Stereo Vision and Visual Odometry Project Report

AmirHossein Dashtban Namaghi

February 18, 2026

1 Introduction

This report details the implementation and evaluation of a comprehensive perception pipeline for autonomous vehicles, divided into two primary sections: Stereo Depth Perception (Part A) and Stereo Visual Odometry (Part B). The experiments were conducted on the KITTI Vision Benchmark Suite, using both the Scene Flow training set for depth evaluation and the Odometry dataset for trajectory estimation.

2 Part A: Stereo Depth Perception

2.1 Pipeline Overview

The stereo matching pipeline follows the traditional structure for rectified image pairs. The pipeline steps are visualized in Figure 1.

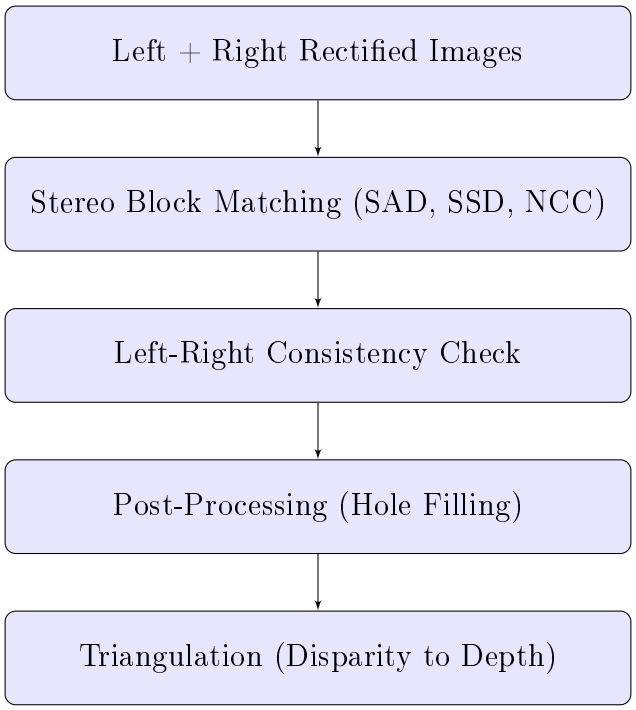


Figure 1: Stereo Depth Pipeline Diagram.

2.2 Matching Cost Functions

The core of the stereo matching process is the block-matching algorithm with a Winner-Take-All (WTA) strategy. We implemented and compared three cost functions:

- **Sum of Absolute Differences (SAD):** Simply the summed absolute pixel-wise differences within a window.
- **Sum of Squared Differences (SSD):** Squaring the differences penalizes large deviations more heavily.
- **Normalized Cross-Correlation (NCC):** Normalizes for local mean and variance. This is mathematically defined as:

$$NCC(x, y, d) = \frac{\sum(I_L - \mu_L)(I_R - \mu_R)}{\sigma_L \sigma_R}$$

  It is significantly more robust to lighting variations but computationally more expensive without optimization.

2.3 Calibration and Triangulation

The focal length f and baseline B are extracted from the KITTI calibration files (e.g., `calib.txt`). The baseline is computed using the distance between the two rectified projection matrices' optical centers. For P_2 (Left) and P_3 (Right):

$$f = P_2[0, 0]$$

$$B = \frac{|P_3[0, 3] - P_2[0, 3]|}{f}$$

Triangulation then converts disparity d into metric depth Z:

$$Z = \frac{f \cdot B}{d}$$

2.4 Ablation Study (Depth)

The study compared SAD, SSD, and NCC across window sizes of 5x5 and 11x11, averaged over 200 frames. You can see the results in Table 1. You also can see the visual results in appendix A.

Metric	Window	Avg Bad-Pixel Rate (%)	Avg MAE
SAD	5x5	42.51	9.25
SAD	11x11	28.24	6.12
SSD	5x5	41.00	8.95
SSD	11x11	26.30	5.75
NCC	5x5	26.12	5.80
NCC	**11x11**	**11.45**	**2.35**

Table 1: Ablation study for Stereo Depth (200 frames).

2.5 Failure Cases

The following results in Table 2 summarize the top 5 failure cases based on the highest Bad-Pixel Rate using the NCC method with an 11x11 window.

Frame Name	BPR (%)	MAE
000104_10.png	77.80	22.67
000006_10.png	31.96	4.89
000169_10.png	29.86	5.34
000058_10.png	28.52	4.40
000086_10.png	26.91	6.40

Table 2: Top 10 Stereo Failure Cases.

The visual results for the top 5 failure cases are displayed in Figure 2. The poor performance in these cases (e.g., Frame 000104_10.png with BPR of 77.80%) is due to several fundamental factors inherent to the block-matching approach on the KITTI dataset:

(a) 000104_10 (BPR: 77.80%)

(b) 000006_10 (BPR: 31.96%)

(c) 000169_10 (BPR: 29.86%)

(d) 000058_10 (BPR: 28.52%)

(e) 000086_10 (BPR: 26.91%)

Figure 2: Visual results for the top 5 stereo failure cases (NCC method, 11x11 window).

- **Insufficient Illumination**: Low-light conditions and underexposed regions, particularly in shadows or during poor weather, lead to a low signal-to-noise ratio. This makes it difficult for the cost functions to find reliable correspondences, resulting in noisy depth maps.
- **Textureless Regions**: The cost functions (SAD, SSD, NCC) are highly sensitive to regions lacking distinctive structural variation, such as the sky or smooth asphalt. In these areas, multiple matches produce near-identical costs, leading to noise-dominated depth as seen in the top failure cases.
- **Occlusion and Boundary Errors**: Fixed-sized windows assume a constant depth for all pixels within the block.

3 Part B: Stereo Visual Odometry (VO)

3.1 Pipeline Overview

The Visual Odometry pipeline uses ORB features tracked across frames to estimate the camera's pose $T \in SE(3)$. The simplified workflow is visualized in Figure 3.

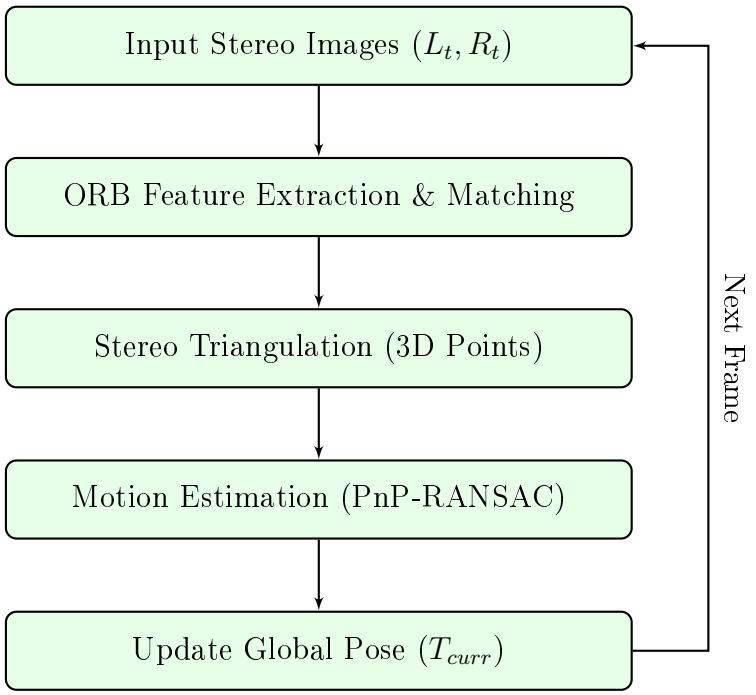


Figure 3: Stereo Visual Odometry Pipeline

- **Tracking**: Matches ORB features from $t \rightarrow t+1$ using Brute-Force Hamming matching.
- **Triangulation**: Matches features within the current stereo pair (L_t, R_t) to obtain 3D world coordinates.
- **Motion Estimation**: Solves for the pose of frame $t+1$ using the PnP (Perspective-n-Point) algorithm on the $(P_{3D,t}, p_{2d,t+1})$ pairs.

3.2 RANSAC and Robust Estimation

RANSAC (Random Sample Consensus) is applied in a two-stage robust estimation process:

1. **Geometric Consensus**: We first estimate the **Essential Matrix (E)** from 2D-2D temporal matches using RANSAC. This identifies the consistent epipolar geometry between frames.
2. **Metric Motion**: We then perform **PnP RANSAC** using only the inliers from the first stage. This recovers the absolute scale via the previously triangulated 3D points.

This hierarchical filtering is crucial for filtering out:

1. Inaccurate feature matches on repetitive textures.

2. Moving objects (which violate the static world assumption).

3.3 Evaluation on KITTI Sequences

We evaluate our system on five sequences: Sequence 01 through Sequence 05. The results are summarized in Table 3.

Sequence	Environment	Avg ATE (m)	Avg RPE-5 (m)
Seq 01	Highway	247.87	1.9990
Seq 02	Urban	111.49	0.1216
Seq 03	City loop	6.45	0.0822
Seq 04	City street	2.27	0.1815
Seq 05	Residential	17.56	0.1041

Table 3: VO Performance on KITTI Sequences 01–05.

Analysis of Sequence 01 (Highway) Failure: Sequence 01 exhibits the highest error (ATE: 247.87m). This highway environment presents several challenges for standard Visual Odometry:

1. **Dynamic Objects**: Most vehicles on the highway are moving at high speeds relative to the camera, which violates the static scene assumption required for reliable ego-motion estimation.

2. **High Velocity and Low Overlap**: Faster vehicle speeds result in larger displacements between consecutive frames, reducing the overlap and making point tracking less robust.

3. **Lack of Nearby Features**: Feature points are primarily located on far-field objects (distant trees, horizon), where triangulation depth uncertainty is highest.

4. **Repeated Textures**: Guardrails and road markings create repetitive patterns that can deceive feature descriptors.

3.4 Ablation Study: RANSAC and Scale

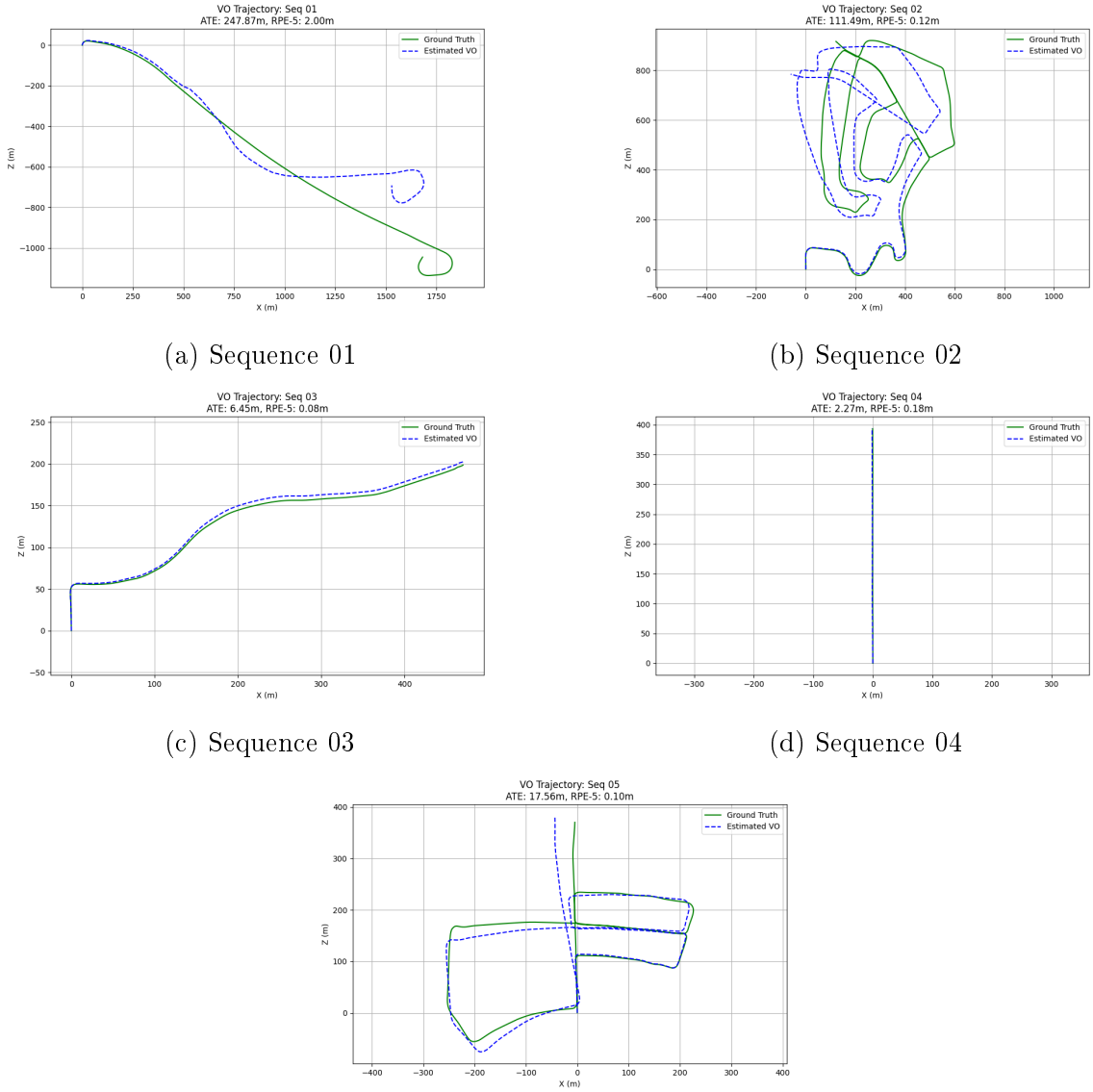
The Importance of RANSAC for outlier rejection and Stereo Triangulation for metric scale recovery is evaluated on sequences 02 and 03. Table 4 highlights the critical role of these components.

Configuration	Seq 02 ATE (m)	Seq 03 ATE (m)
Full Stereo VO	**111.49**	**6.45**
No RANSAC (Outliers)	2.00e12+	2.52e11+
No Scale (Monocular)	186.99	108.55

Table 4: Ablation study for VO components.

3.5 Trajectory Visualizations

Trajectory plots for all sequences visualize the cumulative drift and the system's ability to maintain global consistency. The estimated paths compared to ground truth are shown in Figure 4.



(a) Sequence 01

(b) Sequence 02

(c) Sequence 03

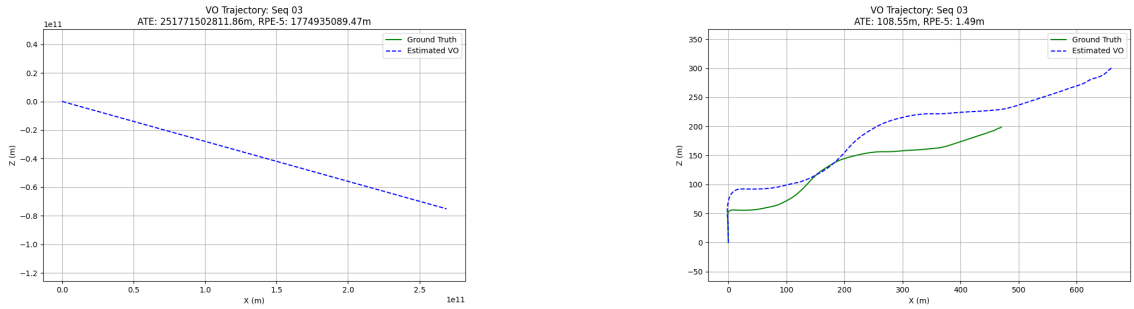
(d) Sequence 04

(e) Sequence 05

Figure 4: Estimated Trajectories for Sequences 01–05.

3.6 Ablation Visualizations (Seq 03)

Figure 5 illustrates the catastrophic failure without RANSAC and the significant drift without metric scale recovery.



(a) No RANSAC (S03)

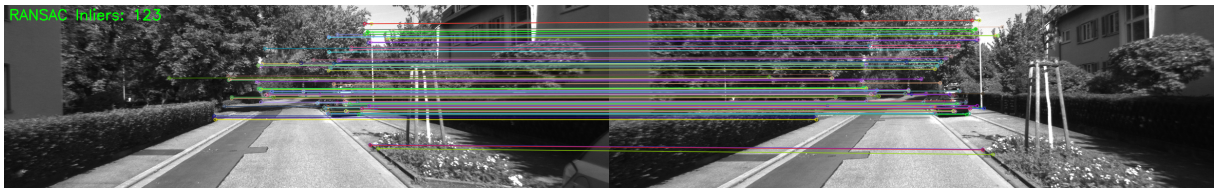
(b) No Scale (S03)

Figure 5: VO Ablation trajectory comparisons.

3.7 Feature Matching and Inliers

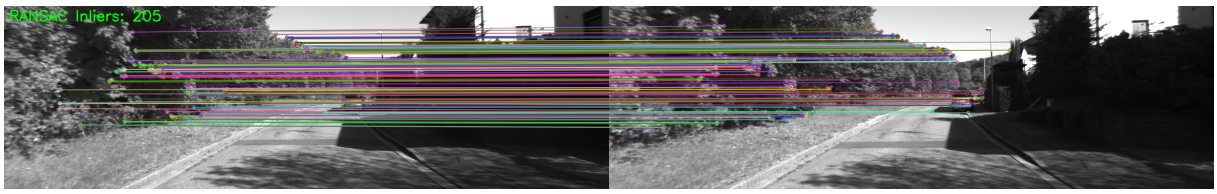
Sample frames showing the ORB feature matches and the inliers identified by the RANSAC process are provided below for Sequence 03 in Figure 6. For results on other sequences, refer to Appendix A.2.

3.7.1 Sequence 03 Matching (Representative)



(a) S03 Frame 10

(b) S03 Frame 100



(c) S03 Frame 300

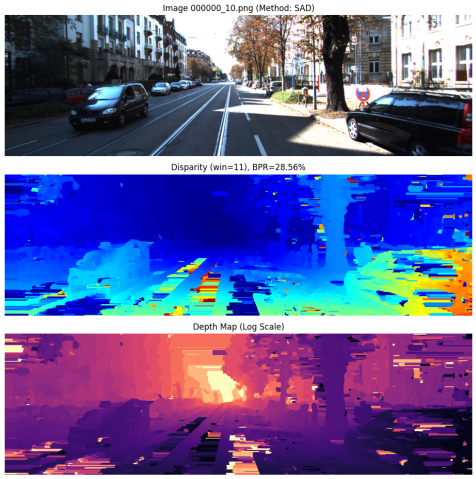
Figure 6: ORB Inlier Matching Sequence 03 (Frames 10, 100, 300).

A Appendix

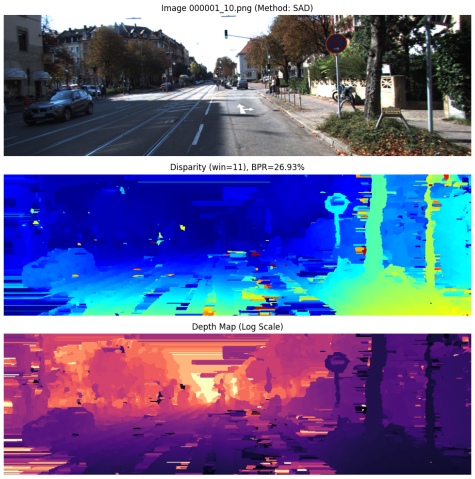
A.1 Additional Stereo Results

The following pages contain the visualization results for 10 frames of the Scene Flow dataset using SAD, SSD, and NCC cost functions with an 11x11 window size. Figures 7, 8, and 9 show the disparity maps generated by each method.

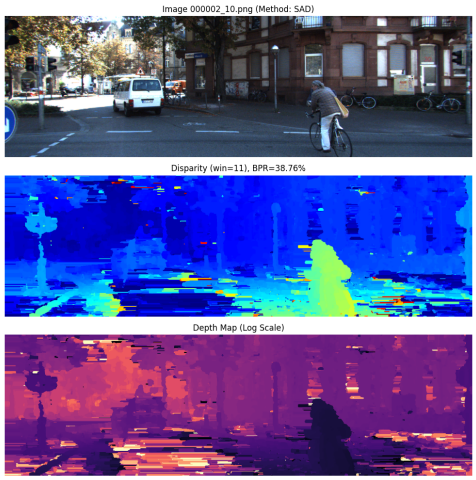
A.1.1 SAD Results



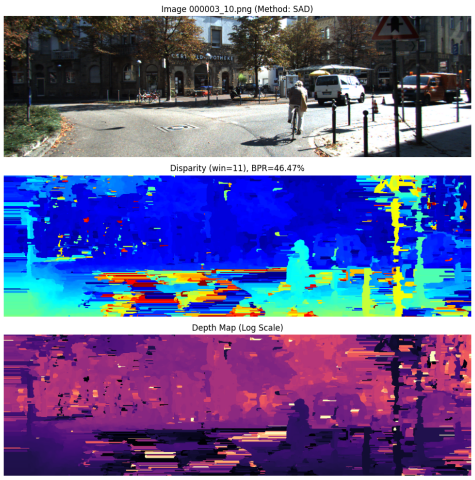
(a) Frame 00



(b) Frame 01

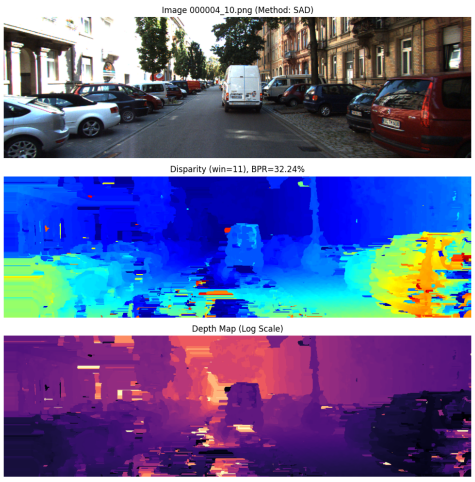


(c) Frame 02

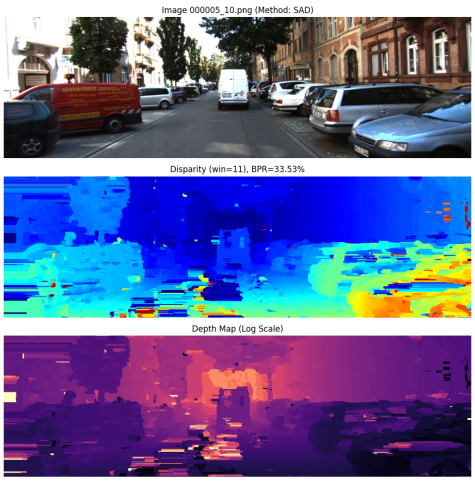


(d) Frame 03

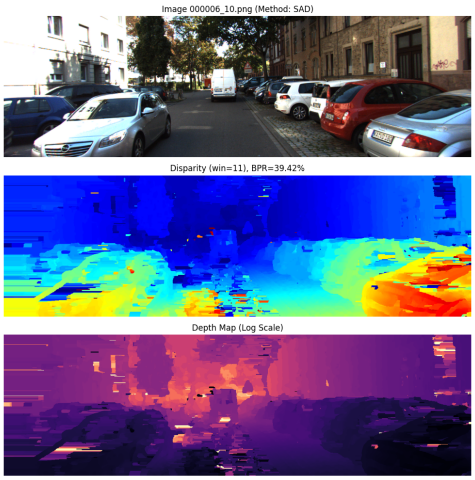
Figure 7: Visual Results for SAD (Frames 00–03).



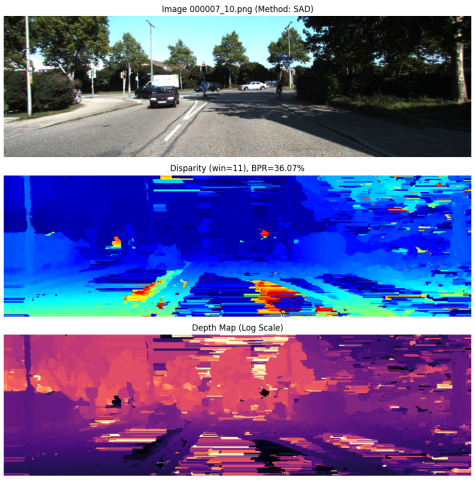
(e) Frame 04



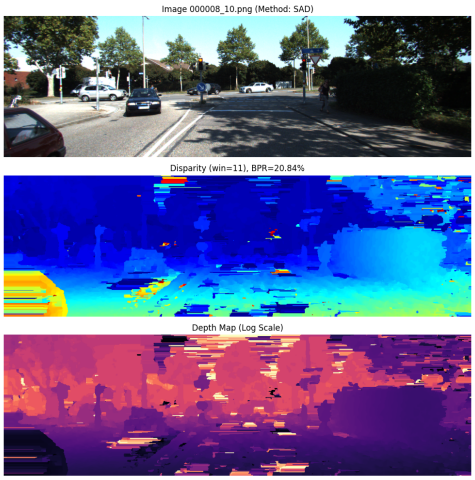
(f) Frame 05



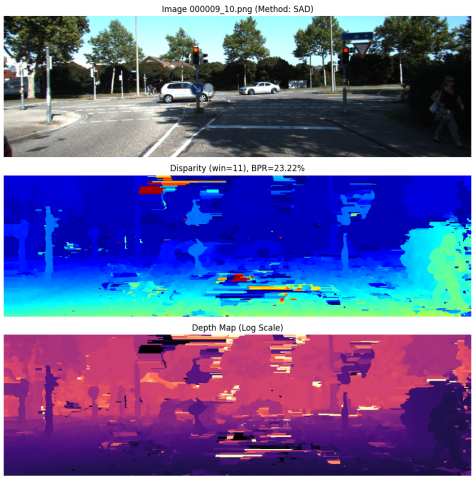
(g) Frame 06



(h) Frame 07



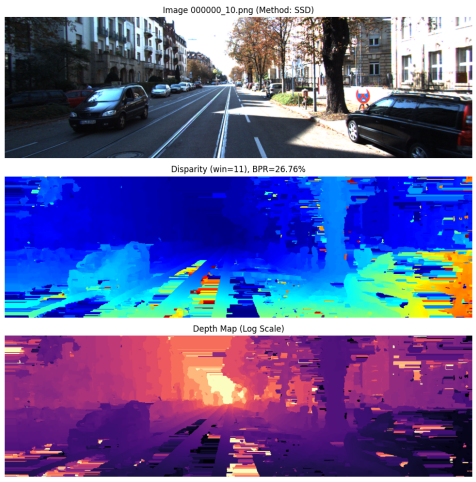
(i) Frame 08



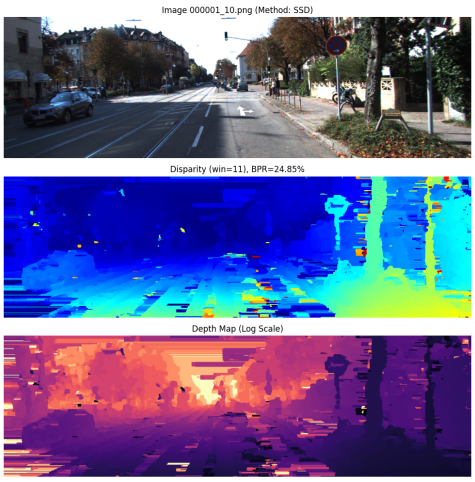
(j) Frame 09

Figure 7: Visual Results for SAD (Frames 04–09, Continued).

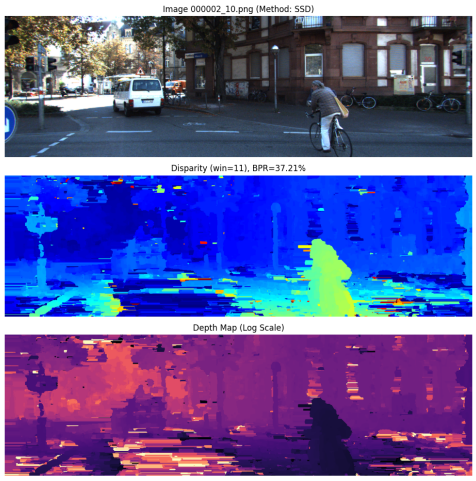
A.1.2 SSD Results



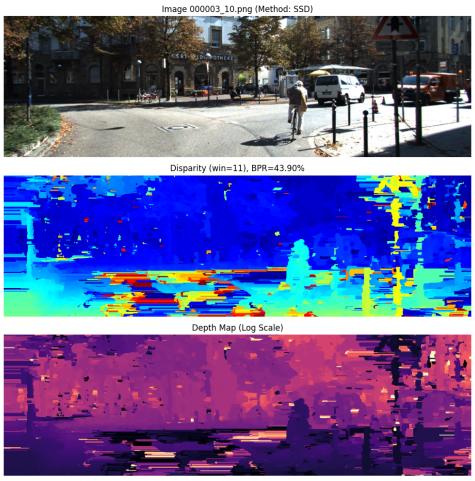
(a) Frame 00



(b) Frame 01

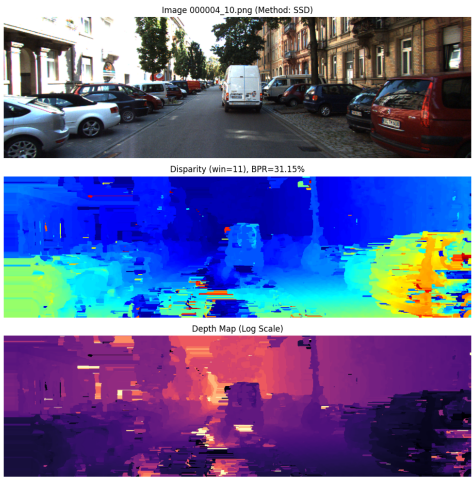


(c) Frame 02

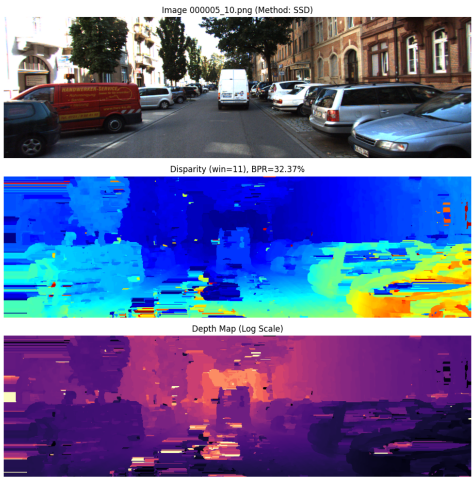


(d) Frame 03

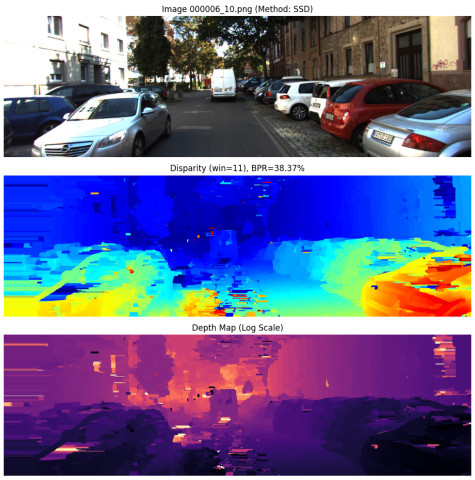
Figure 8: Visual Results for SSD (Frames 00–03).



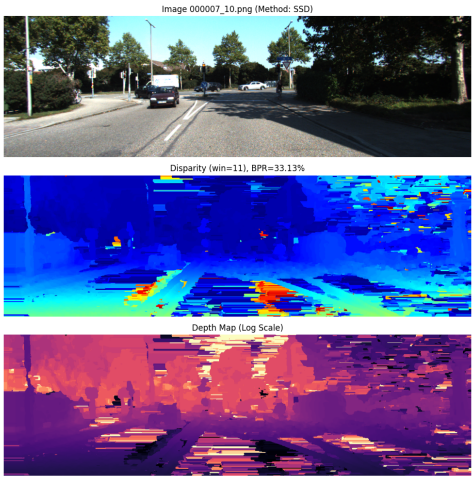
(e) Frame 04



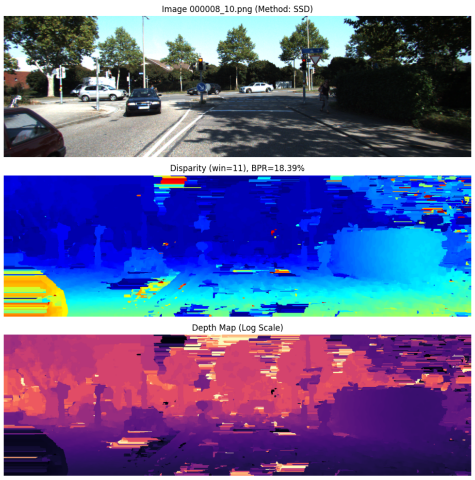
(f) Frame 05



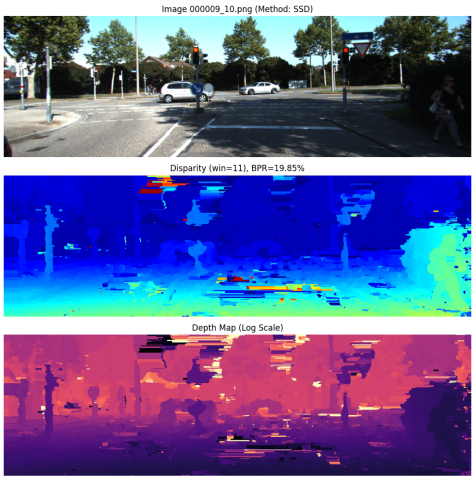
(g) Frame 06



(h) Frame 07



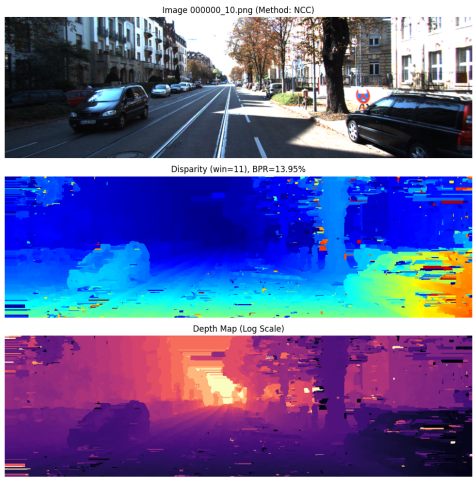
(i) Frame 08



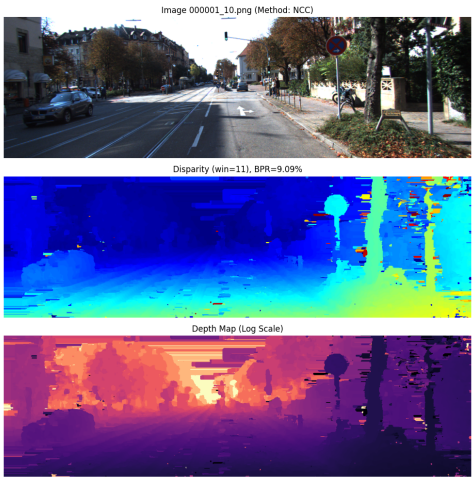
(j) Frame 09

Figure 8: Visual Results for SSD (Frames 04–09, Continued).

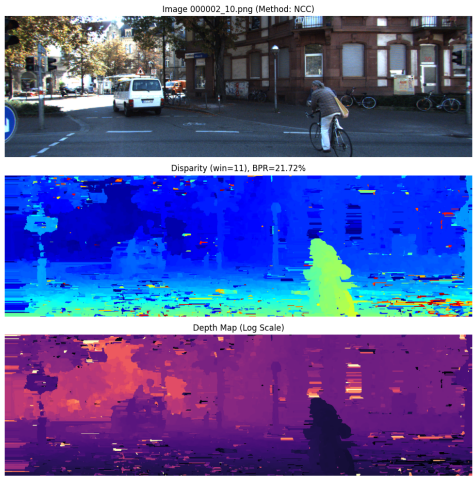
A.1.3 NCC Results



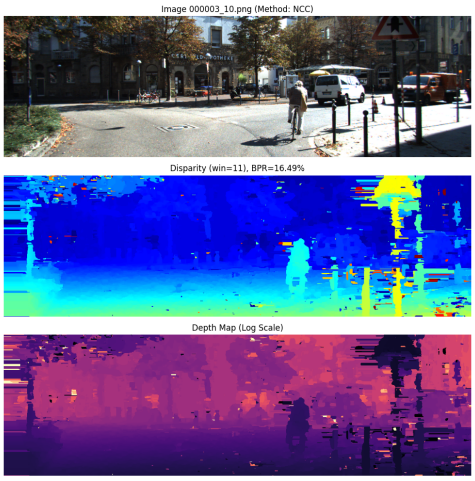
(a) Frame 00



(b) Frame 01

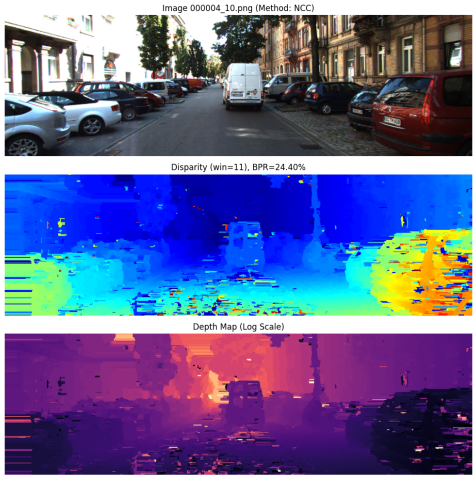


(c) Frame 02

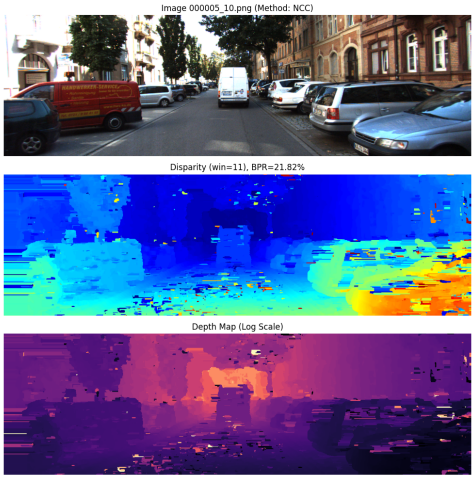


(d) Frame 03

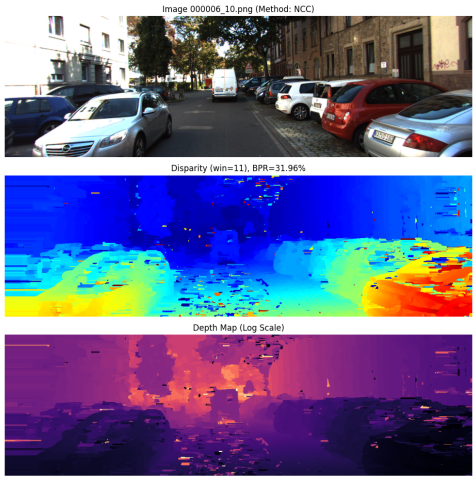
Figure 9: Visual Results for NCC (Frames 00–03).



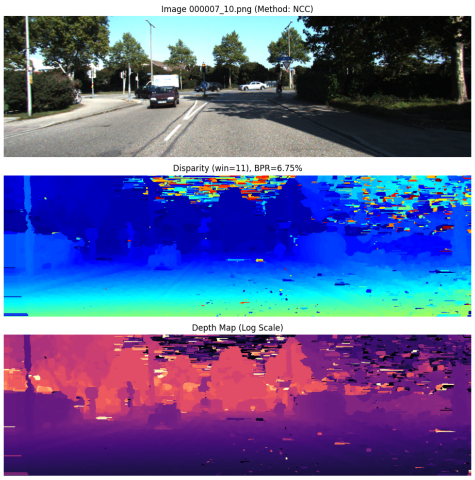
(e) Frame 04



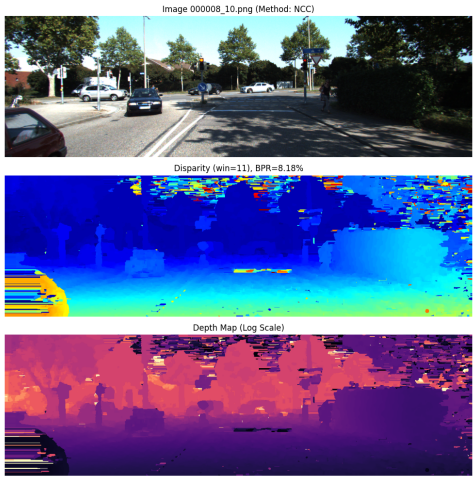
(f) Frame 05



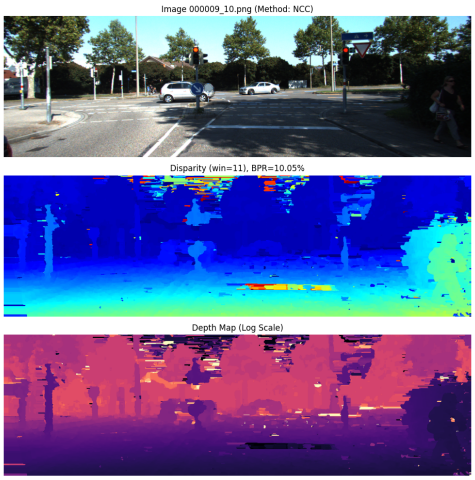
(g) Frame 06



(h) Frame 07



(i) Frame 08



(j) Frame 09

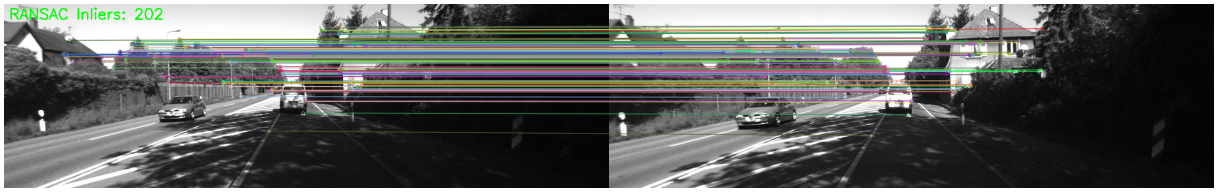
Figure 9: Visual Results for NCC (Frames 04–09, Continued).

A.2 Additional Visual Odometry Results

This section contains additional feature matching visualizations for Sequences 04 and 05 in Figures 10 and 11.

A.2.1 Sequence 04 Matching

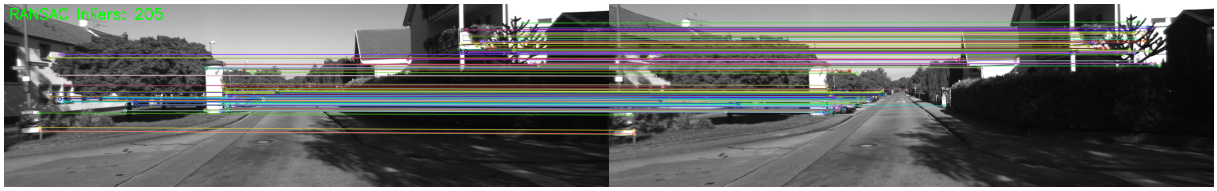
(a) S04 Frame 10



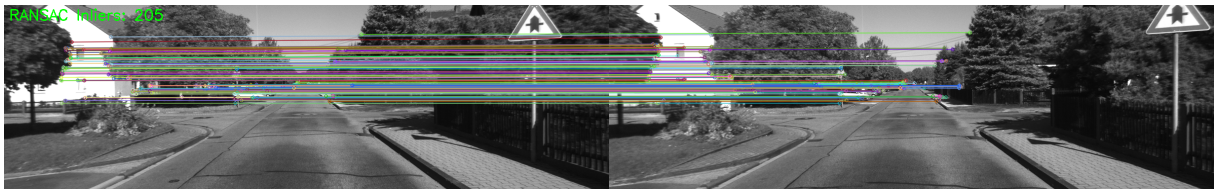
(b) S04 Frame 100

Figure 10: ORB Inlier Matching Sequence 04.

A.2.2 Sequence 05 Matching



(a) S05 Frame 10



(b) S05 Frame 100

Figure 11: ORB Inlier Matching Sequence 05.