Supplementary Materials for "OphNet: A Large-Scale Video Benchmark for Ophthalmic Surgical Workflow Understanding"

Ming Hu[1,2,3*], Peng Xia[1,3*], Lin Wang[5*], Siyuan Yan[1,2], Feilong Tang[1,3]
Zhongxing Xu[7], Yimin Luo[6], Kaimin Song[3], Jurgen Leitner[2]
Xuelian Cheng[1], Jun Cheng[8], Chi Liu[9], Kaijing Zhou[4†], and Zongyuan Ge[1,2,3†]

[1]AIM Lab, Faculty of IT, [2]Faculty of Engineering, Monash University
[3]Airdoc-Monash Research, Airdoc [4]Eye Hospital, Wenzhou Medical University
[5]Bosch Corporate Research [6]King's College London [7]Cornell University
[8]Institute for Infocomm Research, A*STAR
[9]Faculty of Data Science, City University of Macau

1 Training Details

We conducted all experiments using 4 NVIDIA RTX3090Ti GPUs.

1.1 Training Details for Classification Tasks

We employed the officially released codes to train all recognition models. The SlowFast [3], I3D [1] and X3D [2]models were trained for 150 epochs with a batch size of 16, using a base learning rate of 0.001. We employed a cosine decay learning rate scheduler with 34 warmup epochs. We sampled 16 frames per clip with a sampling rate of 16. For the configuration of training MViT v2 [6] model, we apply the base learning 0.0001, cosine decay learning rate scheduler, 200 training epochs, 30 warmup epochs, and the batch size 8. We sample 16 frames per clip with the sampling rate of 16.

1.2 Training Details for Localization Task

Feature extraction. We firstly extract the frames from each video with 25 FPS and also extract the optical flow with TV-L1 [4, 8] algorithm. After that, we finetune an I3D [1] model on Kinetics 400 [5], and then use it to generate the features for each RGB and optical flow frame. Since each video has variable duration, we perform the uniform interpolation to generate 100 fixed-length features for each video. Finally, we concatenate the RGB and optical flow features into a 2048-dimensional embedding as the model input.
Model training. We train all the detection models with their officially released code and the default configurations. For training ActionFormer [11] model, we apply the base learning rate 0.001, cosine decay learning rate scheduler, 30 training epochs, 5 warmup epochs, and the batch size 16. For training TriDet [9]

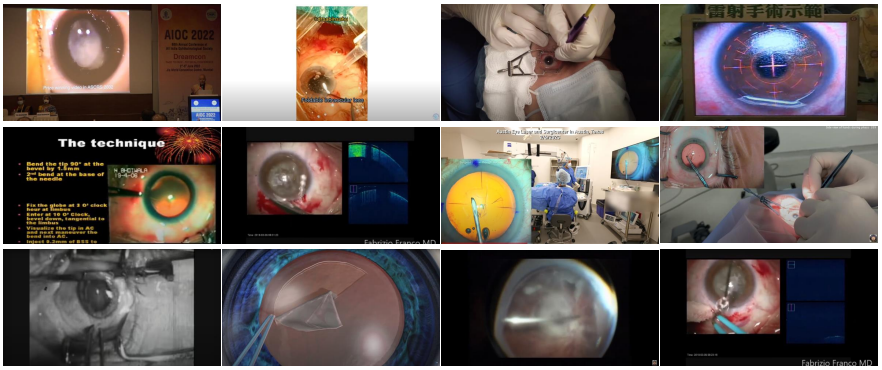


Fig. 1: Examples of filtered videos.

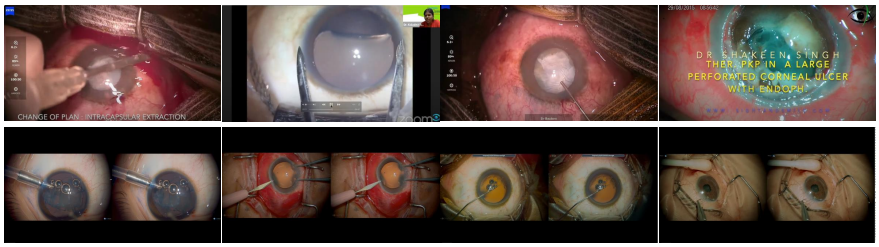


Fig. 2: Examples with minor flaws that were still retained.

model, we apply the base learning rate of 0.0004, step decay learning rate scheduler, 20 training epochs, and the batch size 200. For these two baseline models, we employed three different backbone network settings for performance comparison: CSN [10], SwinViviT [7], and SlowFast [3].

1.3 Training Details for Anticipation Task

We follow the same settings as used in classification experiment.

2 Annotation Interface Demonstration

2.1 Video Filtering

The videos in the OphNet dataset, sourced from YouTube, exhibit a variety of styles, resolutions, and on-screen elements. To ensure quality and relevance, we filtered out videos that do not provide a microscopic perspective (first row of Fig. 1), as well as those with subtitles, additional video windows, or watermarks occupying a significant portion of the frame (second row of Fig. 1). Furthermore, videos depicting unrealistic animations, suffering from poor resolution, displaying grayscale images, or containing OCT imagery (third row of Fig. 1) were

also excluded. However, we retained videos with minimal on-screen text or watermarks (first row of Fig. 2). Additionally, 3D videos recorded using binocular microscopes were preserved, albeit processed to retain only the left-eye perspective in our dataset.

2.2 Classification Annotation Interface

In this stage, we categorize the videos into valid and invalid videos through keypresses, with valid videos further classified based on their primary surgical type. Initially, an attending ophthalmologist categorizes the videos into three types: cataract surgery, glaucoma surgery, and corneal surgery. These are then further distributed for filtering and classification annotation, with each individual responsible for one of the three major surgeries.

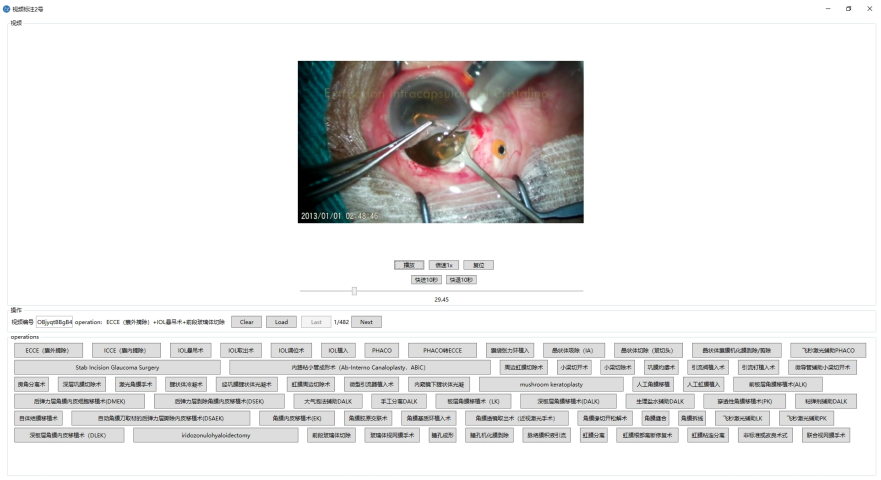
2.3 Hierarchical Localization Annotation Interface

We have designed an interface that supports three levels of annotation: surgery, phase, and operation, and is easy to operate and modify later. The main window plays the video (with features such as speed adjustment, fast forward, rewind, and pause), while the left and right sub-windows display the corresponding frames for the start and end times of the current annotated segment. Additionally, it supports functions such as automatic time positioning and instance insertion.

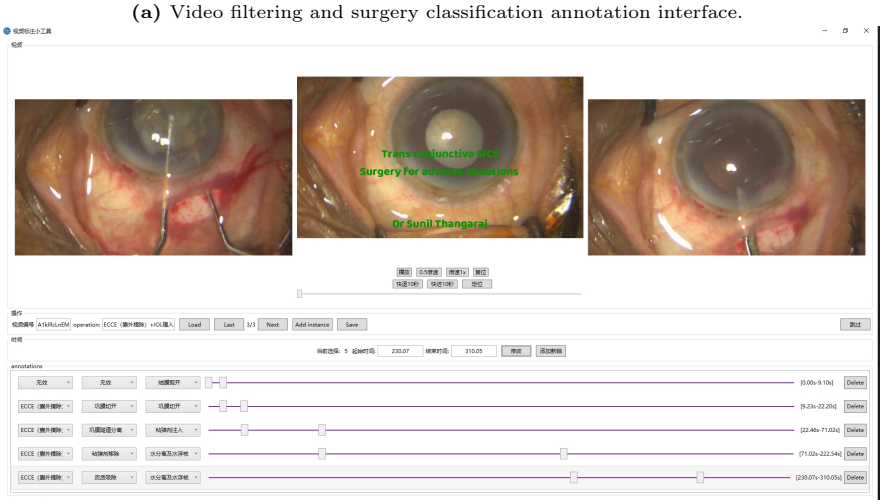
3 Dataset Bias

Dataset Bias. OphNet's videos are sourced from YouTube and exhibit diverse styles, clarity, and screen elements. This diversity can aid detection models in generalization but may affect their effectiveness and performance. Some videos in the dataset include subtitles or additional video windows, such as watermark shown in Fig. 1. Similarly, additional video windows offer another perspective but can make the scene chaotic, making it harder to recognize primary surgical actions. The presence of these factors in OphNet reflects the complexity of real-world surgical environments, because an ophthalmic microscope may inherently display different windows or show parameters during recording. While they pose challenges, they also present opportunities for developing models that can better handle variability and unpredictability, which are crucial aspects of real-world surgical scenarios.

Annotation Bias. OphNet is entirely annotated by ophthalmologists, and while this ensures a high level of expertise, it also introduces the possibility of annotation bias reflecting specific regional practices, terminologies, and interpretations. Despite the universal nature of many ophthalmic procedures, subtle differences in surgical techniques, procedural preferences, and clinical terminologies could lead to inconsistencies in how surgeries are categorized and described across different regions. For instance, the technique for cataract extraction may vary



(a) Video filtering and surgery classification annotation interface.



(b) Hierarchical temporal localization annotation interface for surgery, phase, and operation

Fig. 3: Annotation Interface Design

between phacoemulsification in one region and manual small incision cataract surgery in another, leading to differences in the annotation of surgical phases and operations. Similarly, the terminology used to describe certain procedures might differ, with one region referring to a procedure as "anterior vitrectomy" while another uses "pars plana vitrectomy." To reduce the possibility of biases in precise annotations, we have taken great care to establish a unified definition prior to describing the surgery, phase, and operation. However, potential biases arising from regional variations and individual surgical practices are inevitable. Recognizing these potential biases is crucial for the users of OphNet, as it allows for a more nuanced interpretation of the data and its applicability to different clinical settings. Future work could involve expanding the annotation team to include ophthalmologists from diverse geographical regions and surgical back-

grounds, further mitigating the impact of regional and individual biases on the dataset.

4 OphNet's Extension

Multi-Surgery Recognition. In the realm of surgical procedures, obtaining large-scale, finely annotated video datasets is a formidable challenge due to privacy concerns, the extensive time required for detailed labeling by medical experts, and the complexity of surgical actions. Consequently, weak supervision emerges as a pivotal approach, enabling the utilization of limited or imprecise labels to train robust models capable of understanding and recognizing diverse surgical activities. Looking forward, the integration of domain knowledge, such as surgical ontologies and procedural guidelines, into learning frameworks holds the potential to mitigate the limitations posed by weak labels. Additionally, the exploration of unsupervised and semi-supervised methods, combined with weak supervision, could provide new pathways for leveraging unlabelled video data effectively. Collaboration between computer scientists, clinicians, and domain experts is essential to develop more sophisticated algorithms that can understand and predict surgical dynamics accurately.

Few-shot Learning. Few-shot learning approaches aim to develop models that can generalize from very limited labeled data, a scenario commonly encountered in the medical field due to the high cost, privacy issues, and time constraints associated with annotating surgical videos. In the context of surgery, these methods are particularly valuable as they allow for the recognition and understanding of surgical actions, tools, and phases from only a handful of examples, thereby facilitating broader applicability across diverse surgical procedures and settings.

Domain Generalization. Domain Generalization (DG) techniques are increasingly vital as they allow models to be robust and applicable across different hospitals, surgical procedures, and patient demographics, without the need for retraining. This is particularly crucial in surgical video analysis, where the variance in lighting, surgical techniques, equipment, and individual patient anatomy can vastly differ.

References

1. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017)
2. Feichtenhofer, C.: X3d: Expanding architectures for efficient video recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 203–213 (2020)
3. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6202–6211 (2019)
4. Horn, B.K., Schunck, B.G.: Determining optical flow. Artificial intelligence **17**(1-3), 185–203 (1981)

5. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al.: The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017)
6. Li, Y., Wu, C.Y., Fan, H., Mangalam, K., Xiong, B., Malik, J., Feichtenhofer, C.: Mvitv2: Improved multiscale vision transformers for classification and detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4804–4814 (2022)
7. Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., Hu, H.: Video swin transformer. arXiv preprint arXiv:2106.13230 (2021)
8. Lucas, B.D., Kanade, T., et al.: An iterative image registration technique with an application to stereo vision, vol. 81. Vancouver (1981)
9. Shi, D., Zhong, Y., Cao, Q., Ma, L., Li, J., Tao, D.: Tridet: Temporal action detection with relative boundary modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18857–18866 (2023)
10. Tran, D., Wang, H., Torresani, L., Feiszli, M.: Video classification with channel-separated convolutional networks (2019)
11. Zhang, C., Wu, J., Li, Y.: Actionformer: Localizing moments of actions with transformers. In: European Conference on Computer Vision (2022)