Multi-modal Representations for Fine-grained Multi-label Critical View of Safety Recognition

Britty Baby[1,4], Vinkle Srivastav[1,4], Pooja P. Jain[1], Kun Yuan[1,3], Pietro Mascagni[2,4], and Nicolas Padoy[1,4]

[1] University of Strasbourg, CNRS, INSERM, ICube, UMR7357, Strasbourg, France
[2] Fondazione Policlinico Universitario A. Gemelli IRCCS, Università Cattolica del Sacro Cuore, Rome, Italy
[3] CAMP, Technische Universität München, Munich, Germany
[4] Institute of Image-Guided Surgery, IHU Strasbourg, Strasbourg, France

Abstract

The Critical View of Safety (CVS) is crucial for safe laparoscopic cholecystectomy, yet assessing CVS criteria remains a complex and challenging task, even for experts. Traditional models for CVS recognition depend on vision-only models learning with costly, labor-intensive spatial annotations. This study investigates how text can be harnessed as a powerful tool for both training and inference in multi-modal surgical foundation models to automate CVS recognition. Unlike many existing multi-modal models, which are primarily adapted for multi-class classification, CVS recognition requires a multi-label framework. Zeroshot evaluation of existing multi-modal surgical models shows a significant performance gap for this task. To address this, we propose CVSAdaptNet, a multi-label adaptation strategy that enhances fine-grained, binary classification across multiple labels by aligning image embeddings with textual descriptions of each CVS criterion using positive and negative prompts. By adapting PeskaVLP, a state-of-the-art surgical foundation model, on the Endoscapes-CVS201 dataset, CVS-AdaptNet achieves 57.6 mAP, improving over the ResNet50 image-only baseline (51.5 mAP) by 6 points. Our results show that CVS-AdaptNet's multi-label, multimodal framework, enhanced by textual prompts, boosts CVS recognition over image-only methods. We also propose text-specific inference methods, that helps in analysing the image-text alignment. While further work is needed to match state-of-the-art spatial annotation-based methods, this approach highlights the potential of adapting generalist models to specialized surgical tasks.

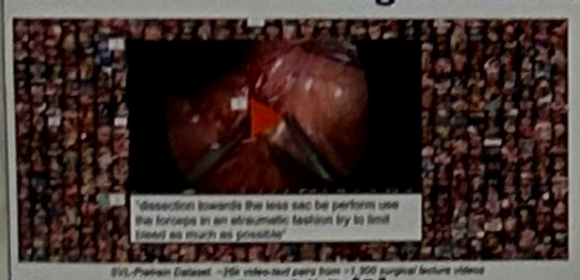
Background

- **Critical View of Safety (CVS)** vital to prevent bile duct injuries
- **Challenges:**
  Visually ambiguous images. Low inter-annotator agreement (Cohen's kappa = 0.38)

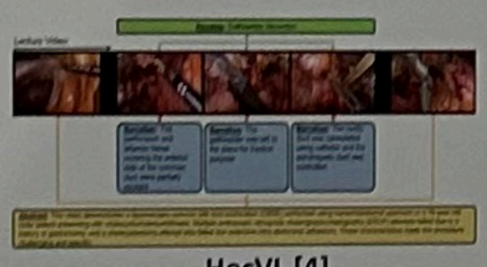
Motivation

- **Limitations of Current Methods**
  - Vision-only models rely on costly, labor-intensive spatial annotations
  - *SurgLatentGraph [1]* drops 8 mAP when cystic duct is removed
- **Multi-modal models** (CLIP [2], SurgVLP [3], HecVL [4], PeskaVLP [5]) offer new potential
  - Can text help in fine-grained surgical recognition? What advantages does text bring?
  - CVS assessment is a multi-label, fine-grained, specialized task. CLIP-like contrastive learning suits multi-class, ***how to adapt?***

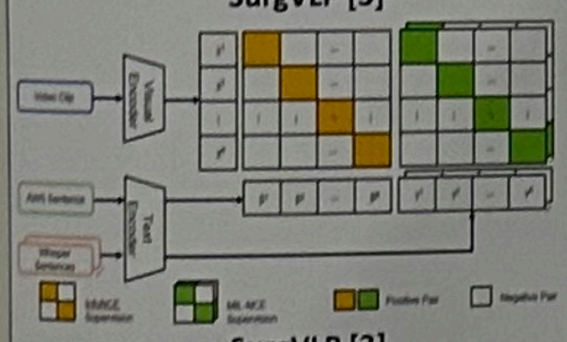
Surgical Foundation Models



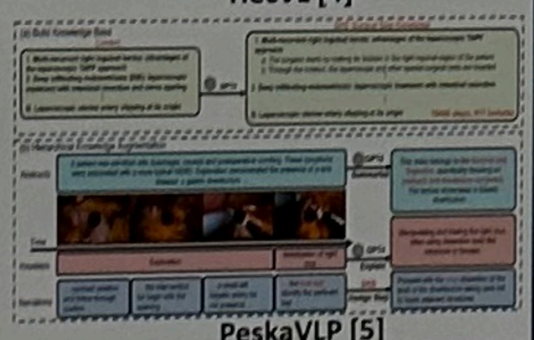
SurgVLP [3]



HecVL [4]

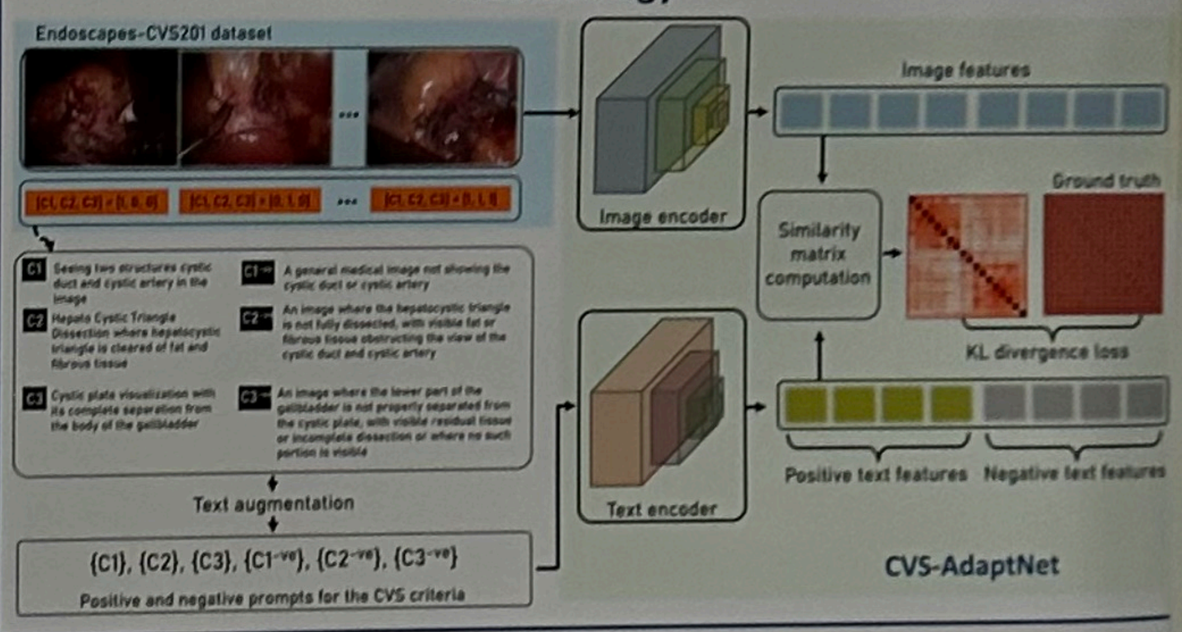


SurgVLP [3]



PeskaVLP [5]

Methodology



Training

- To model the inherent annotator ambiguity and variability in CVS labels, we use **Kullback-Leibler (KL) divergence** as a contrastive loss.
- It accommodates many-to-many matches across prompts and images
- Discriminative separation between matching and non-matching pairs

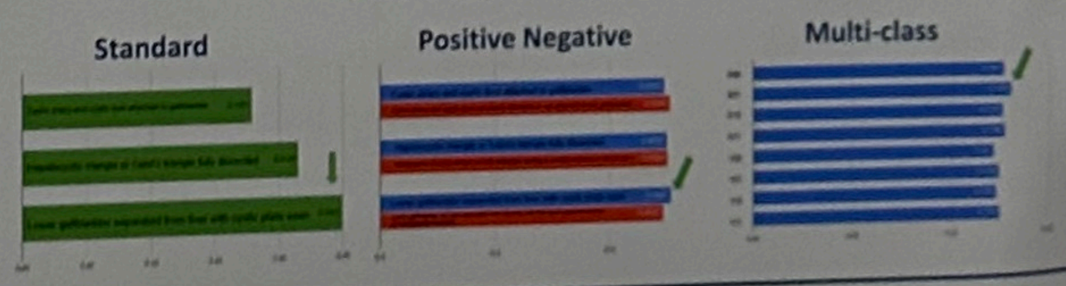
Inference

- **For a given test image:** We propose 3 types of inferences to enhance robustness.
  - ***Standard:*** Cosine similarity to a fixed set of clinician-selected prompts
  - ***Positive- Negative:*** To test the ability to select positive prompt
  - ***Multi-class:*** Prompts combining the aspects of the original criteria

Results

GT: [0,0,1]

Denotes predicted prompt



Conclusion

- Text-augmented, multi-modal models show promise for specialized surgical tasks like CVS recognition.
- Annotation-efficient approach: Leverages natural language prompts and criteria descriptions without extra labeling.

Ablations

- **Random text** → Misaligned text hurts performance (worst results)
- **Generic surgical text** → Slight improvement, still weak.
- **Fixed class prompts** → Class-aligned text helps.
- **Detailed anatomical text** → Too fine-grained, doesn't always improve.
- **Medium-detailed text** → Best (mAP 57.54)

References

1. Murali, A., Alapatt, D., Mascagni, P., Vardazaryan, A., Garcia, A., et al.: Latent graph representations for critical view of safety assessment. TMI (2023)
2. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., et al.: Learning transferable visual models from natural language supervision. ICML (2021)
3. Yuan, K., Srivastav, V., Yu, T., Lavanchy, J.L., Marescaux, J., Mascagni, P., Navab, N., Padoy, N.: Learning multi-modal representations by watching hundreds of surgical video lectures. MedIA (2025)
4. Yuan, K., Srivastav, V., Navab, N., Padoy, N.: Hecvl: Hierarchical video-language pretraining for zero-shot surgical phase recognition. MICCAI (2024)
5. Yuan, K., Navab, N., Padoy, N., et al.: Procedure-aware surgical video-language pretraining with hierarchical knowledge augmentation. NeurIPS (2024)