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.

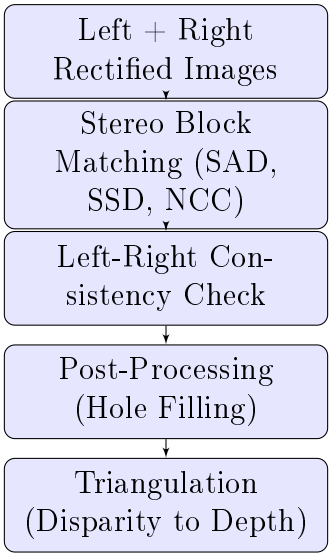


Figure 1: Stereo Depth Pipeline Diagram.

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 5 frames.

Metric	Window	Avg Bad-Pixel Rate (%)	Avg MAE
SAD	5x5	48.99	11.84
SAD	11x11	34.59	8.72
SSD	5x5	47.51	11.55
SSD	11x11	32.78	8.29
NCC	5x5	39.37	8.85
NCC	**11x11**	**17.26**	**3.38**

Table 1: Ablation study for Stereo Depth.

2.5 Failure Cases

- **Occlusions:** Pixels visible in the left image but blocked in the right create invalid disparities. These were handled by the Left-Right check and filled with horizontal streaking.

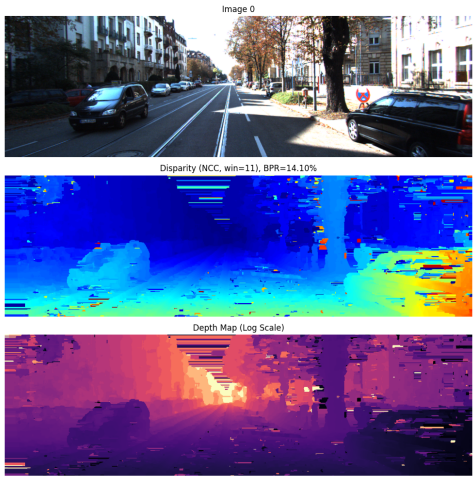
- **Repeated/Low Textures:** Road surfaces and glass reflectances create ambiguity in the NCC/SAD scores, leading to noisy "holes" or misalignments in the disparity map.

KITTI Image ID	BPR (%)	MAE (px)	Likely Failure Reason
000002_10.png	47.11	11.60	Specularity on vehicles
000006_10.png	45.67	8.63	Occlusion boundaries
000007_10.png	43.75	11.82	Overexposure on road surface
000010_10.png	42.44	8.30	Repetitive foliage patterns

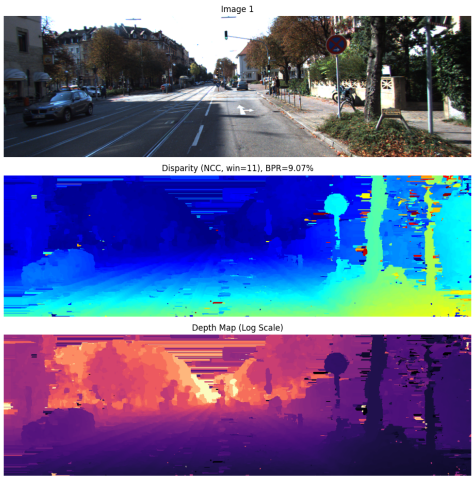
Table 2: Top 5 failure frames from Part A detected using `find_failures.py`.

2.6 Visual Results (10 Examples)

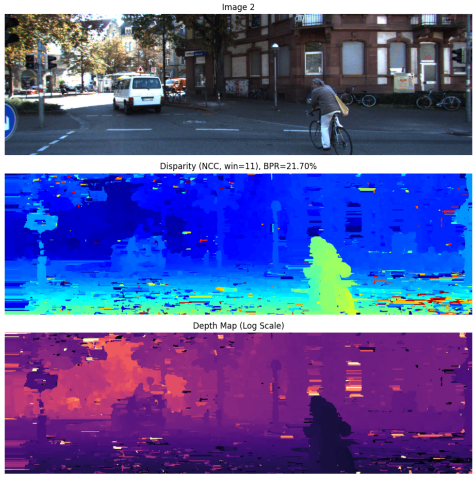
The following figures showcase the Disparity and Depth maps for the first 10 frames of the Scene Flow dataset using the optimized NCC (11x11) algorithm.



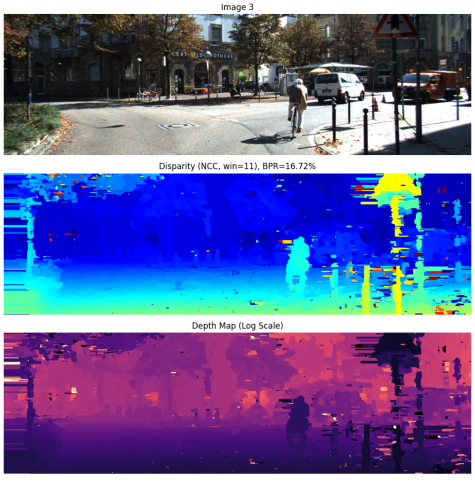
(a) Frame 00



(b) Frame 01

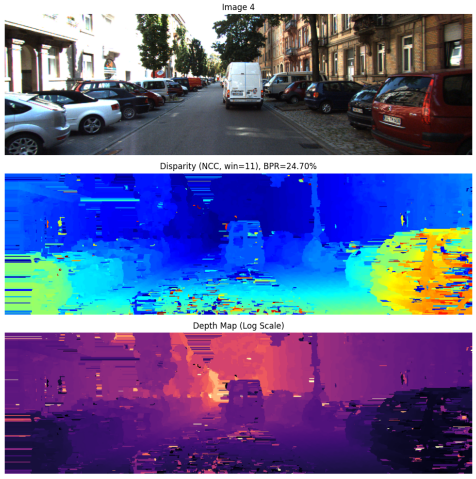


(c) Frame 02

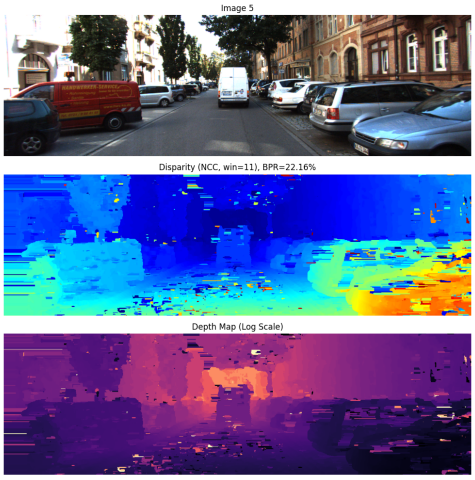


(d) Frame 03

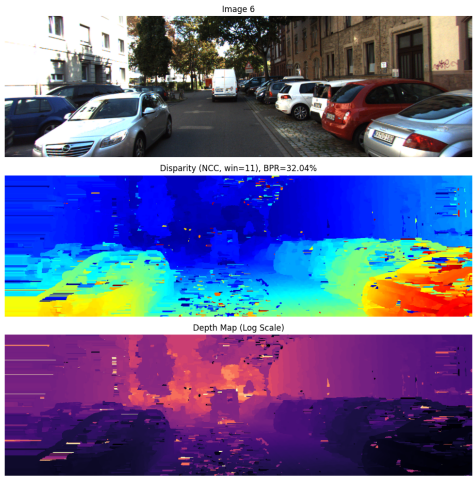
Figure 2: Qualitative Stereo Depth Results (Frames 00–03).



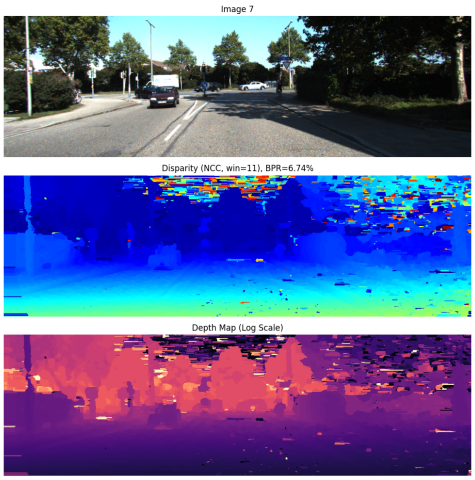
(e) Frame 04



(f) Frame 05

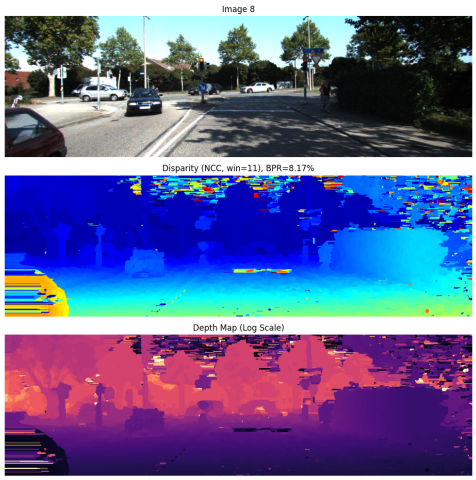


(g) Frame 06

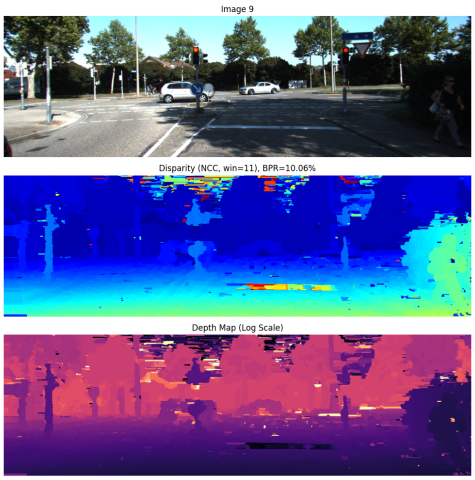


(h) Frame 07

Figure 2: Qualitative Stereo Depth Results (Frames 04–07, Continued).



(i) Frame 08



(j) Frame 09

Figure 2: Qualitative Stereo Depth Results (Frames 08–09, Continued).

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.

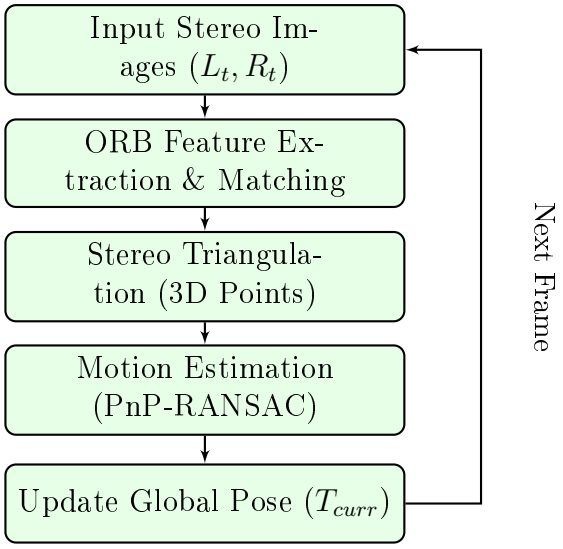


Figure 3: Simplified Stereo Visual Odometry Pipeline.

- **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 two sequences: Sequence 03 (City) and Sequence 01 (Highway, 1101 frames).

Configuration	Sequence	Absolute Trajectory Error (m)	Relative Pose Error (m)
Stereo VO (Full)	03	**6.4498**	**0.0822**
Stereo VO (Full)	01	247.8687	1.9990
Ablation: No RANSAC	03	2.51e11+	N/A
Ablation: No Scale	03	108.5530	N/A

Table 3: Numerical evaluation for KITTI Sequences.

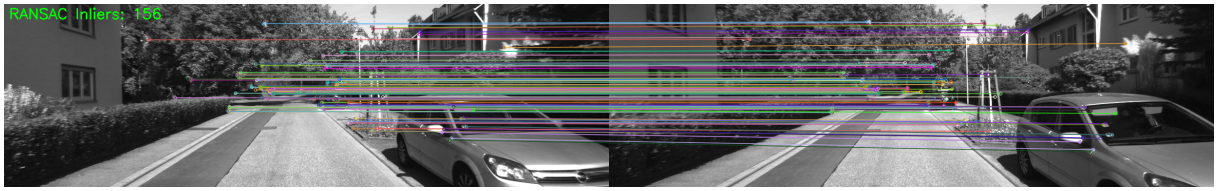
3.4 Ablation Study (VO Importance)

The studies highlight two critical components:

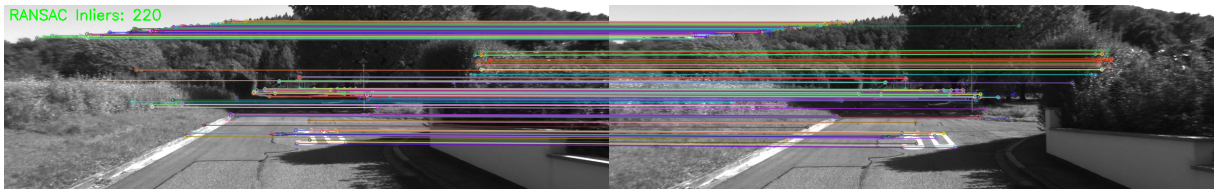
- **RANSAC Importance**: Without RANSAC, a single outlier match causes the pose estimation to explode immediately.
- **Stereo-based Scale Importance**: By using the stereo triangulation, we recover the **absolute metric scale**. Without it (monocular mode), the system defaults to an arbitrary scale factor of 1 per frame, leading to nearly 100 meters of cumulative scale drift on Seq 03.

3.5 Feature Matches and Inliers (Seq 03)

Figure 4 shows the ORB feature matches on Sequence 03.



(a) S03 Frame 5



(b) S03 Frame 500

Figure 4: Temporal ORB Matching with RANSAC: Sequence 03.

3.6 Large-Scale Trajectory Plot (Seq 03)

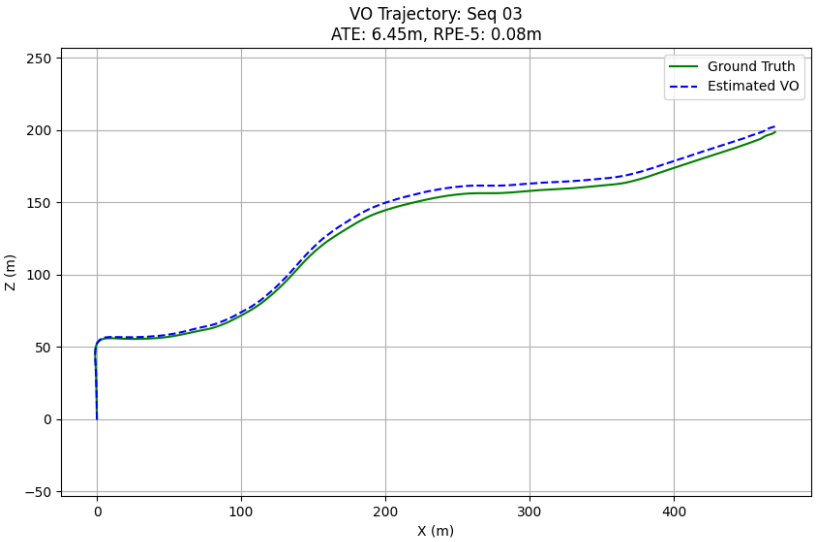
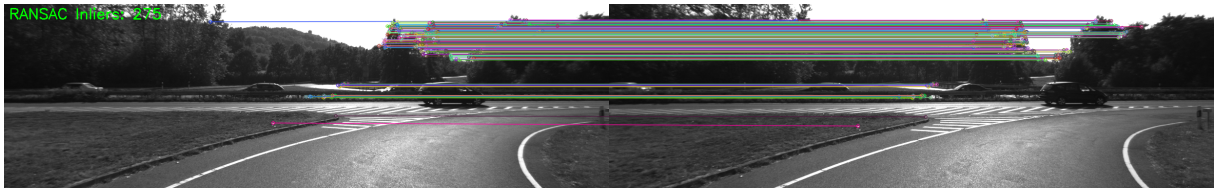


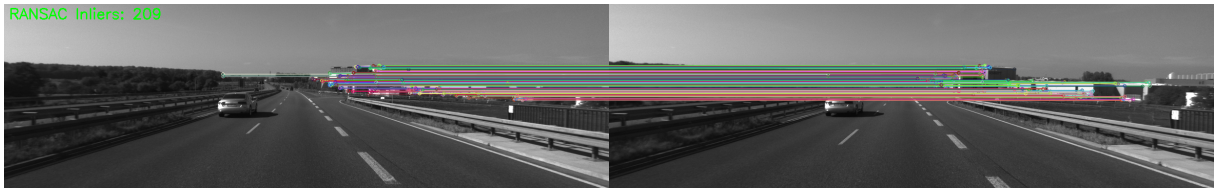
Figure 5: Trajectory for Sequence 03 (City loop).

3.7 Results for Sequence 01 (Highway)

Sequence 01 contains high-speed driving on a highway. Feature tracking is more challenging due to the distance of landmarks and high frame-to-frame displacement.



(a) S01 Frame 5



(b) S01 Frame 500

Figure 6: Temporal ORB Matching: Sequence 01.

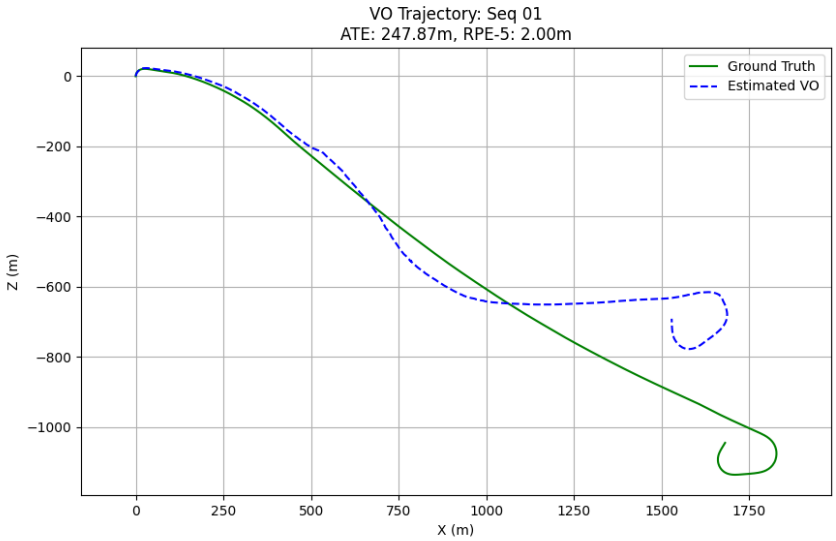


Figure 7: Trajectory for Sequence 01 (Highway, 1101 Frames).