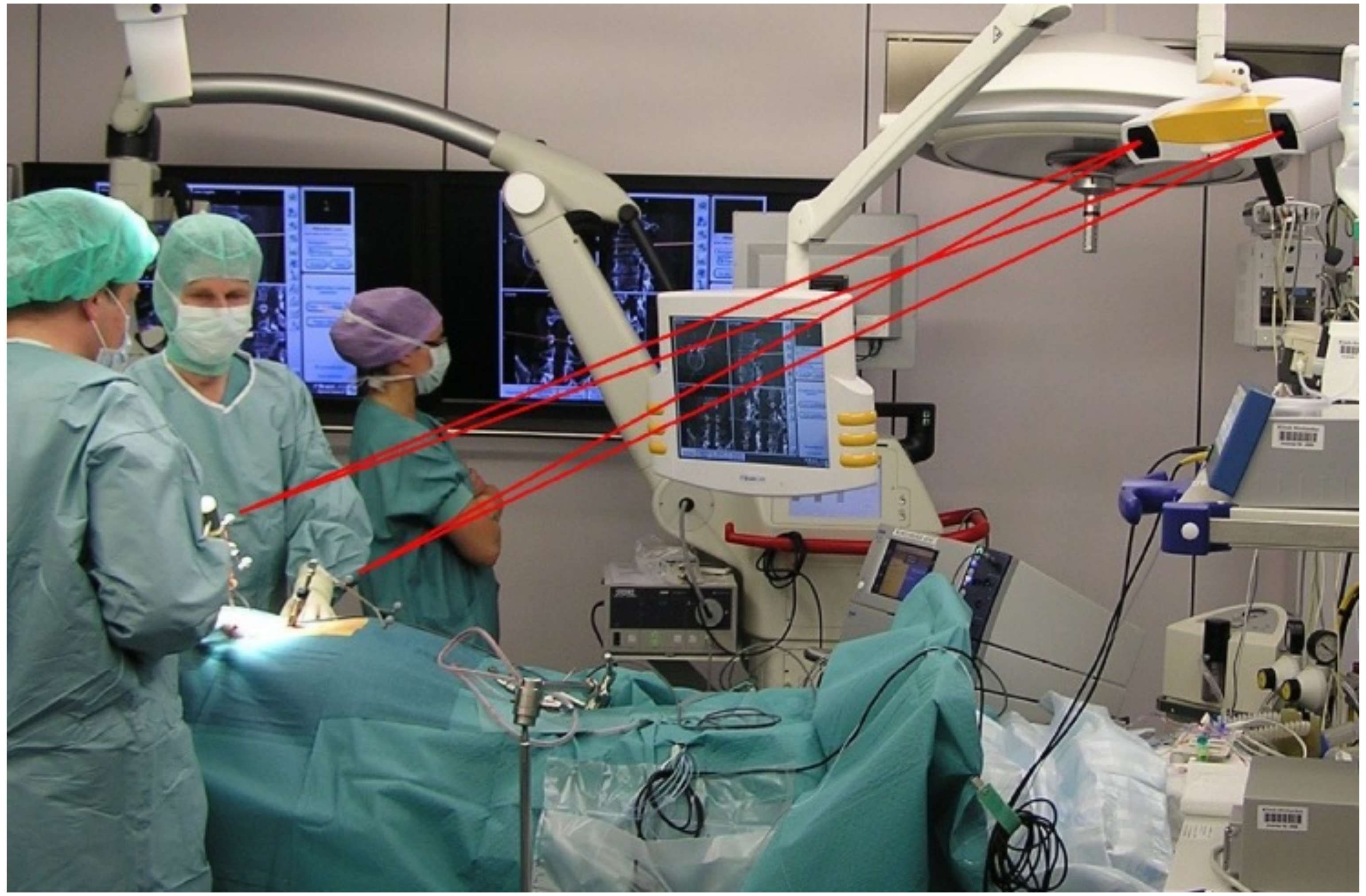
# Recurrent Multi-view 6DoF Pose Estimation for Marker-less Surgical Tool Tracking

Niklas Agethen*, **Janis Rosskamp†,** Tom L. Koller*†,
Jan Klein*, Gabriel Zachmann†

*Fraunhofer MEVIS, Bremen, Germany.
†University of Bremen, Germany.

# Introduction: Surgical Navigation

- Current Approach: Marker-Based
  - Requires line of sight
  - Markers can be contaminated
- Marker-less RGB Approach using DL
  - Fewer occlusions with multiple cameras
  - Cameras can serve additional purposes



Mezger U, Jendrewski C, Bartels M. Navigation in surgery. Langenbecks Arch Surg. 2013 Apr;398(4):501-14. doi: 10.1007/s00423-013-1059-4. Epub 2013 Feb 22. PMID: 23430289; PMCID: PMC3627858.

# Related Work

## 6D Pose Estimation:

- Single-view:
  - GDR-Net [Wang 2021]
- Multi-view:
  - SpyroPose [Haugaard 2023]
  - Surgical Navigation [Hein 2025]
- Sequences:
  - RNNPose [Xu 2024]

## Datasets:

| Properties | Dex-YCBV | T-Less | Hein 2025 | HO3D |
|---|---|---|---|---|
| Hand | ✓ | ✗ | ✓ | ✓ |
| Multi-View | ✓ | ✗ | ✓ | ✗ |
| Small | ✗ | ✗ | ✓ | ✗ |
| OR | ✗ | ✗ | ✓ | ✗ |
| Available | ✓ | ✓ | ✗ | ✓ |

# Our Contributions

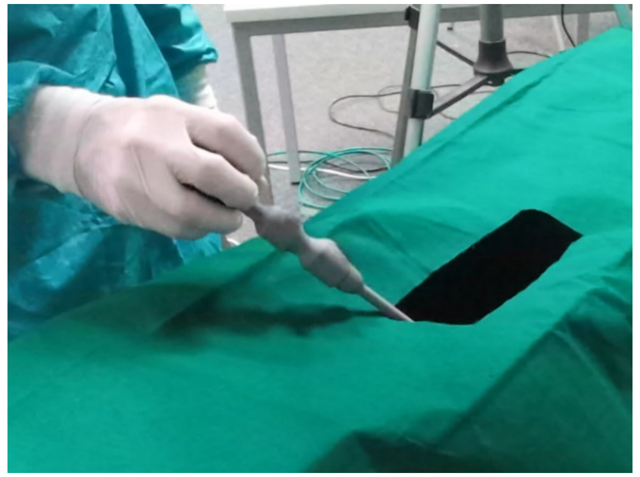
- Novel Architecture: Combines recurrent networks and multi-view input for enhanced spatial-temporal understanding in 6D pose estimation
- Datasets:
  - Real-world dataset for surgical instrument pose estimation
  - Synthetic dataset designed to systematically analyze occlusions

# Dataset - Real

- 2 Objects
- 3 scenes
- 40k images recorded with 4 cameras
- Marker-based annotation [Rosskamp 2024]

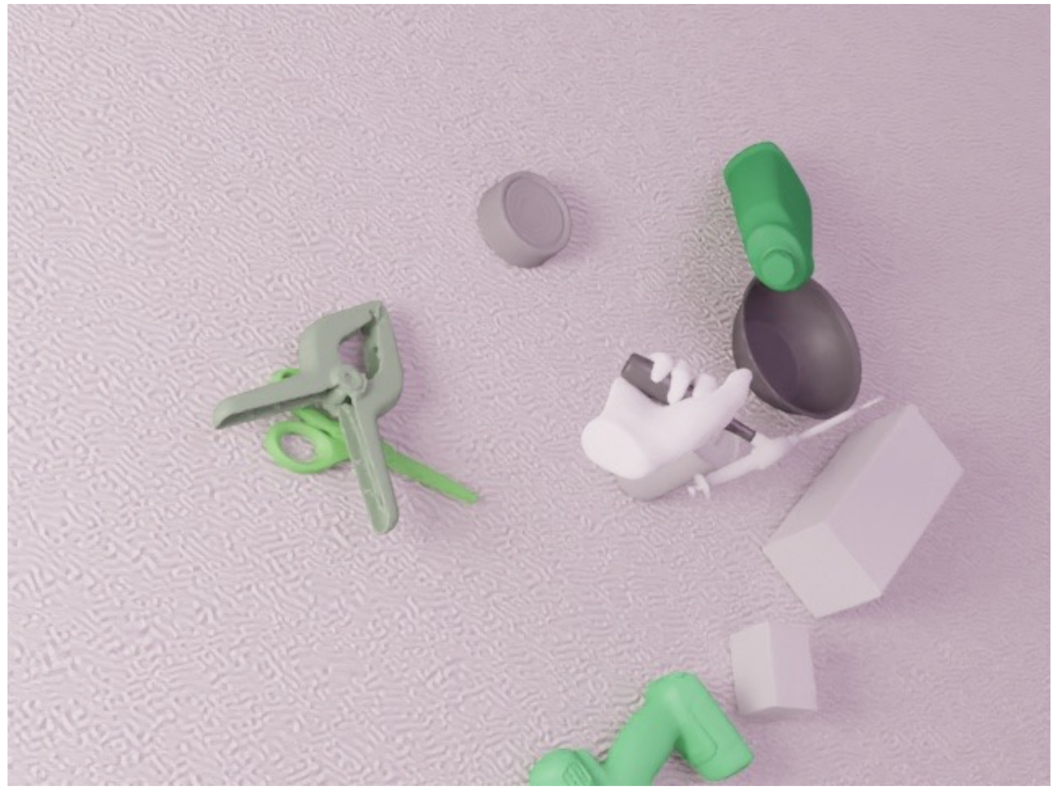


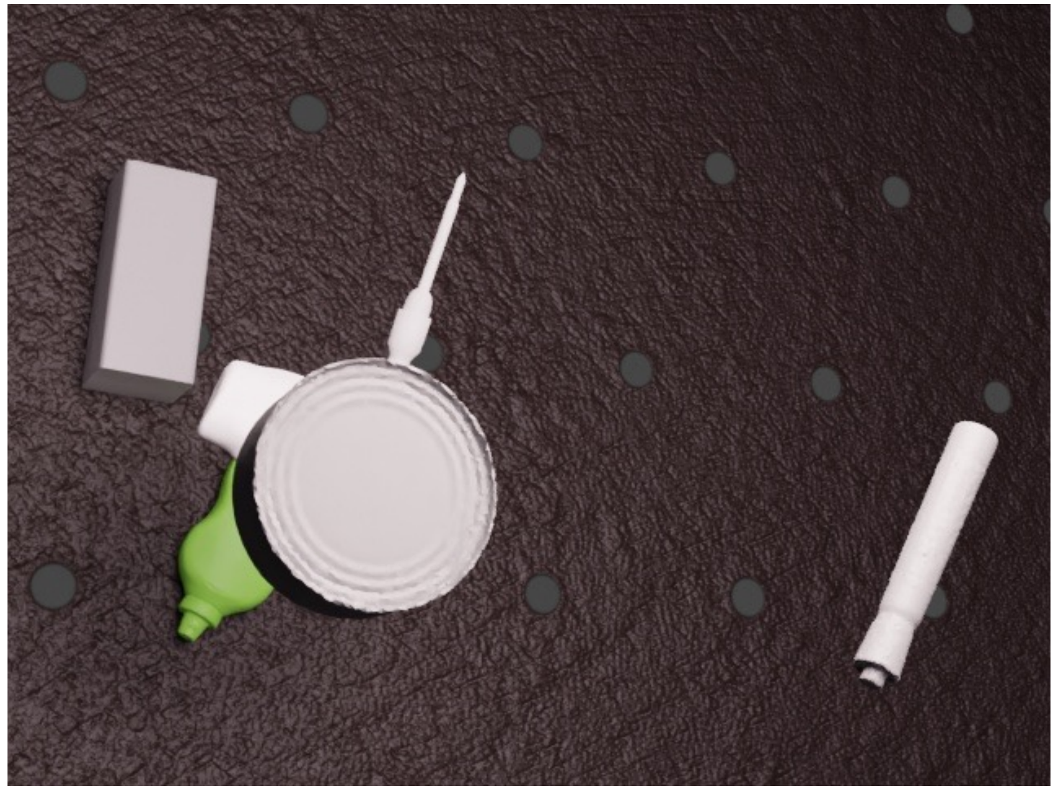





# Dataset - Synthetic

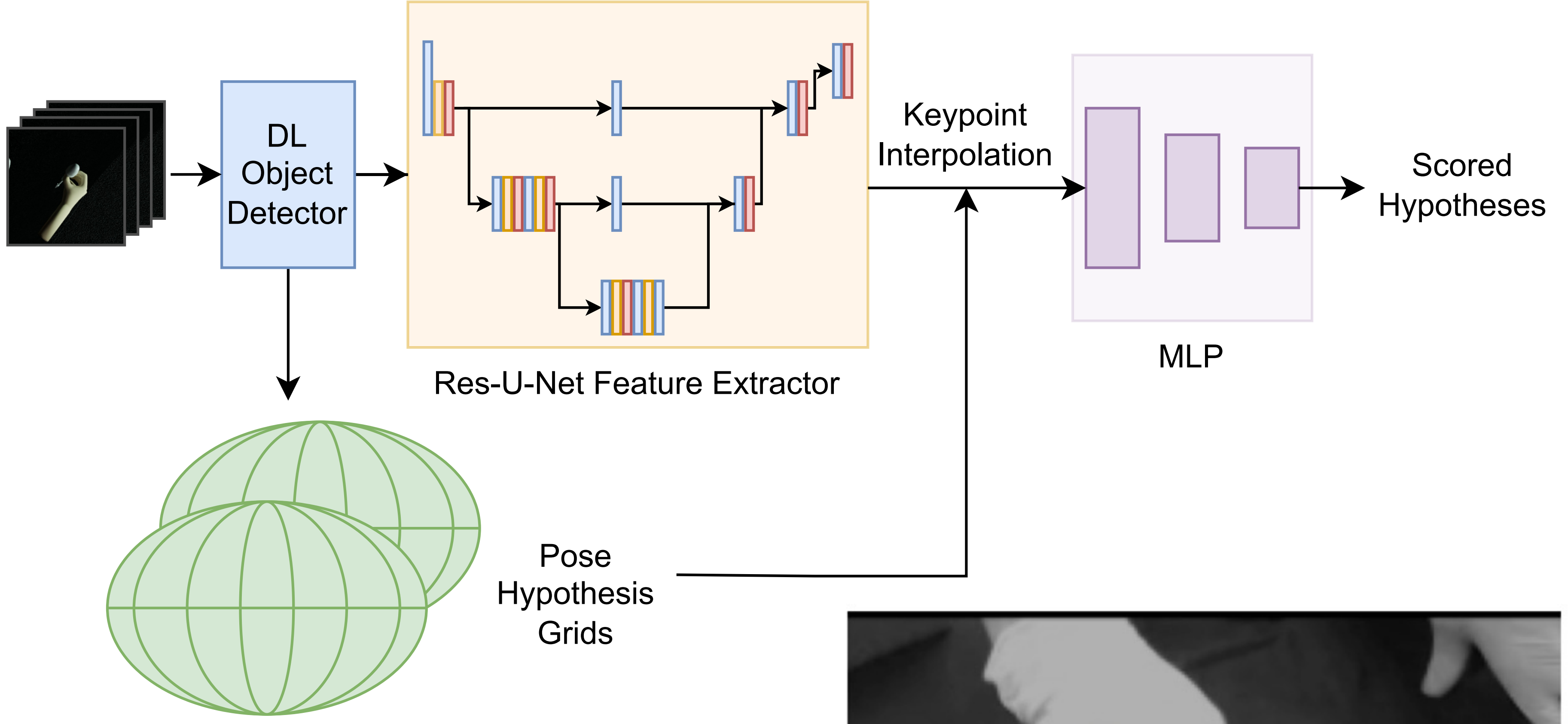
- BlenderProc & Nimble Hand model [Li 2022]
- Object trajectories recorded with MoCap
- 12 frames per scene
- Distractor objects are added from the 6th frame

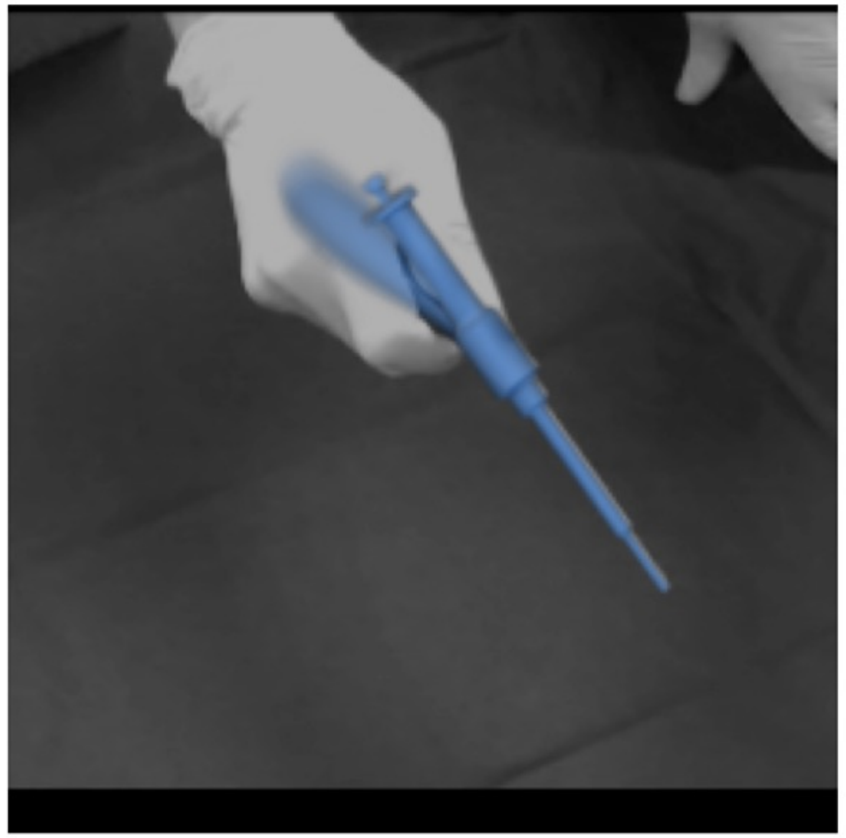




# Architecture - Pose Estimation

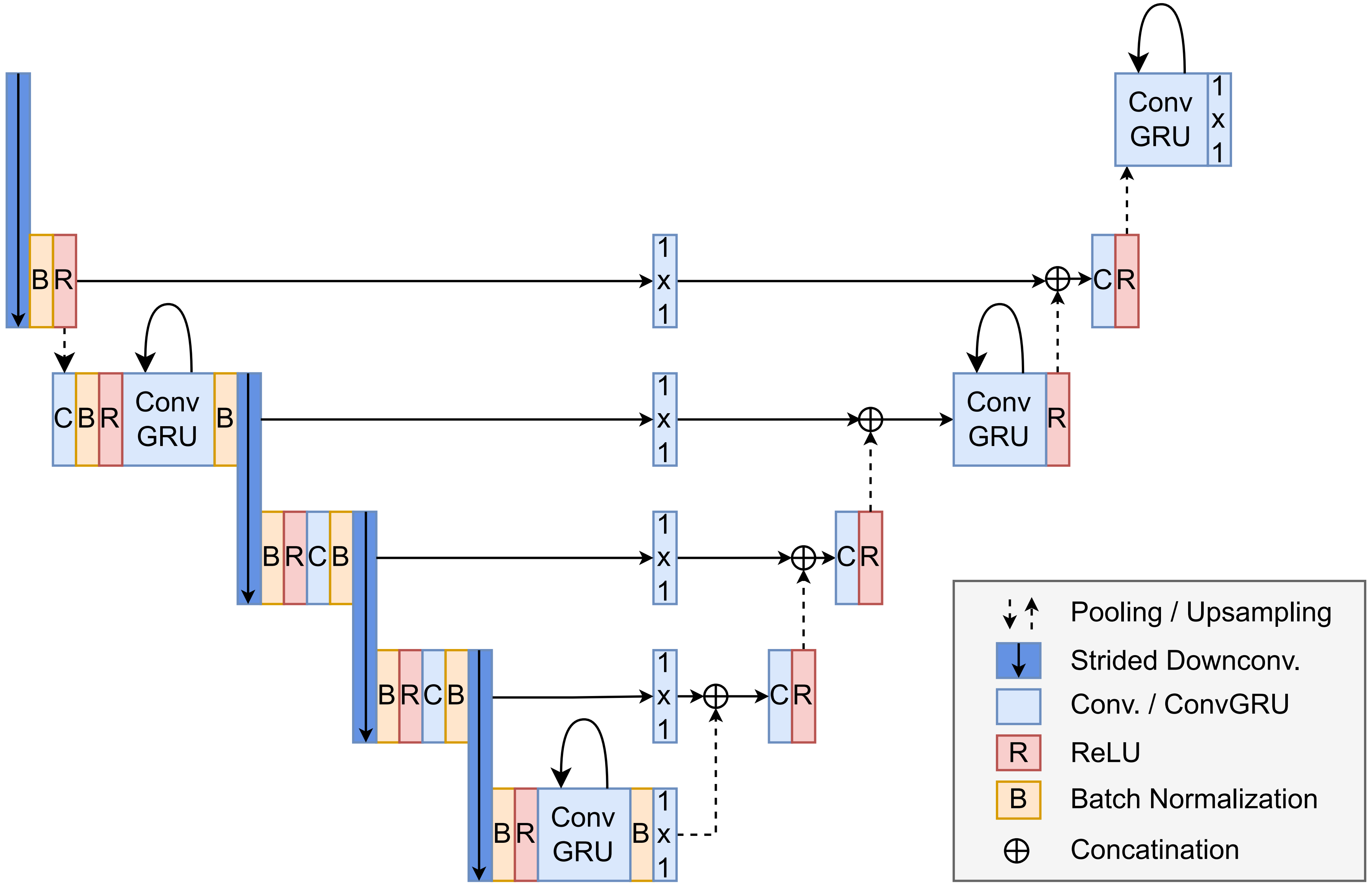
- SpyroPose [Haugaard 2023]
  - U-Net feature extractor
  - Coarse-to-fine hierarchical grids
  - MLP-based scoring of pose



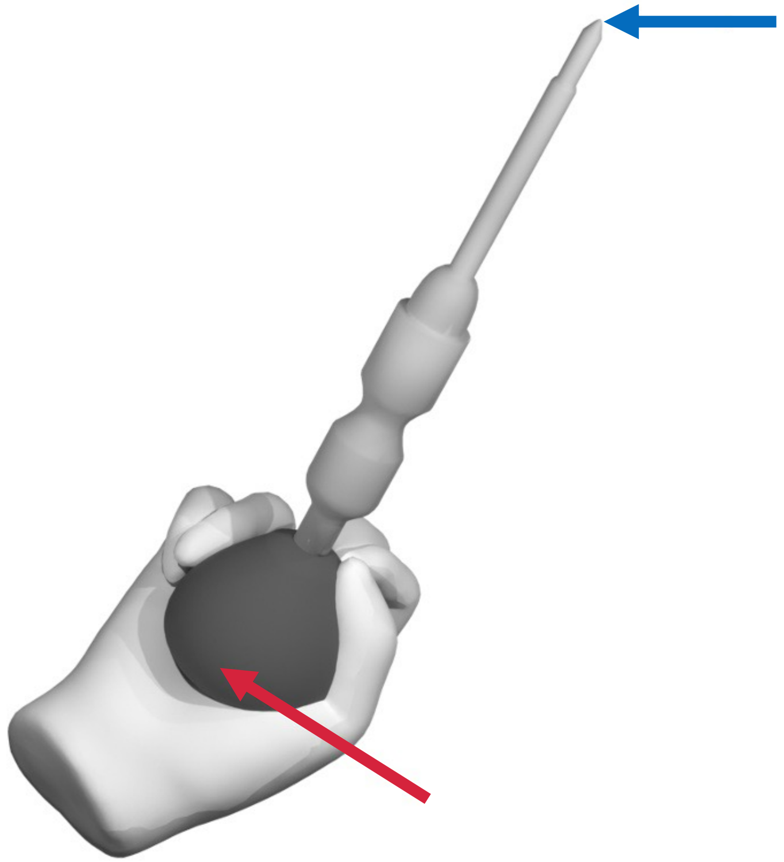



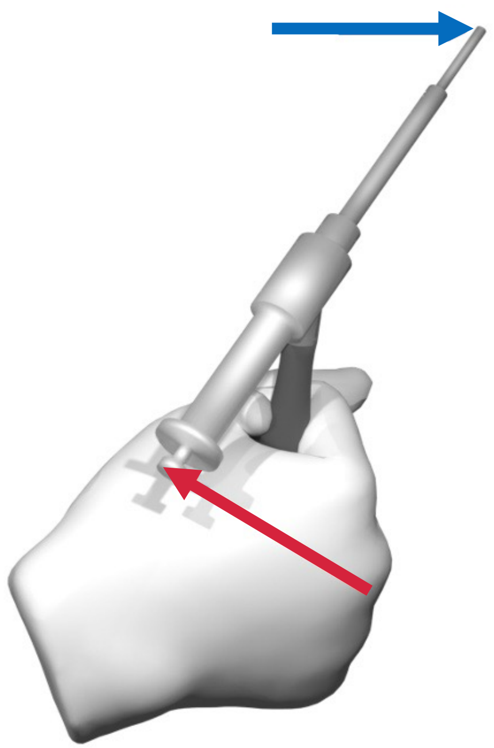
# Our Architecture - Recurrent Pose Estimation

- Apply recurrence in feature extractor (U-Net)
- Gated Recurrent Units (GRU) are not designed for spatial inputs
- **Instead**: ConvGRU [Wang 2018] replace some of convolutional layers



# Results - Multi-View





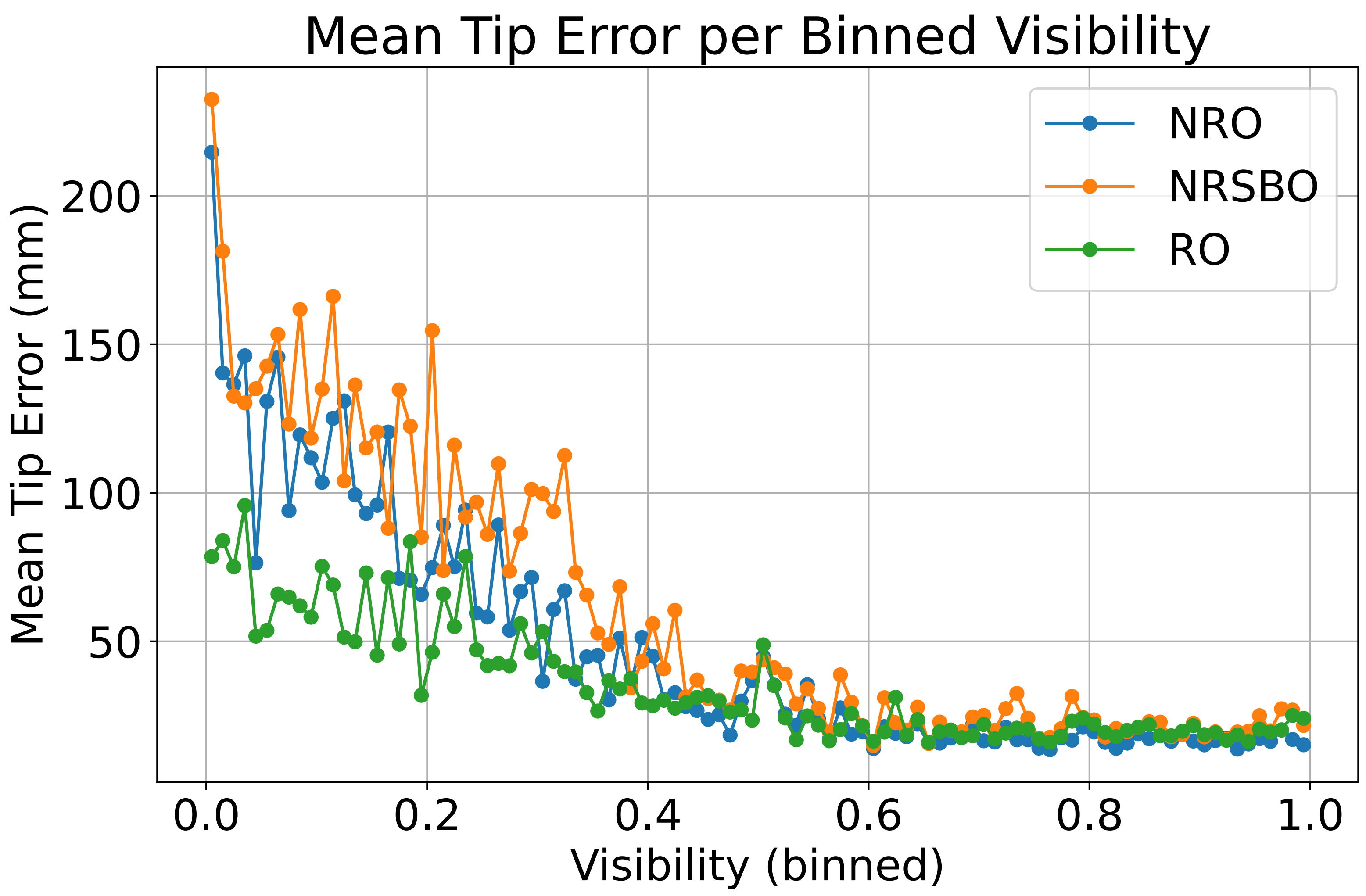
| | Views | Screwdriver | | Drill Sleeve | |
|---|---|---|---|---|---|
| | | Tip Error (mm) | Angle Error (°) | Tip Error (mm) | Angle Error (°) |
| Synthetic | 1 | 15.80 | 1.43 | 11.83 | 1.02 |
| | 2 | 2.37 | 0.47 | 1.90 | 0.47 |
| | 4 | 1.04 | 0.20 | 0.75 | 0.18 |
| | 6 | 0.86 | 0.16 | 0.57 | 0.14 |
| | 8 | 0.83 | 0.15 | 0.55 | 0.13 |
| Real | 1 | 11.50 | 1.87 | 16.05 | 2.05 |
| | 2 | 4.23 | 0.65 | 4.15 | 0.69 |
| | 4 | 2.85 | 0.44 | 2.64 | 0.53 |

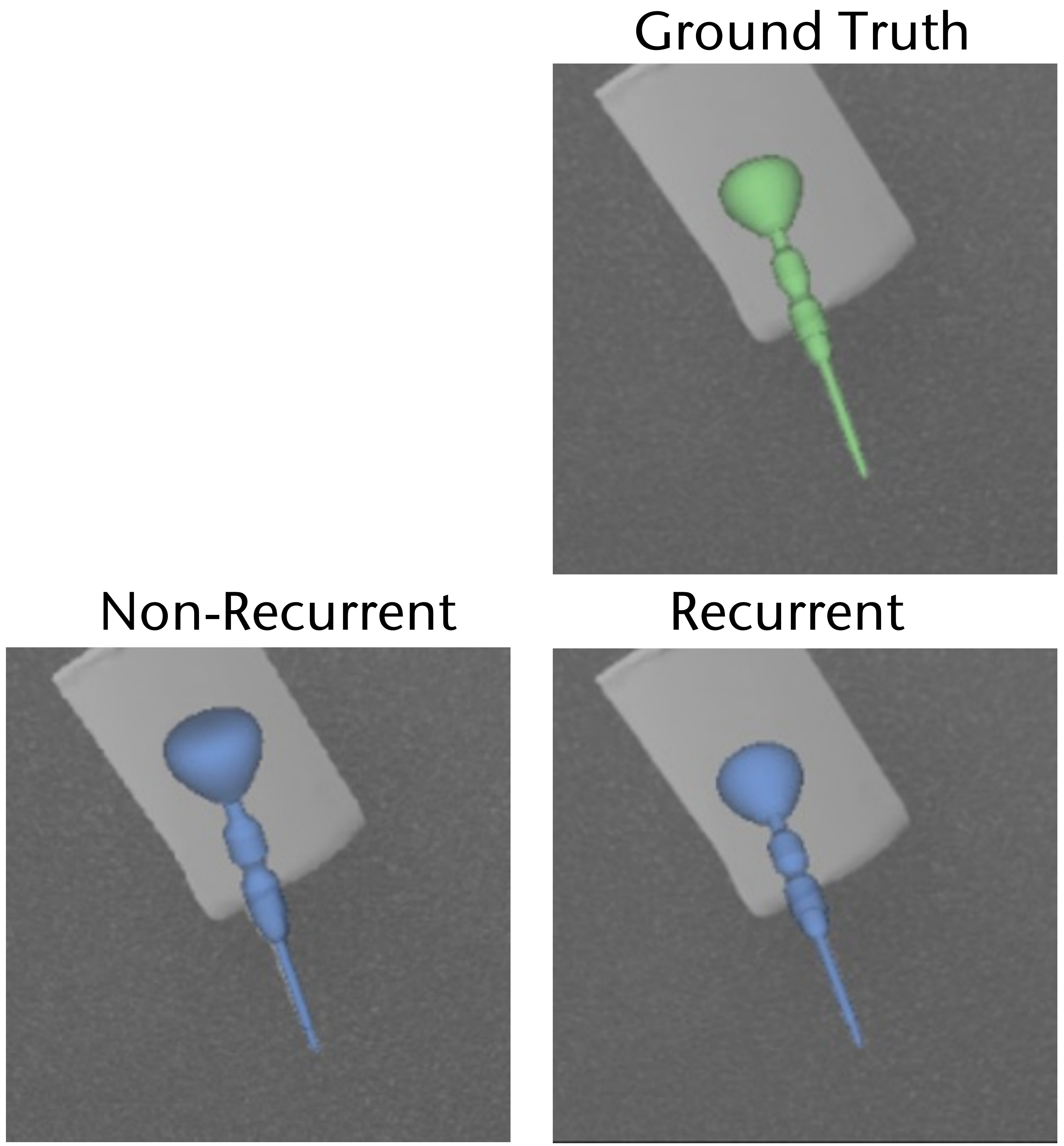
# Results - Recurrent Single-View

NRO: Non-recurrent model

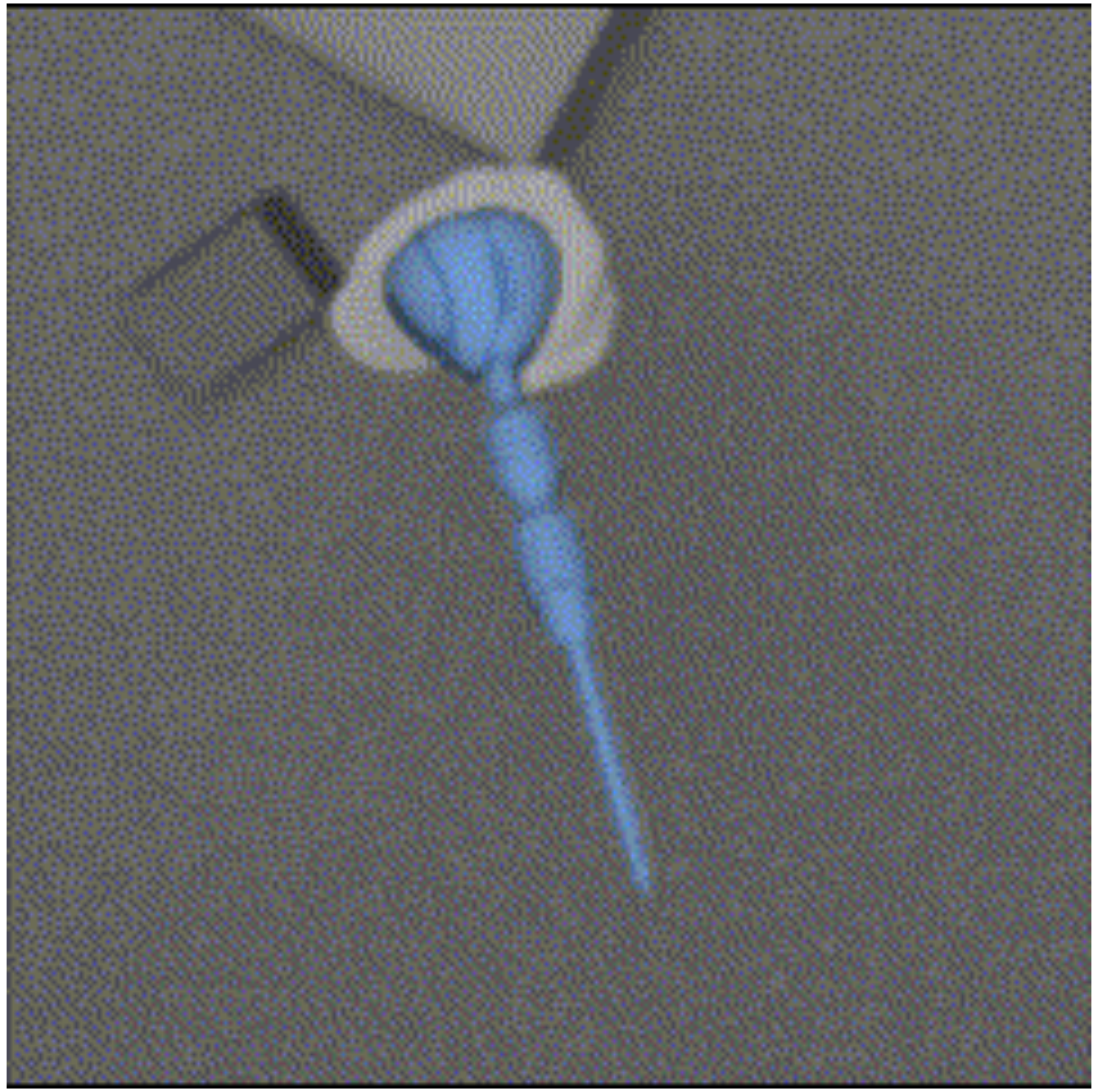
NRSBO: Non-recurrent sequential batch sampling

RO: Recurrent model

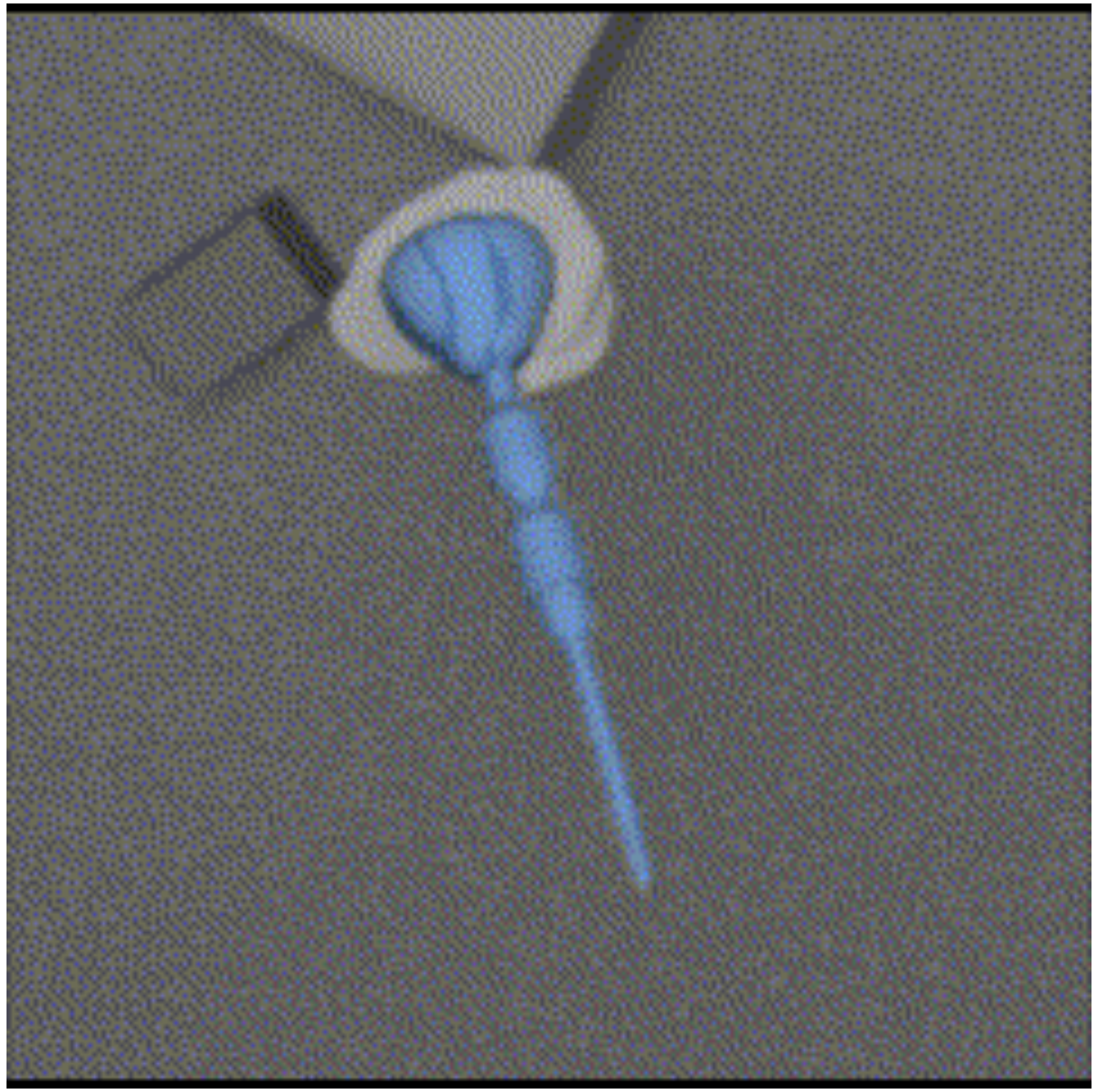


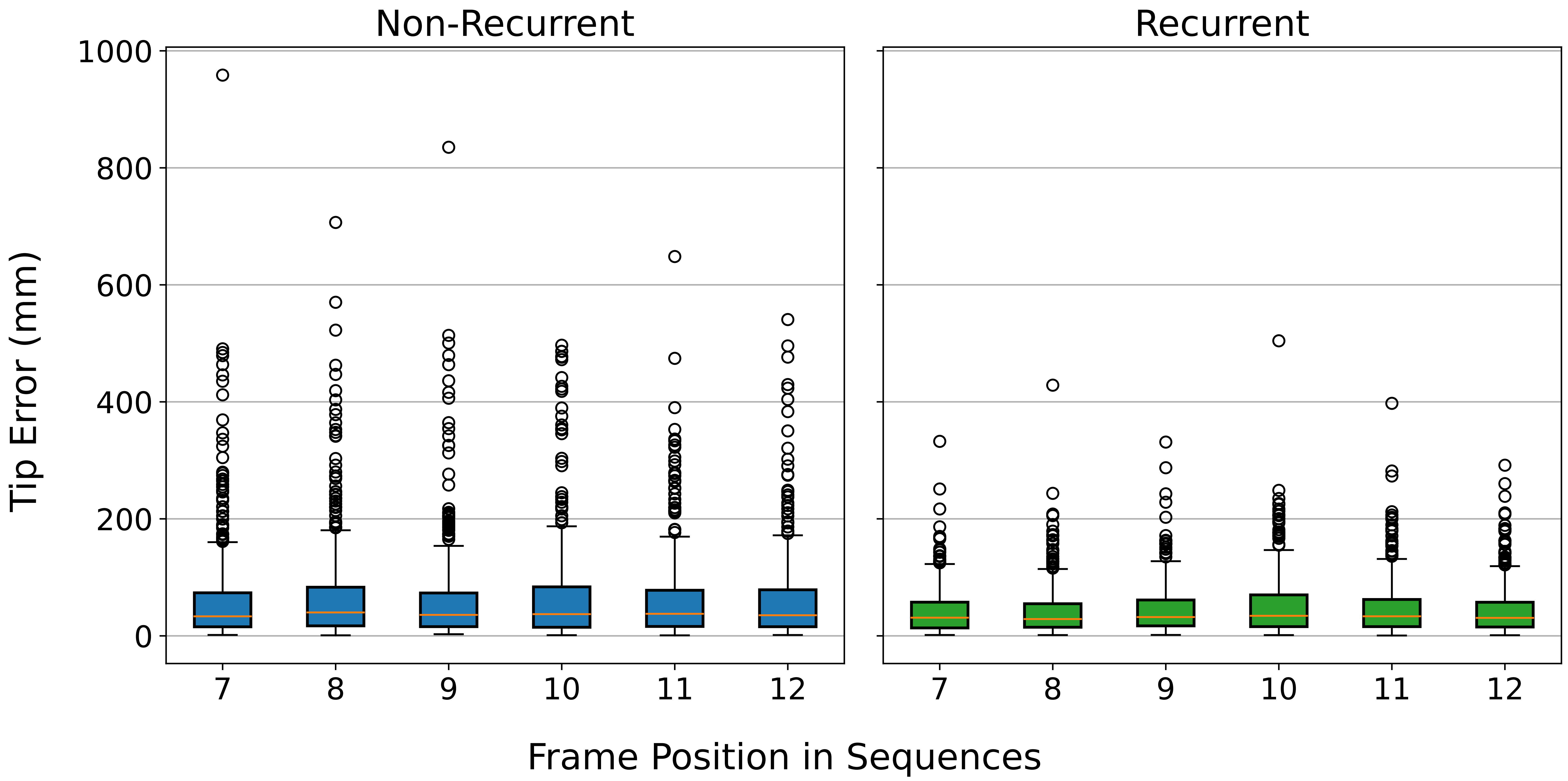



Non-Recurrent



Recurrent




Non-Recurrent
Recurrent
Tip Error (mm)
1000
800
600
400
200
0
7
8
9
10
11
12
Frame Position in Sequences

# Results: Recurrent Multi-View

- Synthetic dataset with two cameras

| | | Mean Tip Error (mm) | Mean Angle Error (degree) |
|---|---|---|---|
| Screw Driver | **NR** | 3.26 | 0.62 |
| | **R** | **3.07** | **0.59** |
| Drill Sleeve | **NR** | 2.73 | 0.55 |
| | **R** | **2.45** | **0.51** |

- Real dataset with two cameras

| | | Mean Tip Error (mm) | Mean Angle Error (degree) |
|---|---|---|---|
| Screw Driver | **NR** | 4.23 | **0.65** |
| | **R** | **3.94** | **0.65** |
| Drill Sleeve | **NR** | **4.15** | **0.69** |
| | **R** | 4.20 | 0.90 |

# Conclusion

- First recurrent multi-view architecture for 6D pose estimation
- Multi-view performance:
  - Tip error < 1 mm on synthetic data
  - Tip error < 3 mm on real-world data
- Under heavy occlusions, recurrence reduces tip error by up to 3 mm
- Achieved marker-less 6D pose estimation of surgical instruments

# Future Work

- Improve training data diversity and realism
- Improvements in network architecture
- Close synthetic to real-world gap