Leveraging Cross-Domain Video Similarity for Fine-Tuning Surgical Models Using Pretrained Hiera

Jessica Tierney and Scott T. Acton
Mentor: Soumee Guha

Objective

Leverage pretrained action recognition vision models to classify surgical tasks such as suturing techniques.

Motivation

- Surgical analysis via machine learning requires large amounts of data to train new models.
- Limited surgical videos, especially open surgery.
- Possibility of using pretrained vision transformers while fine-tuning on a smaller surgical dataset.

One-Handed Half-Hitch (Slip) Knot

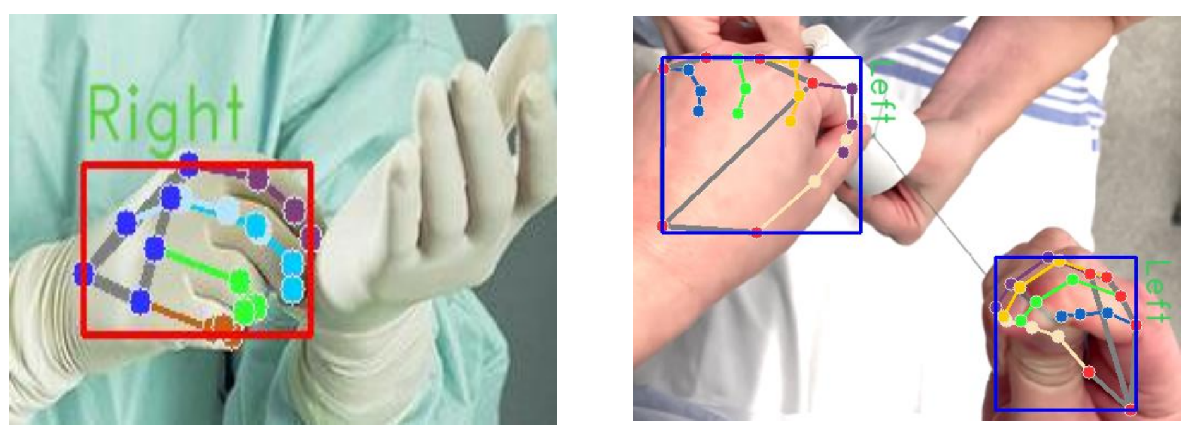
Current Approaches

MediaPipe Hand Landmark Detection

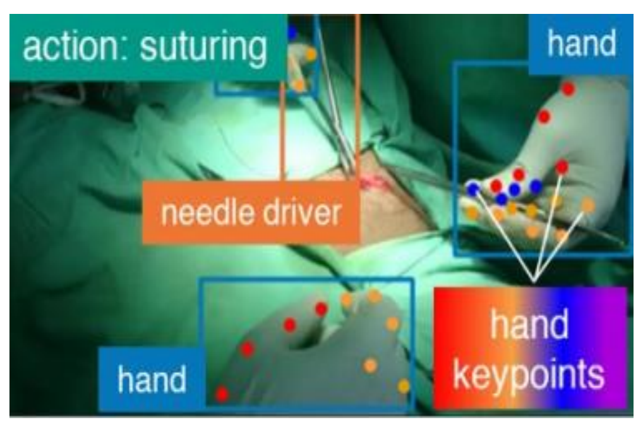
- Developed by google
- Trained on 30k real-world images.
- Detects 21 key hand-knuckle coordinates.

Annotated Videos of Surgery (AVOS)

- AVOS's dataset is already annotated with relevant United Medical Language System (UMLS) tags and spatial and temporal annotations.



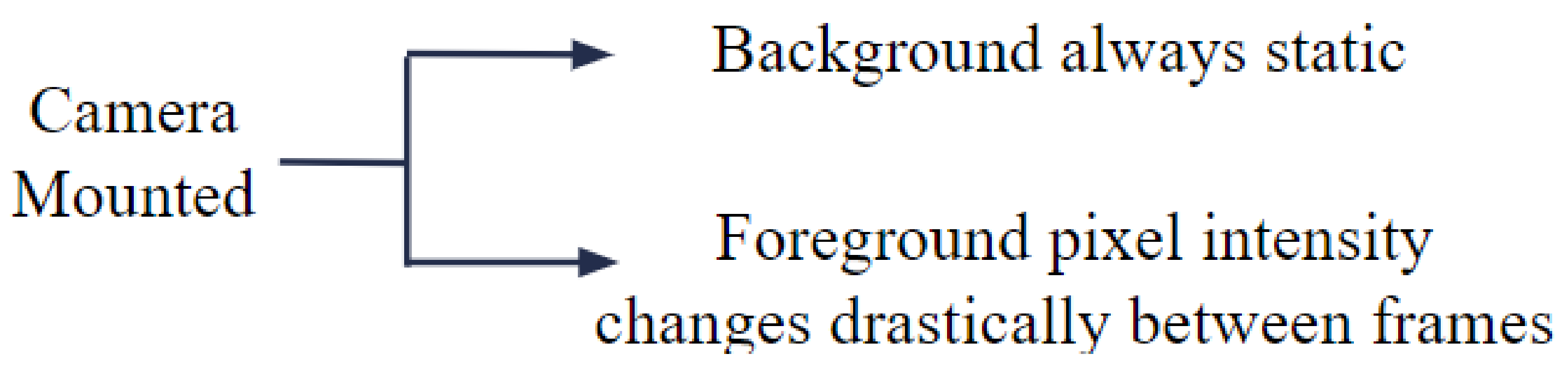
MediaPipe Failure Cases



AVOS sample frame

Preprocessing

Frame Subtraction



Camera Mounted

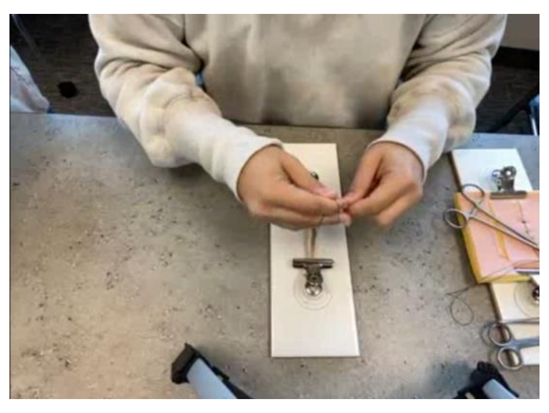
Background always static

Foreground pixel intensity changes drastically between frames

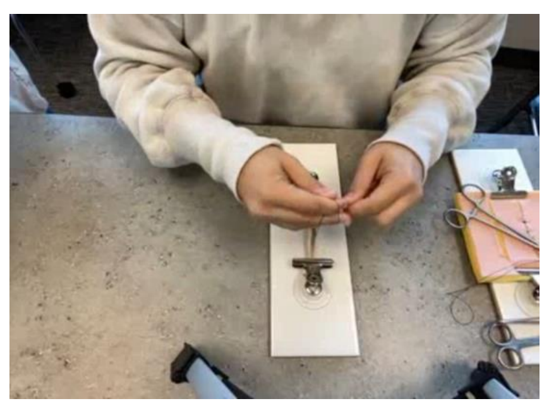
$$\Delta I_t = |I_t - I_{t-1}| \quad \forall t \in [1, T]$$

where T = total number of frames and ΔI_t is the change in pixel intensity from the current frame (I_t) to the prior (I_{t-1})

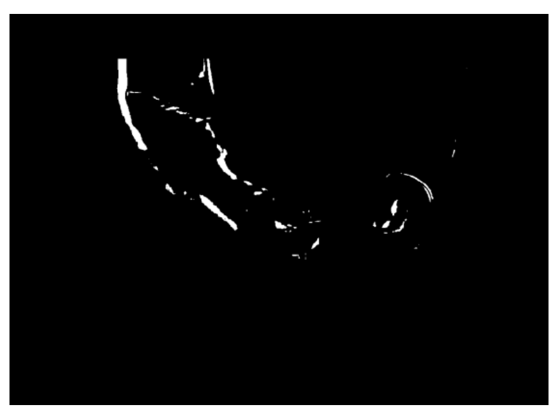
ΔI_t is higher for foreground values



I_{t-1}



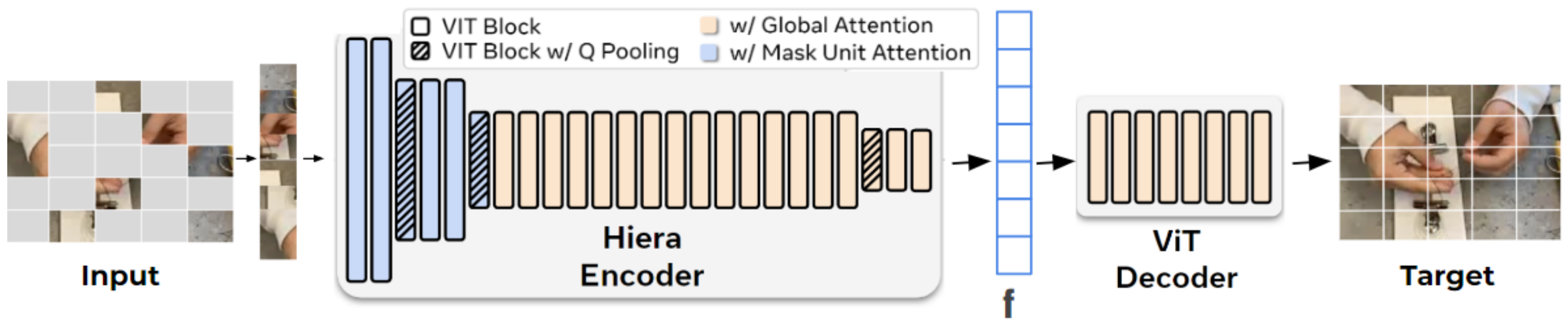
I_t



ΔI_t

Methodology

Hiera Vision Transformer



- Removes 'bells and whistles' in vision transformers.
- 2.4x faster on images and 5.1x faster on video than MViTv and is more accurate.

Cosine Similarity

$$\cos(\theta_{ij}) = \frac{f_i \cdot f_j}{|f_i||f_j|}$$

Dataset

Recorded 28 short shoe tying videos (13 tying and 15 untying), 10 short cooking clips from YouTube, and one sample One Handed Half-Hitch (Slip) Knot suturing video.

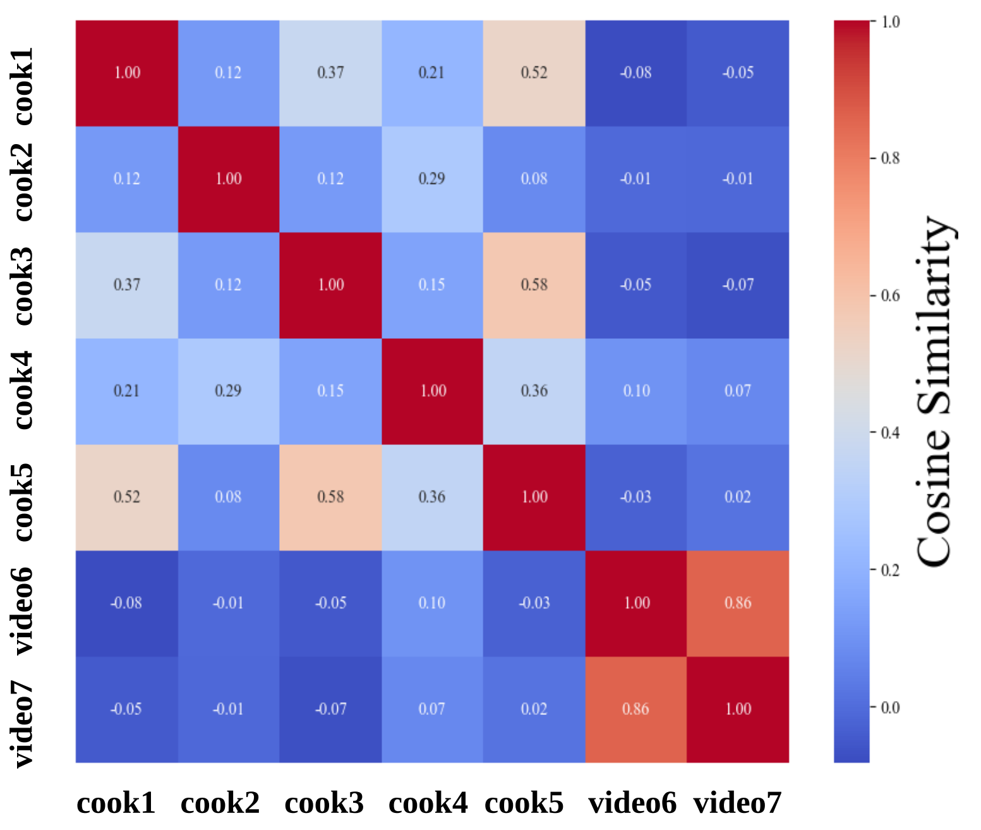
Results

Hiera Common Classifications

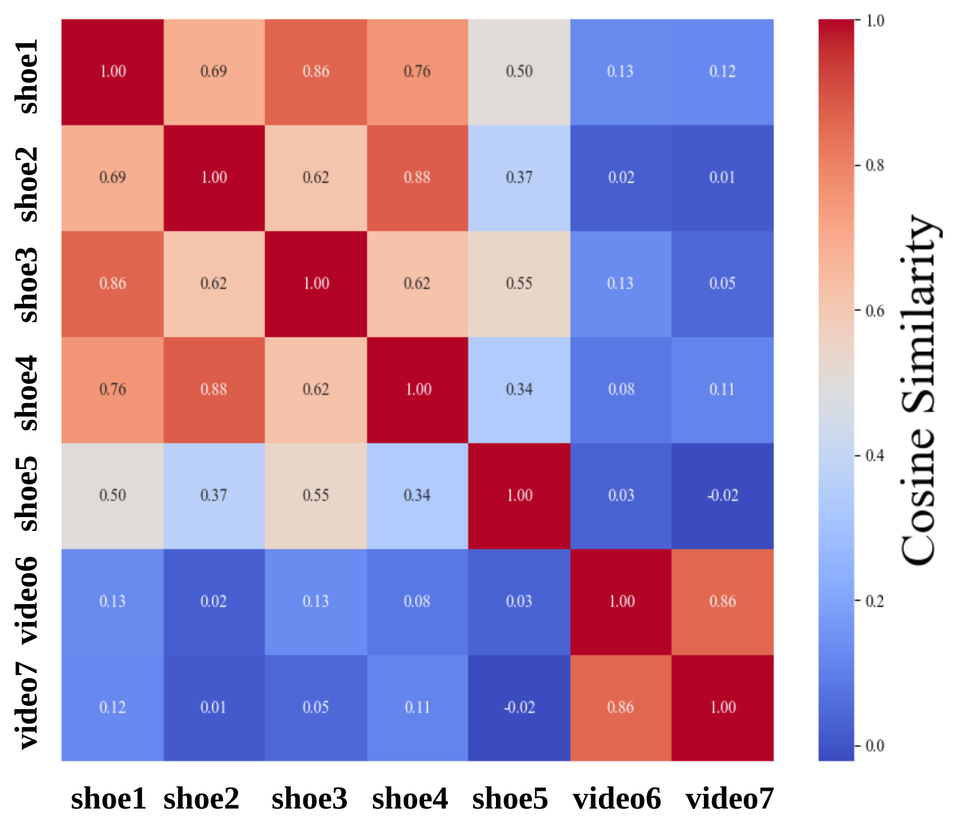
Dataset	Predictions
Shoes	Tying a knot (not on a tie), Folding laundry, Wrapping a present
Cooking	Making a sandwich, Making a cake, Cooking eggs
Suturing	Making jewelry

Cosine Similarity Results

Cooking Cosine Similarity Heatmap



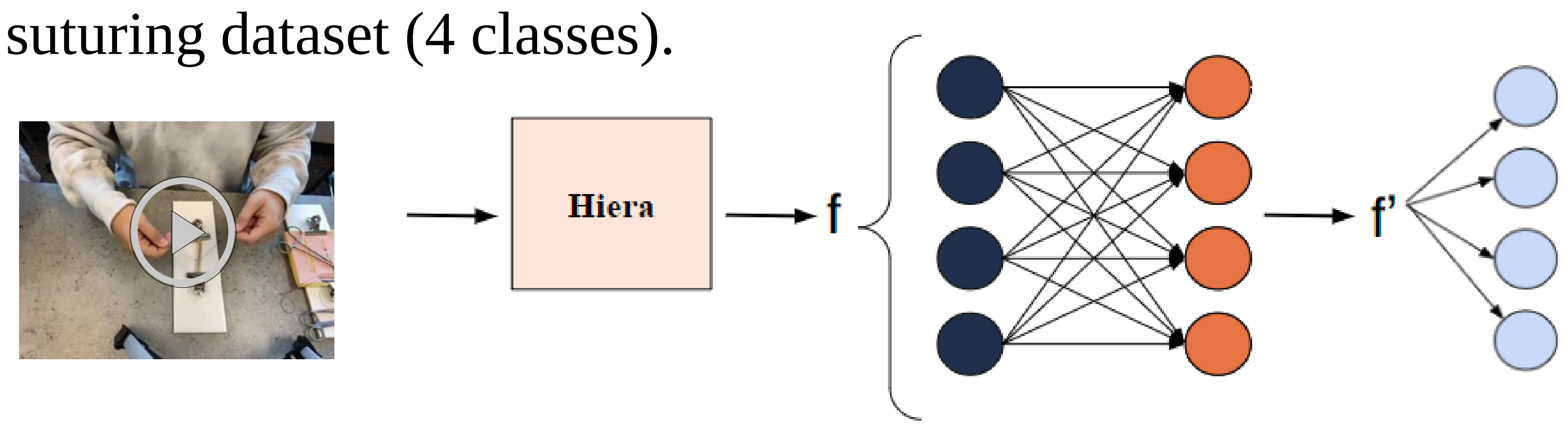
Shoes Cosine Similarity Heatmap



Heatmap: video 6 and 7 are suturing video clips.

Next Steps

Apply feature extraction and similarity analysis to the entire suturing dataset (4 classes).



Apply contrastive learning

References

1. Ryali, Chaitanya, et al. "Hiera: A hierarchical vision transformer without the bells-and-whistles." *International Conference on Machine Learning*. PMLR, 2023.
2. Goodman, Emmett D., et al. "A real-time spatiotemporal AI model analyzes skill in open surgical videos." *arXiv preprint arXiv:2112.07219* (2021).
3. Zhang, Fan, et al. "Mediapipe hands: On-device real-time hand tracking." *arXiv preprint arXiv:2006.10214* (2020).

UNIVERSITY of VIRGINIA
BIOCOMPLEXITY INSTITUTE