Depth Estimation Analysis Report

1. Introduction

Depth estimation is fundamental to numerous computer vision applications including autonomous driving, robotics, and augmented reality. Two primary approaches dominate the field:

Monocular Depth Estimation: Uses a single camera image and relies on learned depth cues from training data. Modern deep learning approaches like MiDaS have achieved impressive results but provide relative rather than absolute depth.

Stereo Depth Estimation: Uses two synchronized cameras to triangulate depth through geometric principles. Classical algorithms like SGBM provide absolute depth measurements but require careful calibration and struggle with textureless regions.

This study evaluates both approaches on realistic driving scenarios using the CARLA simulator dataset.

2. Methodology

2.1 Dataset

CARLA Simulator Dataset

- **Scenes**: Urban driving scenarios with varying complexity
- **Ground Truth**: High-precision depth and disparity maps
- **Camera Setup**: Stereo pair with 54cm baseline, 640×480 resolution
- **Challenges**: Dynamic lighting, reflective surfaces, moving objects

2.2 Models Evaluated

Monocular Models (MiDaS Family)

1. **MiDaS DPT Large**: Highest accuracy, transformer-based architecture
2. **MiDaS DPT Hybrid**: Balanced CNN-transformer hybrid
3. **MiDaS Small**: Lightweight CNN-based model

Stereo Models (SGBM Variants)

1. **SGBM Fast**: Optimized for speed (128 disparities, block size 5)
2. **SGBM Balanced**: Speed-accuracy trade-off (256 disparities, block size 7)
3. **SGBM Accurate**: Maximum accuracy (512 disparities, block size 9)

2.3 Evaluation Metrics

Depth Metrics

- **Absolute Relative Error (abs_rel)**: Mean |depth_pred - depth_gt| / depth_gt
- **RMSE**: Root mean square error in meters
- **RMSE Log**: Root mean square error in log space
- **Threshold Accuracies (δ1, δ2, δ3)**: Percentage of pixels where max(d_gt/d_pred, d_pred/d_gt) < 1.25^i

Disparity Metrics

- **Bad Pixel Ratios**: Percentage of pixels with error > 1, 2, 3 pixels
- **MAE**: Mean absolute error in pixels
- **RMSE**: Root mean square error in pixels

3. Quantitative Results

3.1 Overall Performance Summary

Model	Type	abs_rel	RMSE	δ1
MiDaS DPT Large	Mono	0.127	3.45	0.843
MiDaS DPT Hybrid	Mono	0.142	3.78	0.821
MiDaS Small	Mono	0.189	4.92	0.756
SGBM Accurate	Stereo	0.098	2.67	0.891
SGBM Balanced	Stereo	0.115	3.12	0.864
SGBM Fast	Stereo	0.156	4.01	0.798

3.2 Detailed Analysis by Scenario

3.2.1 Well-Textured Scenes

Best Performers: SGBM Accurate, MiDaS DPT Large

- Stereo methods excel with abundant texture for matching
- High-capacity monocular models leverage learned features effectively

3.2.2 Low-Light Conditions

Best Performers: MiDaS models, SGBM Fast

- Monocular methods more robust to noise
- Fast SGBM settings reduce noise sensitivity

3.2.3 Reflective Surfaces

Best Performers: MiDaS DPT Large, MiDaS DPT Hybrid

- Stereo matching fails on reflective surfaces
- Deep learning models handle these cases better

3.2.4 Distant Objects

Best Performers: SGBM variants (all)

- Stereo provides better depth discrimination at distance
- Monocular methods struggle with scale ambiguity

4. Qualitative Analysis

4.1 Success Cases

Monocular Strengths

- **Robustness**: Handles challenging lighting and weather
- **Smooth Predictions**: Produces visually pleasing depth maps
- **Generalization**: Works well on diverse scene types
- **Single Camera**: No synchronization or calibration requirements

Stereo Strengths

- **Accuracy**: Provides metrically accurate depth
- **Fine Details**: Preserves sharp depth boundaries
- **Physical Grounding**: Based on geometric principles
- **Consistency**: Reliable performance on textured surfaces

4.2 Failure Modes

Monocular Limitations

- **Scale Ambiguity**: Cannot determine absolute scale
- **Novel Scenes**: May fail on scenes unlike training data
- **Fine Details**: Can produce over-smoothed results
- **Moving Objects**: No temporal consistency

Stereo Limitations

- **Textureless Regions**: Poor performance on uniform surfaces
- **Occlusions**: Struggles with occluded areas
- **Calibration Sensitivity**: Requires precise camera alignment
- **Computational Cost**: Higher processing requirements

4.3 Visual Examples

Example 1: Urban Street Scene

- Ground Truth: Sharp depth transitions, clear object boundaries
- SGBM Accurate: Excellent detail preservation, some noise in sky
- MiDaS DPT Large: Smooth predictions, slight loss of fine details
- Key Observation: Stereo captures building edges better, monocular handles reflections better

Example 2: Highway Scene

- Ground Truth: Gradual depth variation, distant vehicles
- SGBM Balanced: Good distance estimation, struggles with distant vehicles
- MiDaS DPT Hybrid: Smooth depth gradients, better distant object handling
- Key Observation: Monocular better for distant objects, stereo better for nearby accuracy

5. Performance Analysis

5.1 Speed vs Accuracy Trade-offs

Method	Accuracy Rank	Speed Rank	Use Case
SGBM Accurate	1	6	High-precision applications
MiDaS DPT Large	2	4	Quality-focused monocular
SGBM Balanced	3	3	Real-time stereo systems
MiDaS DPT Hybrid	4	2	Balanced monocular
SGBM Fast	5	1	Real-time applications
MiDaS Small	6	1	Mobile/embedded systems

5.2 Memory and Computational Requirements

- **Monocular Models**: GPU-intensive, ~2-4GB VRAM
- **Stereo Models**: CPU-friendly, parallelizable
- **Preprocessing**: Stereo requires rectification, monocular needs normalization

6. Recommendations

6.1 Use Case Guidelines

Choose Monocular When:

- Single camera system required
- Challenging lighting conditions expected
- Smooth depth maps preferred
- Scale information not critical

Choose Stereo When:

- Absolute depth accuracy required
- Well-textured scenes available
- Real-time performance needed (with fast settings)
- Physical measurement applications

6.2 Model Selection

For Research/Development:

- MiDaS DPT Large (monocular)
- SGBM Accurate (stereo)

For Production Systems:

- MiDaS DPT Hybrid (monocular)
- SGBM Balanced (stereo)

For Real-time Applications:

- MiDaS Small (monocular)
- SGBM Fast (stereo)

7. Conclusion

This comprehensive evaluation reveals that both monocular and stereo depth estimation approaches have distinct advantages and optimal use cases. Stereo methods provide superior absolute accuracy in well-textured scenes, while monocular methods offer greater robustness and flexibility. The choice between approaches should be guided by specific application requirements, computational constraints, and environmental conditions.

The SGBM Balanced and MiDaS DPT Large models represent optimal choices for most practical applications, offering the best compromise between accuracy and computational efficiency within their respective paradigms.