MICCAI Doctoral Consortium 2025

Doctoral Symposium

Clinical Video Analysis with Geometric Feature Enhanced Deep Learning

Francis Xiatian Zhang
Department of Computer Science, Durham University
Supervisors: Prof. Hubert P. H. Shum and Dr Noura Al Moubayed

Durham University

Motivations

- Clinical videos are vital for intervention, diagnosis, and training — but are often noisy, occluded, and visually degraded.
- Conventional deep learning relies on RGB only, which struggles with poor lighting, cluttered backgrounds, and occlusions.
- Geometric features (bounding boxes, depth maps, skeletons) provide structured spatial and motion cues that boost robustness and interpretability.
- Our goal: integrate geometry with deep learning to make clinical video analysis more accurate, reliable, and clinically meaningful.

Background

- RGB-only models: Powerful but brittle in clinical settings → fail under occlusion, smoke, variable lighting.
- Geometric features add structure:
  - Bounding boxes: capture tool–anatomy interactions.
  - Depth maps: reveal spatial layout in cluttered endoscopic views.
  - Skeletons: encode fine-grained motion and skill cues.
- Prior works explore these features in isolation.
- This research: systematic integration of geometry across three clinical video tasks → anticipation, video quality, and skill assessment.

Scientific Approach

- We integrate geometric priors into deep learning to overcome noise, occlusion, and variability in clinical videos.
- Selection Criteria:
  - Task relevance → captures structure needed (interactions, layout, motion).
  - Practical feasibility → no extra hardware or heavy annotation.
  - 2D–3D balance → rich structure with efficient computation.
- Chosen Features:
  - Bounding boxes → model surgical workflow via tool–anatomy graphs.
  - Depth maps → guide realistic endoscopic video inpainting.
  - 3D skeletons → capture motion for skill assessment (e.g., acupuncture, CPR).

Proposed Solution

- Surgical Workflow Anticipation (Bounding Boxes → Graphs)
  - Represent instruments & anatomy with bounding boxes.
  - Build dynamic interaction graphs capturing tool–tissue relationships.
  - Adaptive graph learning supports long-horizon prediction of surgical steps.

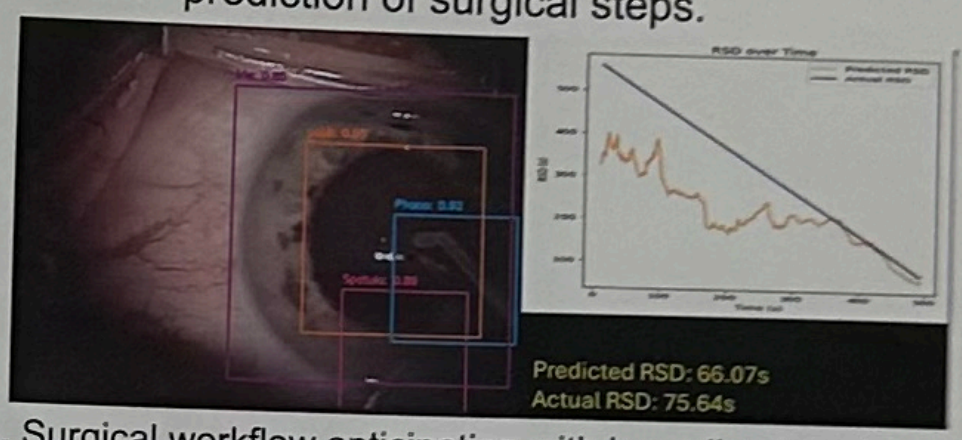

Figure 1. Surgical workflow anticipation with bounding box–based graphs.

- Endoscopic Video Inpainting (Depth Maps → Autoencoder)
  - Estimate monocular depth to provide coarse 3D structure.
  - Depth-aware autoencoder with spatial–temporal fusion and depth-guided discriminator.
  - Produces more realistic reconstructions in occluded, smoky, or cluttered scenes.

Figure 2. Depth-aware inpainting achieves more realistic reconstructions of occluded or corrupted endoscopic frames compared to prior methods.

- Clinical Skill Assessment (3D Skeletons → Multi-view Fusion)
  - Fuse video + pose features via cross-attention.
  - Multi-view alignment enables view-invariant skill evaluation.
  - Works with single-view input at inference, supporting practical training scenarios.

Figure 3. Multi-view clinical skill assessment using synchronized CPR and acupuncture datasets for robust, view-invariant evaluation.

Discussion & Long-Term Goals

Our work highlights the promise of geometric features for clinical video analysis, though challenges remain in adaptive feature selection, safe deployment, and human-in-the-loop integration. Long-term, we aim to develop geometry-aware AI that is robust, interpretable, and clinically deployable, enabling trustworthy and personalized support for medical decision-making.

Contact and More Information

Personal Website | PhD Research Page | LinkedIn
Email: francis.xiatian.zhang@outlook.outlook