ACS 2025 Surgeons and Engineers: A Dialogue on Surgical Simulation Meeting

O-08

Research Abstracts

Foundational Model for Cataract Surgical Scene Understanding

Nisarg Anish Shah; Chaminda Bandara; Shameema Sikder, MD, FACS; S. Swaroop Vedula, MBBS, PhD, MPH; and Vishal Patel

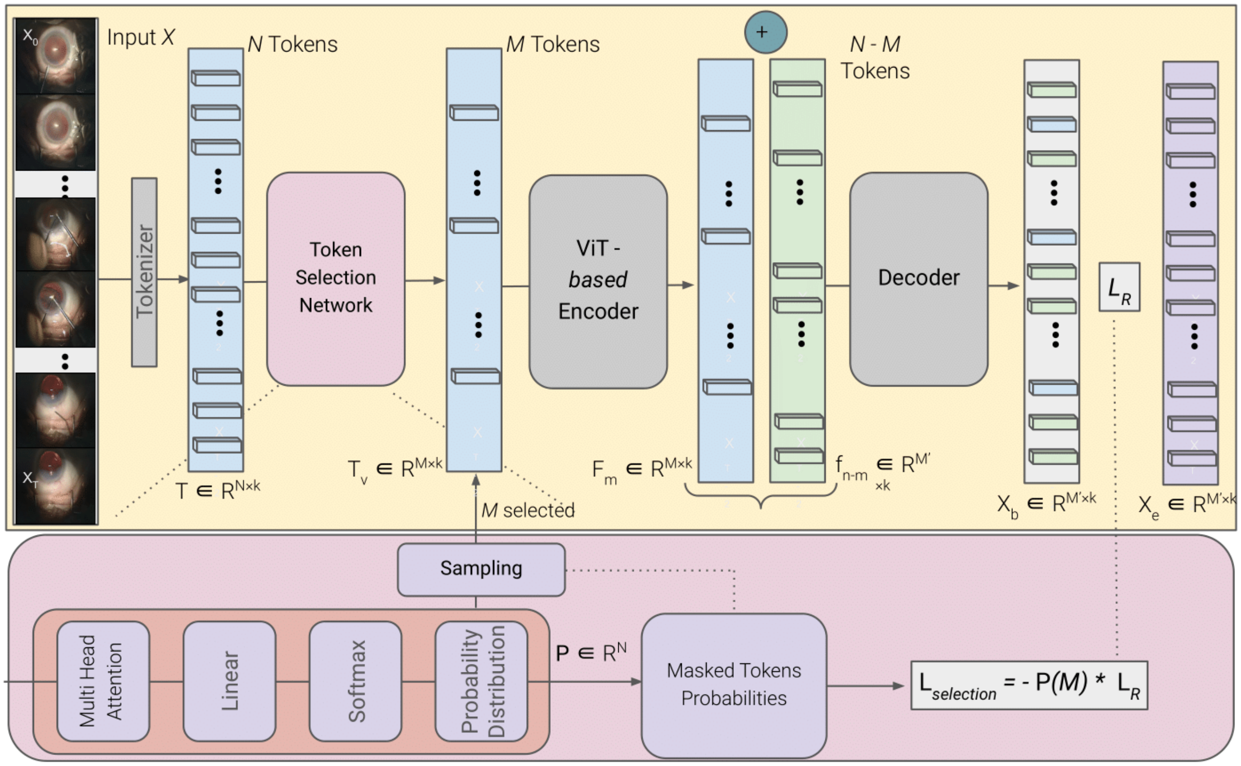
Johns Hopkins University, Baltimore, MD

Introduction: The foundation model facilitates automated analysis of cataract surgery videos, crucial for standardizing surgical training and improving postoperative outcomes. This research addresses the challenge of variability in surgical techniques and the scarcity of annotated data. We introduce a Masked Autoencoder (MAE)--based pre-training methodology for cataract surgeries, enabling learning from a comprehensive dataset of surgery videos through a novel token masking strategy. Our model excels in low-data regimes, setting new standards in step recognition and surgical skill assessment.

Methods: Our MAE model comprises four key components: the Tokenizer, Encoder, Decoder, and Token Selection Network. The Tokenizer processes input video frames into tokens, which are then selectively masked based on their informational value, calculated using a lightweight multi-head attention network. This selection process is guided by a categorical distribution over the tokens' probability scores. The visible tokens are encoded using a vision transformer, and their representations are then combined with placeholders for the masked tokens. The combined data is decoded by a lightweight transformer to reconstruct video frames. We optimize using Mean Squared Error loss between reconstructed and actual frames, focusing on masked regions. This efficient setup allows significant computational savings, crucial for managing video datasets without extensive manual annotations.

Results: Evaluating MAE on Cataract-101 and D99 datasets for step recognition and skill assessment showed significant improvements. Achieving a Jaccard index up to 10% higher and 15% better precision on D99, MAE excelled in low-data scenarios. In skill assessment, it improved accuracy and sensitivity by over 5-10%. These results highlight MAE's robust performance in managing surgical video variability and low-data regimes.

Conclusions: We introduced an MAE-based pre-training method for cataract videos, achieving superior feature representation with a 95\% masking ratio. Demonstrating robust performance in low-data regimes, it significantly outperforms existing methods in clinical applications.


x_0
x_T
Input X
Tokenizer
N Tokens
T ∈ R^{N×k}
Token Selection Network
M Tokens
T_v ∈ R^{M×k}
M selected
ViT - based Encoder
+
N - M Tokens
F_m ∈ R^{M×k}
f_{n-m} ∈ R^{M'×k}
Decoder
L_R
X_b ∈ R^{M'×k}
X_e ∈ R^{M'×k}
Sampling
Multi Head Attention
Linear
Softmax
Probability Distribution
P ∈ R^N
Masked Tokens Probabilities
L_selection = - P(M) * L_R