# Direct Multichannel Tracking


Carlos Jaramillo* Yuichi Taguchi & Chen Feng

CUNY Graduate Center Mitsubishi Electric Research Laboratories (MERL)

*This work was done while interning at MERL


## Contributions

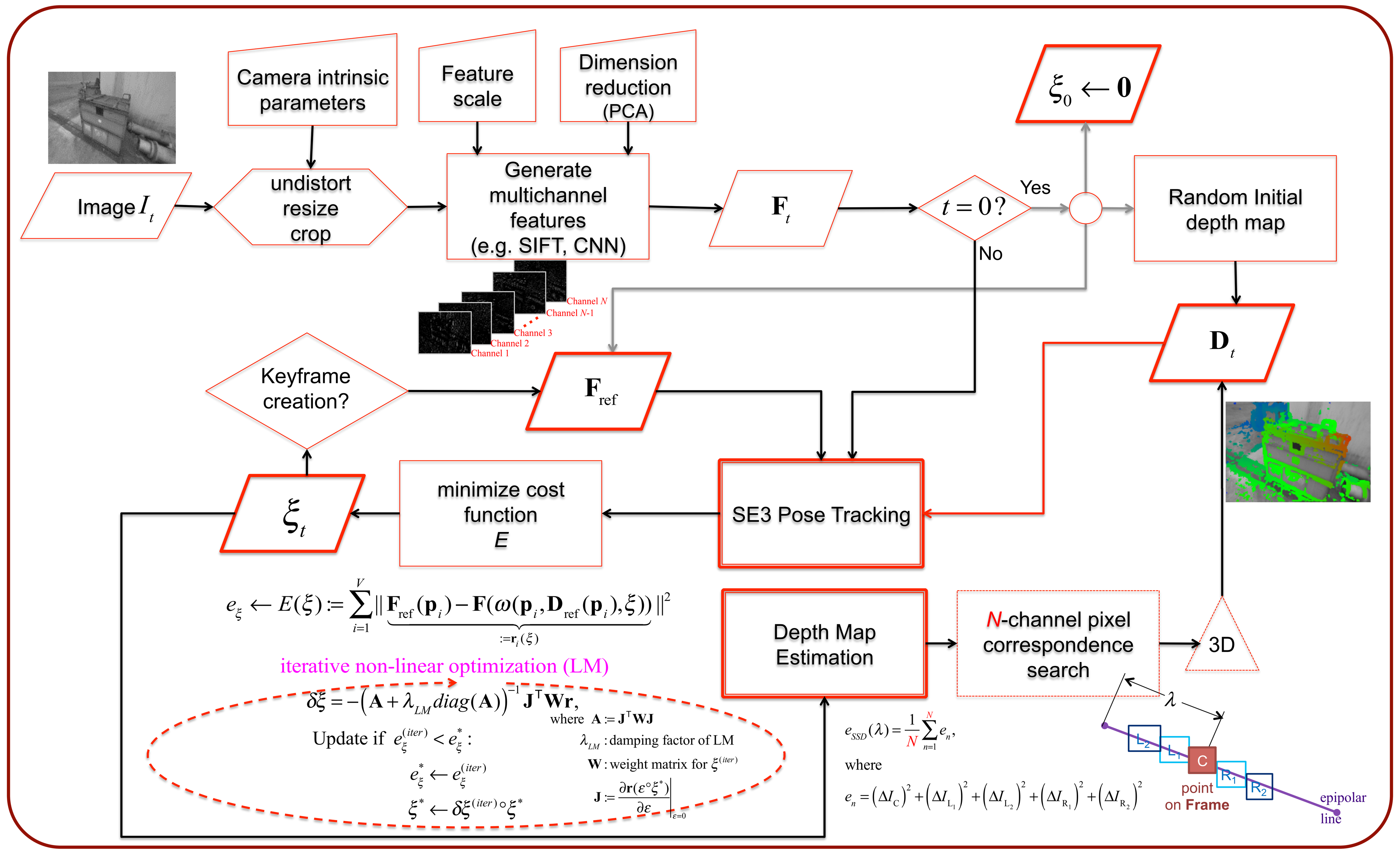
We extend direct visual odometry algorithms [1][2], which use the single grayscale channel and assume intensity constancy, by:

- Multichannel feature generation (feature constancy assumption)
  - **Dense-SIFT** features (a SIFT descriptor for each pixel)
  - **CNN** features from pre-trained models:
    - **AlexNet** (*conv1* layer)
    - **Siamese** (**"*siam*"** at *SpatialConvolution_7* layer )
- Direct image alignment of multichannel features
  - Multichannel SE3 pose tracking
  - Multichannel depth estimation
- **Denser** 3D reconstruction and **more accurate** camera pose tracking
- Real-time via PCA-based dimensionality reduction

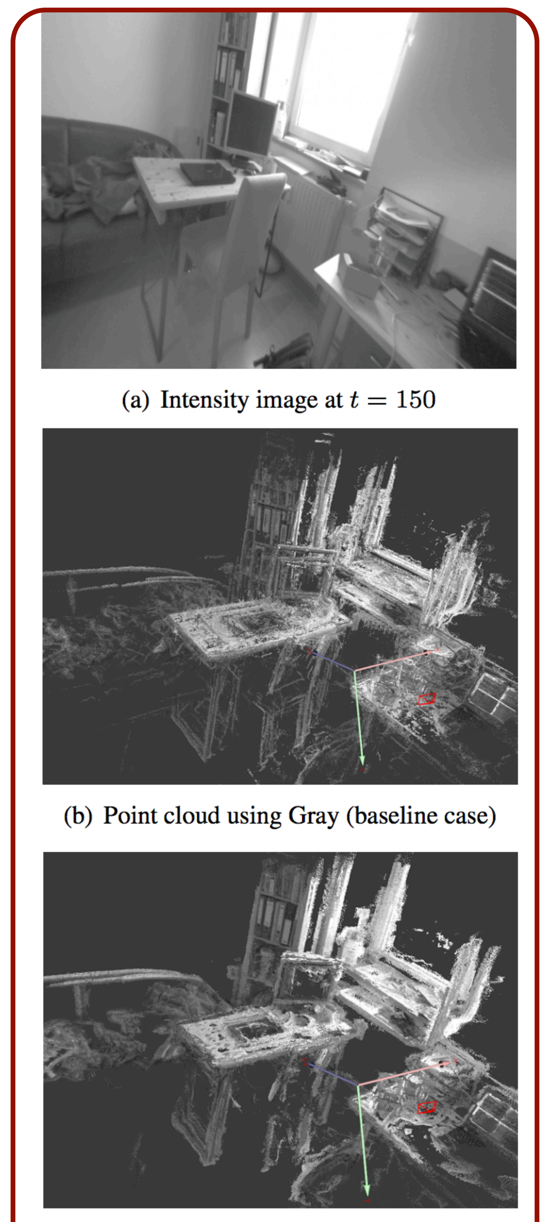
**Evaluation** with 5 public datasets:

- TUM DSO (50 *real* sequences with GT)
- KITTI Odometry (11 *real* seqs with GT)
- ICL-NUIM (8 *synthetic* seqs)
- Tsukuba (New 4 *synthetic* seqs)
- LSD-SLAM (4 *real* seqs, without GT)

## DMT Pipeline




## 3D Comparison



(a) Intensity image at $t = 150$

(b) Point cloud using Gray (baseline case)

(c) Point cloud using AlexNet

## Quantitative Experimental Results

TUM DSO dataset: all 50 sequences

### Overall Average Errors

| Average Metric \ Feature | Gray | Gray-A | Bit-Planes | SIFT | Siamese | AlexNet |
|---|---|---|---|---|---|---|
| Number of Tracked Frames | 2336 | 2604 | 2095 | 2361 | 636 | **3131** |
| RPE Rotation Error [degree/meter] | 7.44 | 7.31 | 13.73 | 8.17 | 8.10 | **6.14** |
| RPE Trans. Ang. Error [degree/meter] | 18.05 | 16.06 | 38.02 | 21.86 | 20.90 | **12.93** |
| ATE [meters] | 0.38 | 0.46 | 0.45 | 0.37 | **0.07** | 0.65 |

**Note:** RPE here uses the normalized errors by the path length

### Average Rankings

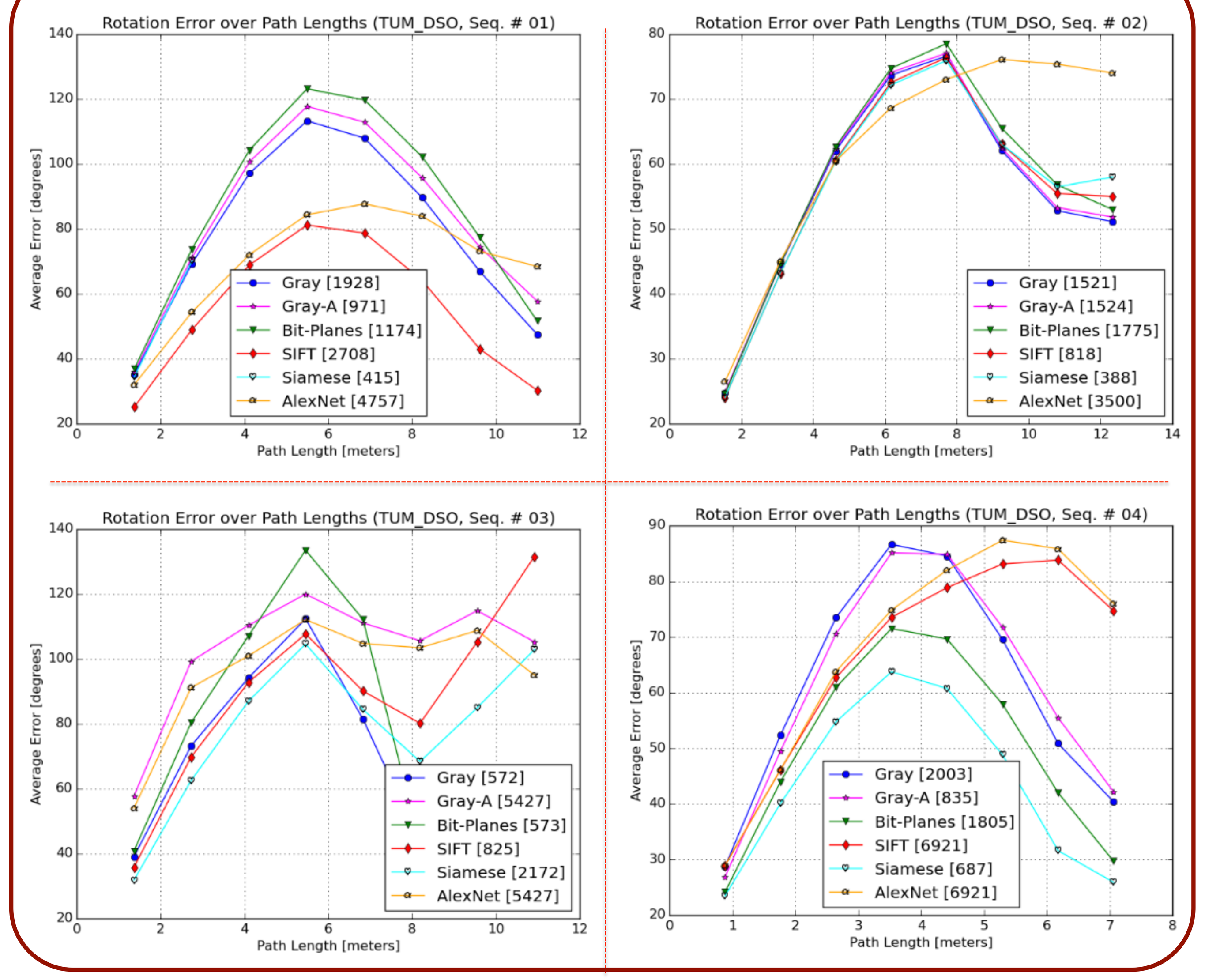
| Feature | Tracked Frames | | Rotation Error | | Trans. Ang. Error | | ATE | |
|---|---|---|---|---|---|---|---|---|
| | Wins Count | Average Ranking | Wins Count | Average Ranking | Wins Count | Average Ranking | Wins Count | Average Ranking |
| Gray | 21 | 2.76 | 3 | 3.58 | 7 | 3.24 | 1 | 3.92 |
| Gray-A | 21 | 2.60 | 3 | 3.62 | 4 | 3.40 | 5 | 3.80 |
| Bit-Planes | 15 | 3.18 | 3 | 4.40 | 5 | 4.14 | 3 | 4.16 |
| SIFT | 20 | 2.82 | 12 | 3.18 | 6 | 3.72 | 8 | 3.42 |
| Siamese | 0 | 5.54 | 10 | 3.72 | 4 | 4.06 | **26** | **1.88** |
| AlexNet | **37** | **1.74** | **19** | **2.50** | **24** | **2.44** | 7 | 3.82 |

Wins Count: Larger is better | Average Ranking: Smaller is better

- Gray: The baseline single-channel algorithm [1][2]
- Gray-A: Gray with the affine lighting model [2]
- Bit-Planes: The 8-bit binary descriptor proposed by [3]

## RPE Rotation Angle

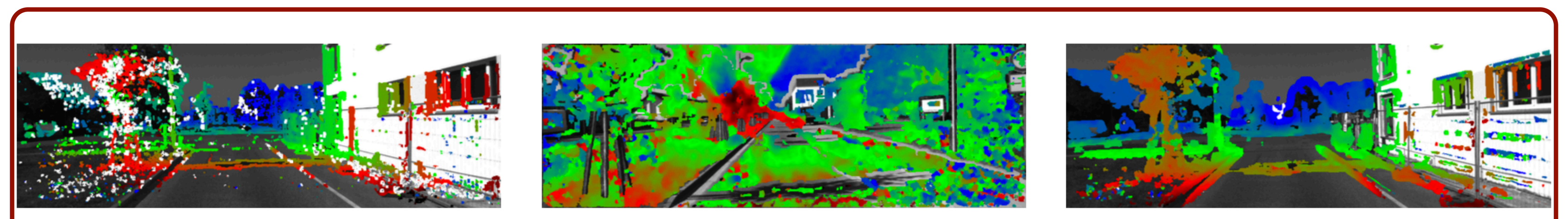
TUM DSO dataset: first 4 sequences




Number in [brackets] specify the successful number of tracked frames

## Depth Maps Comparison

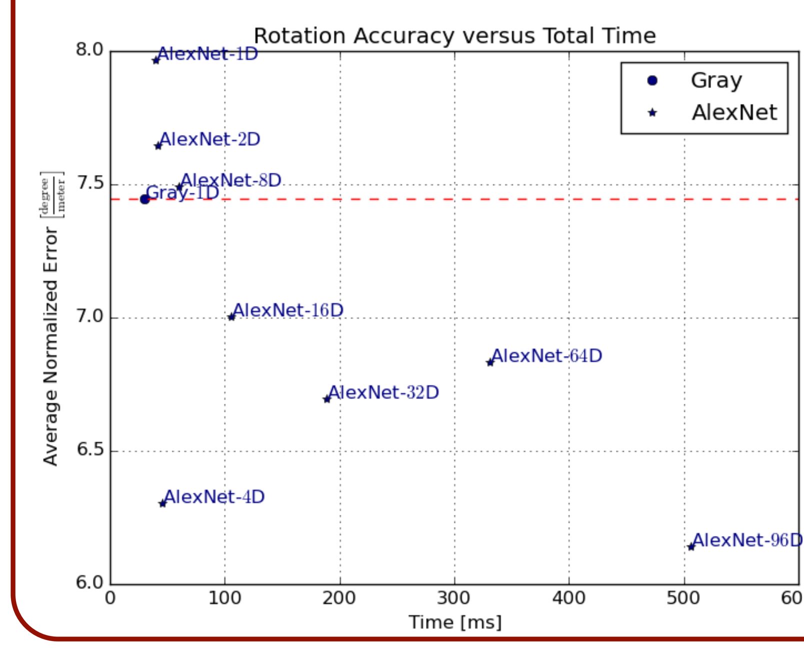
KITTI dataset: sequence No. 6 at $t = 250$



(a) Gray (b) SIFT (c) AlexNet

## Dimensionality Reduction for Real-time Tracking

### Average Accuracy and Timing




| | Feature | Generation (ms) | Tracking (ms) | Total (ms) | Total (fps) |
|---|---|---|---|---|---|
| | Gray (1D) | $1.0 \times 10^1$ | $2.0 \times 10^1$ | 30 | 33 |
| | Gray-A (1D) | $1.0 \times 10^1$ | $2.1 \times 10^1$ | 31 | 32 |
| | Bit-Planes (8D) | $2.0 \times 10^1$ | $1.2 \times 10^2$ | 140 | 7 |
| | SIFT (128D) | $5.3 \times 10^2$ | $7.4 \times 10^2$ | 1270 | 1 |
| | Siamese (256D) | $8.0 \times 10^3$ | $1.0 \times 10^3$ | 9000 | $\frac{1}{9}$ |
| | AlexNet (96D) | $7.9 \times 10^1$ | $4.3 \times 10^2$ | 509 | 2 |
| Reduced | AlexNet-64D | $6.1 \times 10^1$ | $2.7 \times 10^2$ | 331 | 3 |
| Reduced | AlexNet-32D | $4.0 \times 10^1$ | $1.5 \times 10^2$ | 190 | 5 |
| Reduced | AlexNet-16D | $2.9 \times 10^1$ | $7.6 \times 10^1$ | 105 | 10 |
| Reduced | AlexNet-8D | $2.4 \times 10^1$ | $3.6 \times 10^1$ | 60 | 17 |
| Reduced | AlexNet-4D | $2.2 \times 10^1$ | $2.3 \times 10^1$ | 45 | 22 |
| Reduced | AlexNet-2D | $2.1 \times 10^1$ | $2.1 \times 10^1$ | 42 | 24 |
| Reduced | AlexNet-1D | $2.0 \times 10^1$ | $2.0 \times 10^1$ | 40 | 25 |

**References:**


[1] J. Engel, J. Sturm, and D. Cremers. Semi-dense visual odometry for a monocular camera. In Proc. IEEE Int'l Conf. Computer Vision (ICCV), pages 1449–1456, 2013.
[2] J. Engel, T. Schops, and D. Cremers. LSD-SLAM: Large-scale direct monocular SLAM. In Proc. European Conf. Computer Vision (ECCV), Sept. 2014.
[3] H. Alismail, M. Kaess, B. Browning, and S. Lucey. Direct Visual Odometry in Low Light Using Binary Descriptors. IEEE Robotics and Automation Letters, 2(2):444–451, 2017.