Detection and Localization of Robotic Tools Using Deep Neural Networks for Region Proposal and Detection

Duygu Sarikaya, Jason J. Corso and Khurshid A. Guru

Roswell Park Cancer Institute

University at Buffalo, The State University of New York

Problem

Modeling the gestures and skill level of surgeons presents an interesting problem. The insights drawn may be applied in effective skill acquisition, objective skill assessment, real-time feedback, and human-robot collaborative surgeries.

Challenges with tool detection and localization:

1. Camera movement and zoom
2. Appearance changes due to lighting and pose
3. Occlusions, deformations

Contributions

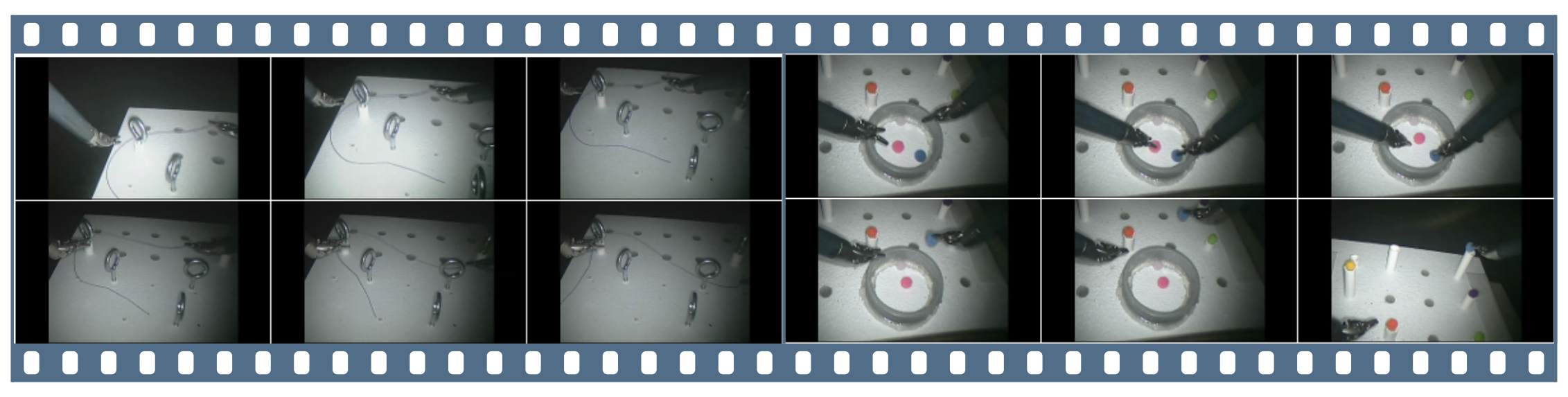
Our main contributions are

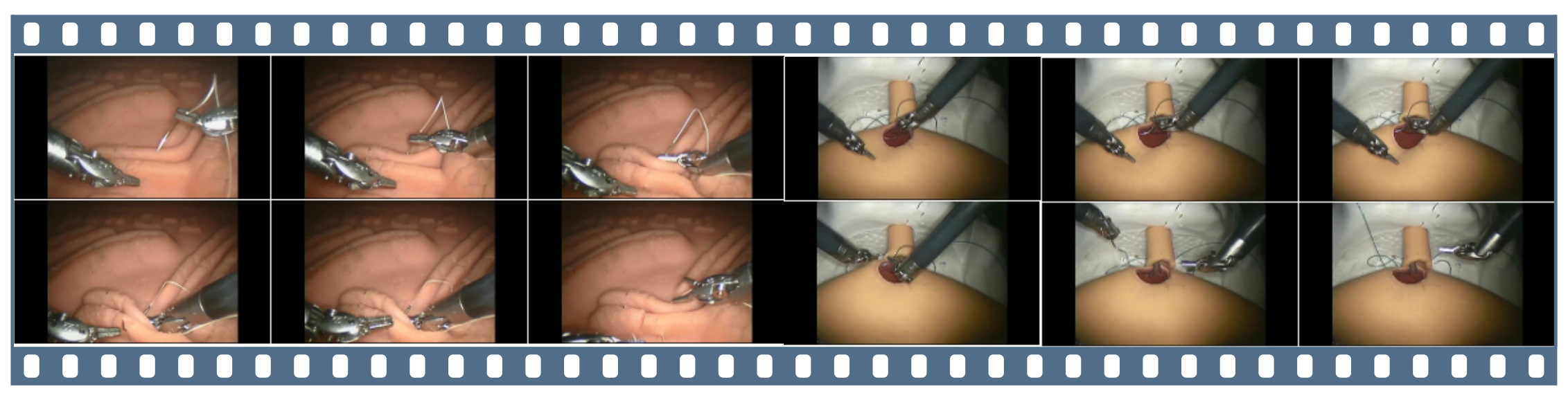
1. Strictly computer vision approach
2. No manual initialization is required
3. First to incorporate deep neural networks, adopts RPN
4. Jointly predicts objectness and localization
5. Multi-modal architecture
6. A fusion of image and temporal motion cues

Dataset

ATLAS Dione Dataset for Robot-Assisted Surgery (RAS) Video Understanding

1. IRB-approved study (I 228012)
2. 86 full subject task study videos, 910 subtask clips with a total of 5 hours
3. Subjects: Ten surgeons Roswell Park Cancer Institute (RPCI) Buffalo, NY
4. Tool annotations in VOC format
5. Timestamps of subtasks
6. Expertise levels of the subjects





Publicly available for download @

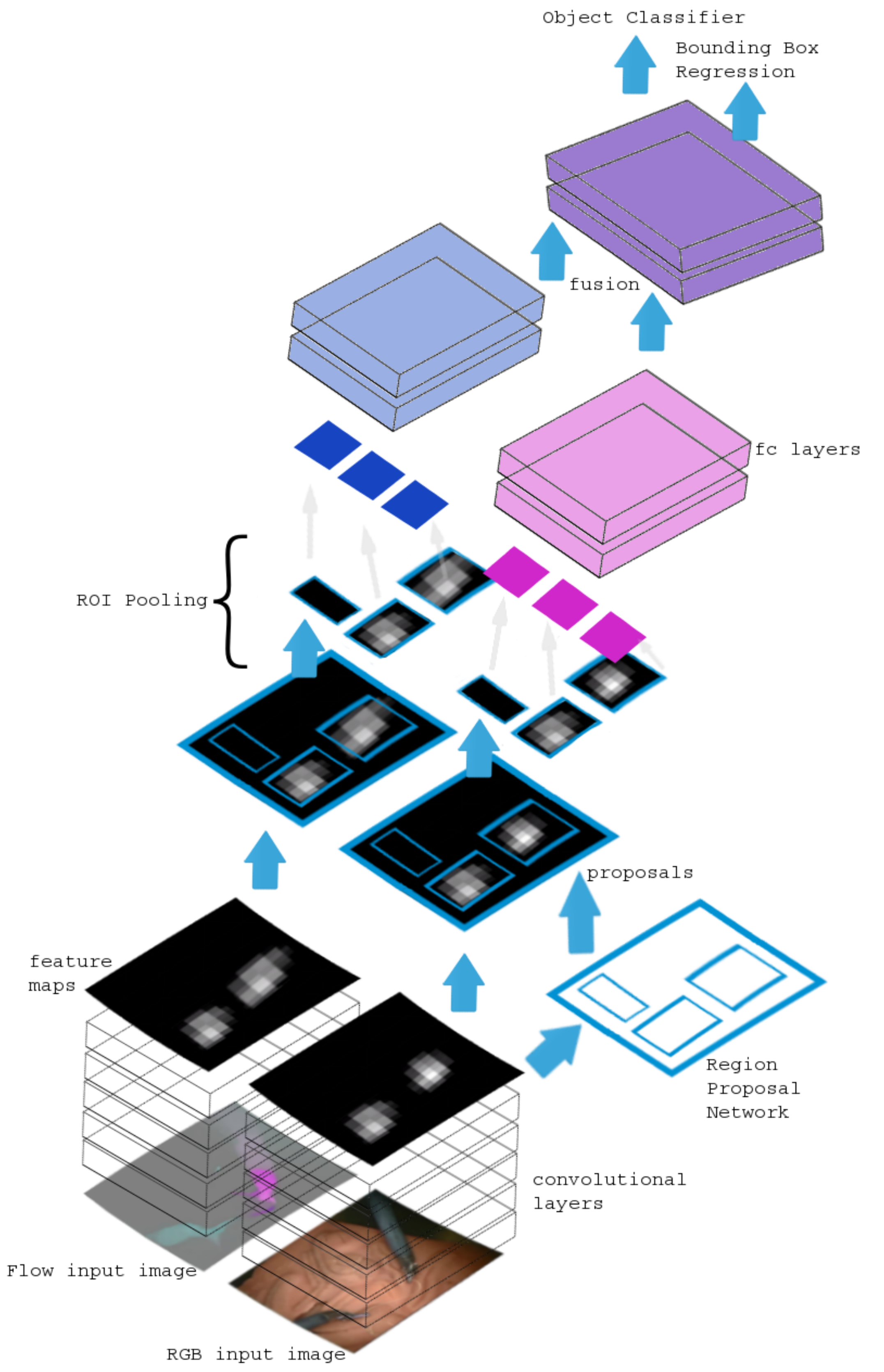
www.roswellpark.edu/education/atlas-program/dione-dataset

References

[1] M. D. Zeiler and R. Fergus Visualizing and understanding convolutional networks In *ECCV '14*

[2] S. Ren, K. He, R. Girshick, J. Sun Faster R-CNN: Towards real-time object detection with region proposal networks. In *NIPS '15*

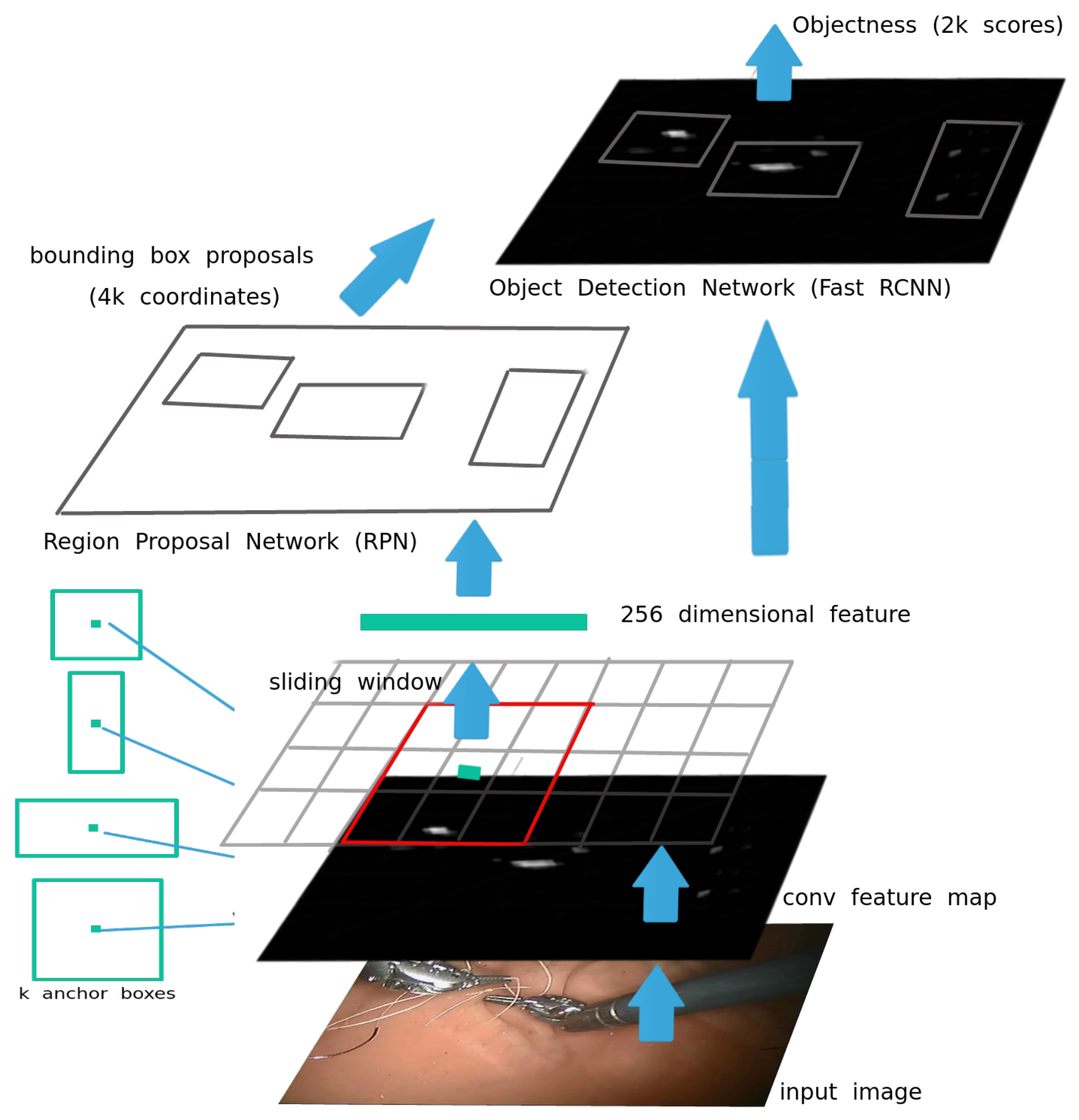
Overview of Our Model



We propose an end-to-end deep learning approach for tool detection and localization in RAS videos. Our architecture, based on the work of Zeiler *et al.* [1], has two separate CNN processing streams on two modalities: the RGB video frame and the RGB representation of the optical flow information of the same frame.

We first convolve the two separate input modalities and get their convolutional feature maps. Using the RGB image convolutional features, we train a Region Proposal Network (RPN) [2] to generate object proposals. We use the region proposals and the feature maps as input to our object classifier network. The last layer features of these streams are later fused together before being classified.

Region Proposal Network



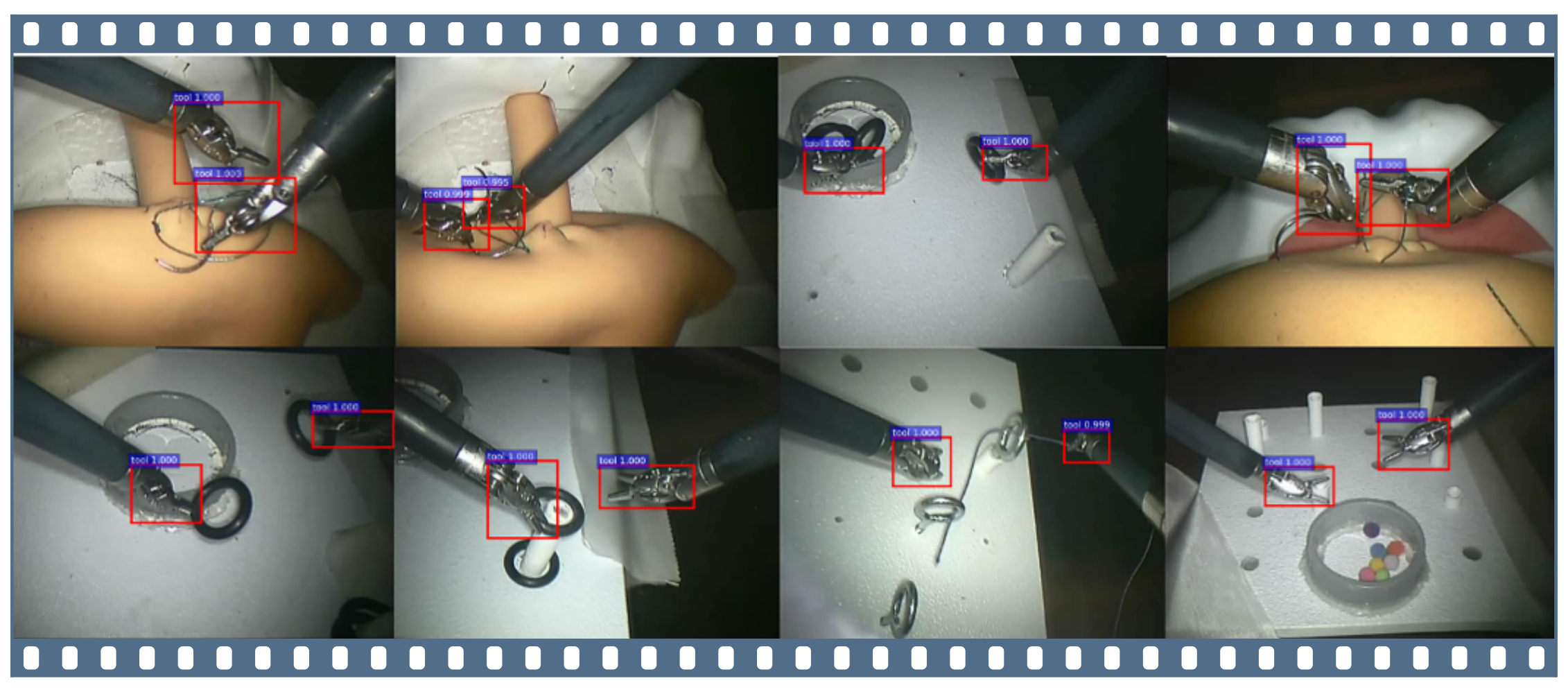
We use a Region Proposal Network that proposes regions, which are used by a detector network. We slide a small network over the convolutional feature map. Each spatial window over the feature map is mapped to a 256 dimensional feature which is fed into the system. RPN outputs object region proposals, each with an objectness score on whether the region is of a tool or not.

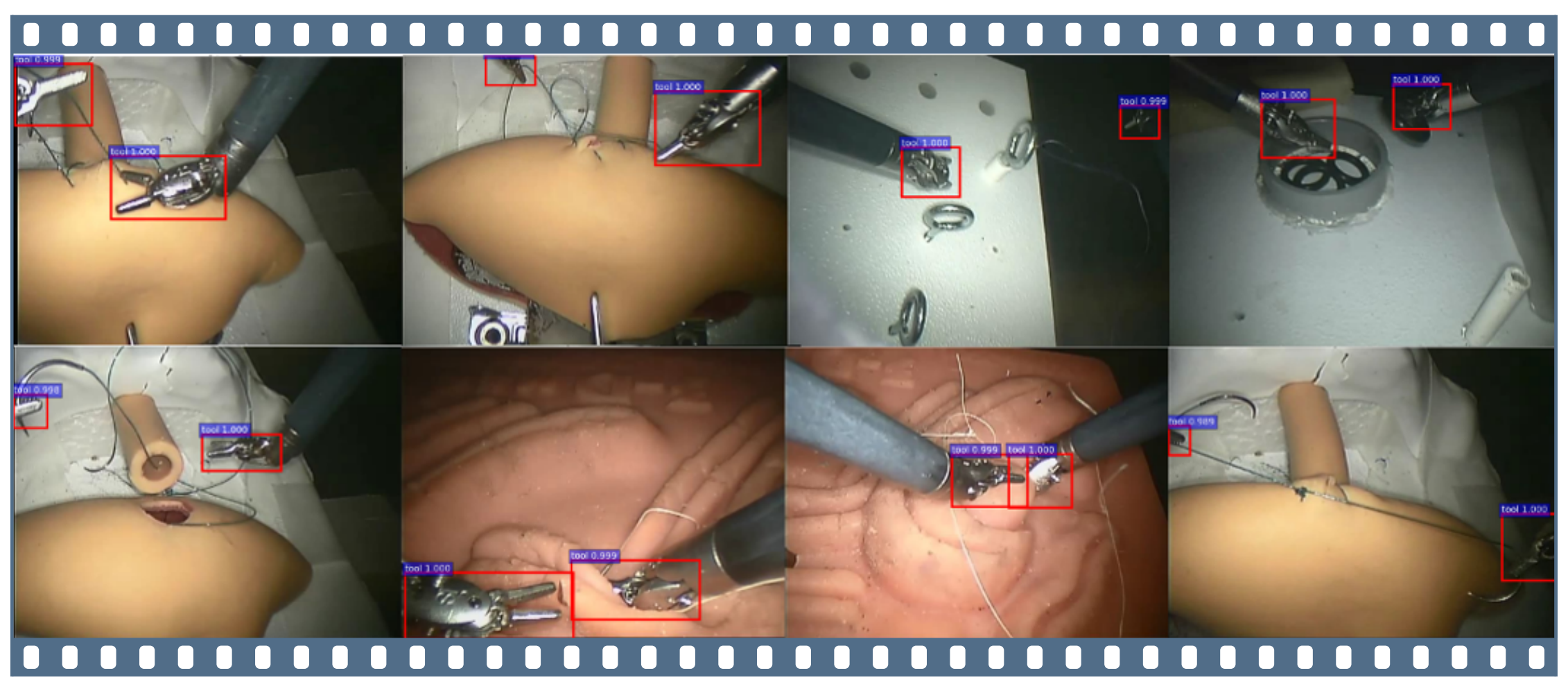
A Future Direction

Recognizing activities that reoccur across different tasks

1. Temporal dynamics
2. Long-term Short Memory Networks (Recurrent)
3. Fusion of shared and task-specific representations of hierarchical tasks

Results





Method	*Mean Average Precision*	*Detection Time (per frame)*
RPN+Fast R-CNN Multimodal	91% (90.65%)	0.103 seconds + optical flow computation (a few seconds)
RPN+Fast R-CNN (Faster R-CNN)	90% (90.39%)	0.059 seconds
Edge Boxes+Fast R-CNN (Fast R-CNN)	20%	0.134 seconds for detection + 2 seconds for region proposal
Deformable Parts Model (DPM)	76% (83% with bounding box regression)	2.3 seconds

Availability

Published in: IEEE Transactions on Medical Imaging (Volume: 36, Issue: 7, July 2017)

http://ieeexplore.ieee.org/abstract/document/7847313/