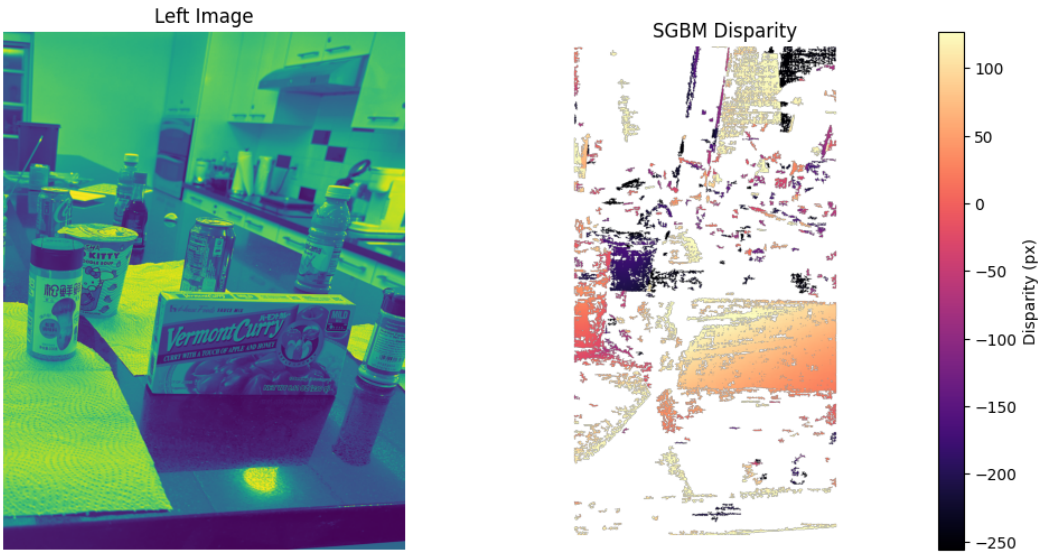
(m) GT (n) SAD (o) Census + Guided Filter (p) Scanline DP

Scene 4

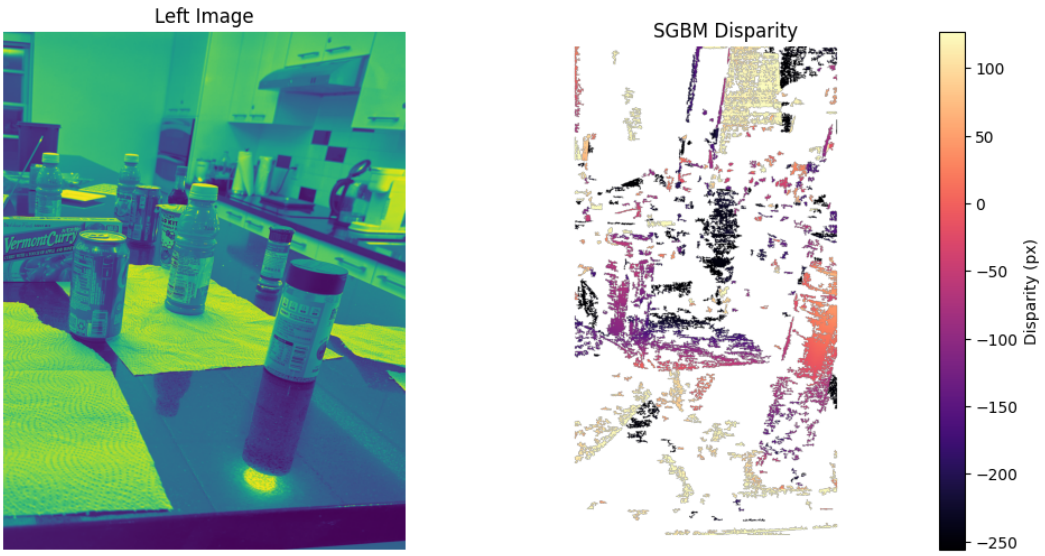
Figure 7. Qualitative comparison of three stereo matching algorithms (SAD, Census+Guided Filter, and Scanline DP).



(a) SGBM (Scene 1)

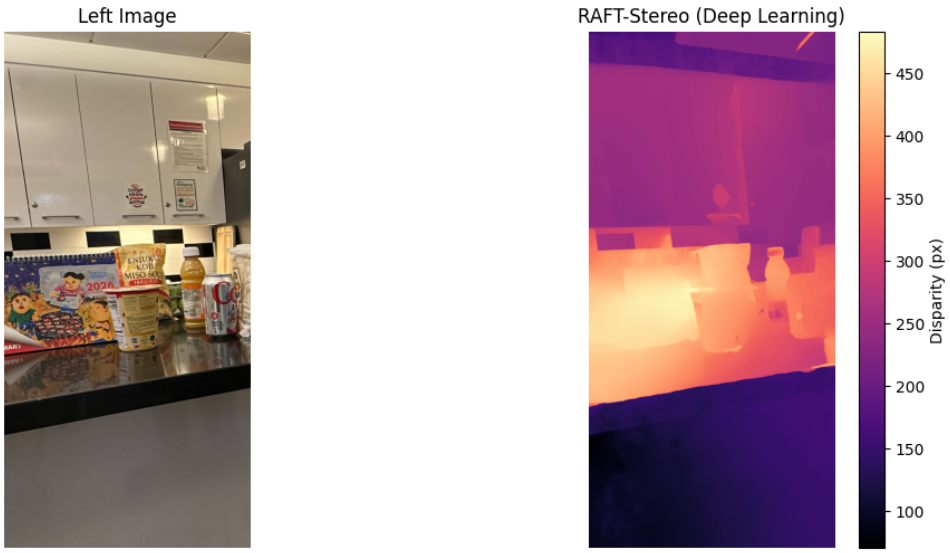


(b) SGBM (Scene 2)

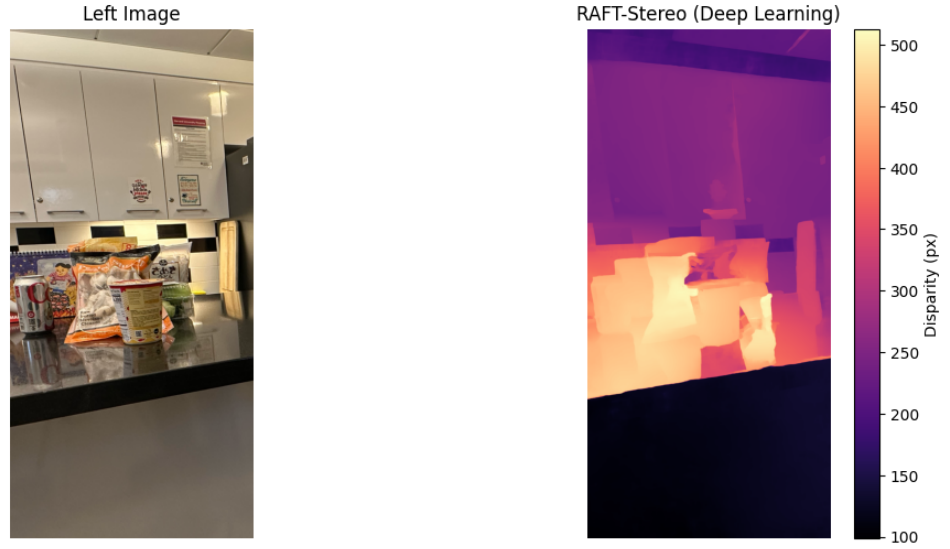


(c) SGBM (Scene 3)

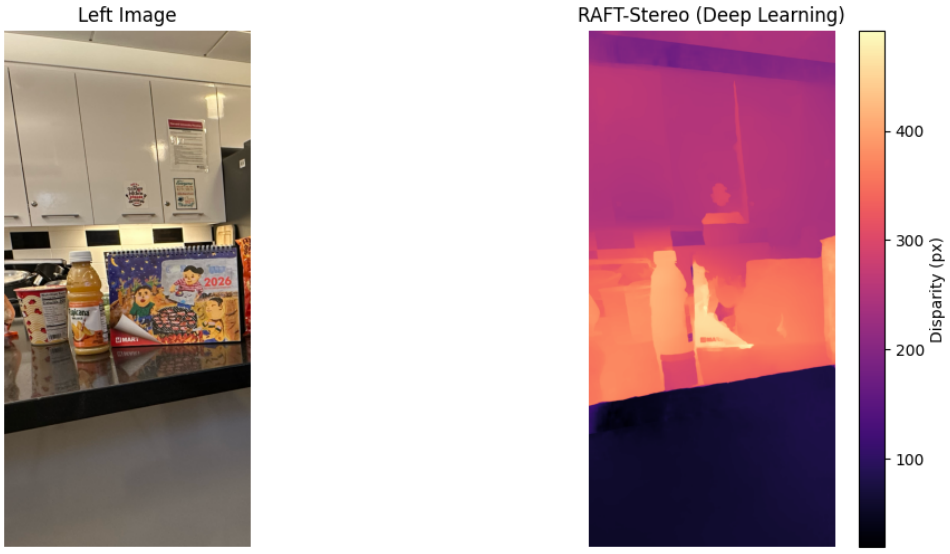
Figure 8. Qualitative results of SGBM



(a) RAFT (Scene 1)



(b) RAFT (Scene 2)



(c) RAFT (Scene 3)

Figure 9. Qualitative results of RAFT on mirror scenes.