# 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

**Abstract.** Surgical scene perception via videos is critical for advancing robotic surgery, telesurgery, and AI-assisted surgery, particularly in ophthalmology. However, the scarcity of diverse and richly annotated video datasets has hindered the development of intelligent systems for surgical workflow analysis. Existing datasets face challenges such as small scale, lack of diversity in surgery and phase categories, and absence of time-localized annotations. These limitations impede action understanding and model generalization validation in complex and diverse real-world surgical scenarios. To address this gap, we introduce OphNet, a large-scale, expert-annotated video benchmark for ophthalmic surgical workflow understanding. OphNet features: 1) A diverse collection of 2,278 surgical videos spanning 66 types of cataract, glaucoma, and corneal surgeries, with detailed annotations for 102 unique surgical phases and 150 fine-grained operations. 2) Sequential and hierarchical annotations for each surgery, phase, and operation, enabling comprehensive understanding and improved interpretability. 3) Time-localized annotations, facilitating temporal localization and prediction tasks within surgical workflows. With approximately 285 hours of surgical videos, OphNet is about 20 times larger than the largest existing surgical workflow analysis benchmark. Code and dataset are available at: https://minghu0830.github.io/OphNet-benchmark/.

**Keywords:** Surgical Workflow Understanding · Ophthalmic Surgery · Medical Image Analysis · Video Benchmark

## 1 Introduction

As surgical robot platforms such as the da Vinci® surgical system become increasingly sophisticated, there is growing interest in integrating enhanced intelli-

∗ Equal contribution † Corresponding author

gence into scenarios like minimally invasive surgery [20,32,68]. The advancements in machine vision perception empower these robotic systems to autonomously recognize and adapt to the intricacies of surgical environments, without relying on binary instrument usage signals, RFID tags, sensor data from tracking devices, or other signal information that necessitates laborious manual annotations or additional equipment installations [28]. This autonomy includes the capacity to identify anatomical structures, detect anomalies, and adjust surgical plans in real-time, which is crucial in dynamic and unpredictable surgical settings. In recent years, especially in endoscopy and ophthalmic surgery, the application of deep learning has demonstrated considerable promise in bolstering these autonomous capabilities. This encompasses the analysis of surgical workflows [6, 28, 74], segmentation of instruments and anatomy [4, 24, 43], and depth estimation [78], among others.

Automatic video surgical workflow understanding is a fundamental yet challenging problem for developing computer-assisted and robotic-assisted surgery, which can be divided into internal (e.g., laparoscopic and endoscopic [28,47,57]) and external (e.g., operating room and nursing procedure [41]) analysis. In addition to promoting the development of intelligent surgery, it also greatly benefits surgical documentation, education, and training [11, 12, 65]. Baret et al. [6] showed networks, like Inflated 3D ConvNet (I3D) [10] that utilize spatiotemporal convolutions, require a relatively extensive dataset for effective training. In their study, the model achieves an accuracy exceeding 80% when trained on 100 videos, with a progressive improvement as the sample size surpasses 700. However, the highly efficient and rapidly evolving deep learning technologies for surgical workflow analysis are currently limited by the following shortcomings in current video benchmarks: **1) Small-scale:** the majority of video datasets contain no more than 100 videos. For example, the CATARACTS [23] and CatRelDet [23] datasets contain only 50 and 21 surgical videos, respectively. These datasets are relatively small, insufficient for large-scale validation. **2) Limited categories of surgeries and phases:** almost all ophthalmic surgical video datasets only include cataract surgery and do not further classify specific types of surgeries. Additionally, the number of phase categories is also limited, like CatRelDet [23] only contains 4 different phase labels, which is insufficient to meet the requirements for evaluation in real clinical environments. **3) Coarse-grained annotation:** due to annotation costs, existing benchmarks often have coarse-grained action definitions. For example, adhesive injection may occur in two different phases: main incision and capsulorhexis, so it may be classified into different phase categories. Coarse-grained action definitions may lead to annotation bias. **4) Single time-boundary annotation:** they only annotate designated phases in the videos, ignoring the continuity across different stages of ophthalmic surgery, as well as the hierarchical relationship between surgery, phase, and operation. Simpler datasets, such as LensID [22], are limited to binary classification tasks distinguishing lens implantation from other irrelevant phases. **5) Uniform domain:** the videos are meticulously collected, and

| Protocol | Dataset Properties | | | | | | Tasks | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Datasets | No. of Videos | No. of Action Segments | No. of Surgery Categories | No. of Action Categories | Total Duration | Multi-Surgery Presence Recognition | Phase Recognition | Phase Localization | Phase Prediction |
| Endo&Lap | Cholec120 [45] | 120 | - | 1 | 7 | 76.2h | ✗ | ✓ | ✗ | ✗ |
| | SurgicalActions160 [54] | 160 | 160 | 1 | 16 | 0.2h | ✗ | ✓ | ✗ | ✗ |
| | HeiCo [39, 50] | 30 | - | 3 | 14 | 2.8h | ✗ | ✓ | ✓ | ✓ |
| | EndoVis 2021 [67] | 33 | 250 | 1 | 7 | 22.0h | ✗ | ✓ | ✗ | ✗ |
| | PitVis [3] | 25 | 287 | 1 | 17 | 33.3h | ✗ | ✓ | ✗ | ✗ |
| | CholecT50 [46] | 50 | - | 1 | 10 | 44.7h | ✗ | ✓ | ✓ | ✓ |
| | AutoLaparo [70] | 21 | 300 | 1 | 7 | 23.1h | ✗ | ✓ | ✓ | ✓ |
| OphScope | LensID [22] | 100 | 2,440 | 1 | 2 | 11.7h | ✗ | ✓ | ✗ | ✗ |
| | Cataract-101 [55] | 101 | 1,266 | 1 | 10 | 14.0h | ✗ | ✓ | ✓ | ✓ |
| | CatRelDet [23] | 21 | 2,400 | 1 | 4 | 2.0h | ✗ | ✓ | ✗ | ✗ |
| | CATARACTS [4] | 50 | 1,536 | 1 | 19 | 20.0h | ✗ | ✓ | ✓ | ✓ |
| | Cataract-1K [21] | 1,000 | 931 | 1 | 12 | 118.7h | ✗ | ✓ | ✓ | ✓ |
| | **OphNet(Ours)** | **2,278** | **9,795** | **66** | **150** | **284.8h** | ✓ | ✓ | ✓ | ✓ |

**Table 1:** The statistics comparison among existing workflow analysis datasets and our OphNet. Compared to other datasets, OphNet focuses on more comprehensive coverage of various surgery, phase and operation categories, collects a large number of videos, totaling 284.8 hours, and also enables a variety of recognition, localization and prediction tasks. OphNet demonstrates considerable competitiveness in both its scale and the richness of its labels. For instance, Cholec120 [45], Cholec80 [65], m2cai-workflow and LapChole [62] form one series, whereas CholecT50 [46], CholecT45 [46], and CholecT40 [44] comprise another series. We have excluded the following scenarios from our comparison: (1) non-open-source datasets such as Bypass170 [66], ESD [30], Yu's [74], etc.; (2) a superset of multiple open-source or non-open-source datasets, like Cholec207 [6], etc.; (3) datasets employed for lesion, anatomy, and instrument classification and segmentation, such as SUN-SEG [27], CVC-ClinicDB [7], ROBUST-MIS [52], Mesejo's [40], Cata7 [43], etc., anomaly detection such as PolypDiag [63] (from Hyper-Kvasir [9] and LDPolypVideo [38]), Kvasir-Capsule [59], etc., and other datasets not dedicated to workflow analysis. It's worth mentioning that even in comparison with the above datasets, OphNet demonstrates considerable competitiveness in both its scale and the richness of its labels. *Endo&Lap* denotes the endoscopic and laparoscopic protocol, *OphScope* denotes the ophthalmic microscope protocol. We choose the latest version for comparison in cases where datasets have multiple supplementary updates.

while this ensures video quality, the uniform style is not conducive to testing the model's domain generalization ability.

While some works have explored semi-supervised and self-supervised learning strategies [8, 51, 73, 75] to alleviate the cost of annotations or use only a small fraction of available labels, these approaches still lack competitiveness in performance compared to fully supervised learning. This deficiency in performance is impeding the widespread clinical application of these strategies. To address the shortage of sufficient labeled datasets, we construct OphNet, a large-scale

and expert-level video benchmark with high diversity, for ophthalmic surgical workflow understanding. The main advantages of OphNet are as follows:

- **Largest scale and diversity:** to the best of our knowledge, OphNet is currently the largest and most richly labeled dataset for surgical workflow analysis. It contains a number of videos 20 times greater than the current largest benchmark in ophthalmic surgery and far exceeds datasets in more established fields such as endoscopy. Additionally, OphNet includes the greatest variety of different types of surgeries, encompassing 66 different surgeries such as cataract, glaucoma, and corneal surgeries, along with 102 unique surgical phases and 150 distinct operations. This diversity significantly surpasses that of previous research.
- **Fine-grained, sequential and hierarchical annotation:** we have meticulously selected a subset of videos for annotation localization, with each video being annotated for an average of 22 operations. Additionally, we provide exquisite annotations at the levels of surgery, phase, and operation, catering to the requirements for training specific challenge models. This annotation design aims to offer a multifaceted understanding of surgical protocols, accommodate the nuances of each distinct surgery, and enhance the usability and interpretability of our dataset.
- **Expert-level manual annotation:** the annotation work for OphNet was completed by ten experienced ophthalmologists and five individuals with ophthalmic experience, encompassing, but not limited to, standardization of definitions for surgery, phase, and operation labels, video filtering, classification and localization annotations, and secondary verification, among others. Expert-level annotators ensure the quality and professionalism of OphNet.

## 2 Related Work

**Surgical Workflow Understanding.** Beyond its therapeutic advantages, minimally invasive surgery also provides the capability for operative video recording. These videos can be stored and later utilized for various purposes such as cognitive training, skill assessment, and surgical workflow analysis [76,77]. Techniques derived from the broader field of video content analysis and representation are increasingly being incorporated into the surgical realm [6,36,74]. A typical surgical workflow can be defined by a sequence of tasks or events, including patient positioning, incision, dissection, and suturing. These events are influenced not only by the specific type of surgery but also by the individual surgeon's proficiency and technique. Consequently, a comprehensive understanding of the surgical workflow necessitates a thorough examination of the temporal, spatial, and contextual facets of these tasks.

**Weakly-Supervised Video Learning.** Substantial pioneering work has been undertaken in the realms of video understanding [5,16,33,53,69,71,72,79]. Since even a small amount of videos easily comprises several million frames, methods that do not rely on a frame-level annotation are of special importance. Weakly-supervised video learning makes use of loosely labeled data to train models,

thereby obviating the necessity for exhaustively annotated training data. These weak labels might manifest in a variety of forms, including video-level labels or partial labels, which are less specific than frame-level or pixel-level annotations. The objective of weakly-supervised learning is to utilize these coarse labels to produce models capable of delivering fine-grained predictions, such as temporally precise action localization or detailed semantic segmentation [15, 35]. The large-scale annotation of surgical videos demands a significant investment of invaluable medical time resources. In this context, weakly supervised learning emerges as a viable solution to this bottleneck.

## 3 Dataset Construction

In this section, we detail the construction of our dataset, which involved meticulous data collection and preprocessing. YouTube was leveraged as a primary source to circumvent privacy issues while ensuring a broad representation of ophthalmic surgeries. Our selection criteria aimed to capture a wide range of video qualities and styles, specifically targeting cataract, glaucoma, and corneal surgeries due to their prevalence in clinical settings. We refined the dataset by excluding videos of inadequate quality or those depicting non-human subjects. The annotation process was designed to reflect the complex nature of eye surgeries, incorporating hierarchical classification to account for multiple conditions often treated within a single procedure. This was complemented by detailed localization annotations, delineating the distinct phases and techniques characteristic of ophthalmic operations, all undertaken by a dedicated team of ophthalmologists to ensure accuracy and relevance.

### 3.1 Data Collection & Preprocessing

**Collection.** Medical data exhibit unique privacy considerations and tend to be of smaller data volumes, characteristics that are particularly salient in the case of video data. By capitalizing on the wealth of surgical videos available on YouTube, we are able to obviate potential ethical and privacy concerns while concurrently enabling the rapid procurement of a substantial corpus of videos for further screening [1, 2, 17, 29]. To fulfill this objective, we deploy text-based search algorithms to probe each surgery on YouTube, obtaining videos with titles that incorporate the requisite surgical keywords. We select cataract, glaucoma, and corneal surgery—three of the most commonly performed ophthalmic surgeries in actual clinical environments—as the central subjects of our research. To expand our video collection, we bolster our search queries by integrating synonyms and abbreviations associated with each type of ophthalmic surgery. For instance, *cataract surgery* encompasses various types: *Phacoemulsification (abbr. PHACO)*, *Intraocular Lens implantation (abbr. IOL)*, and *Extracapsular Cataract Extraction (abbr. ECCE)*, etc.

**Preprocessing.** Given the nature of text-based retrieval and the varying quality, style, and filming methods of surgical videos influenced by different YouTube

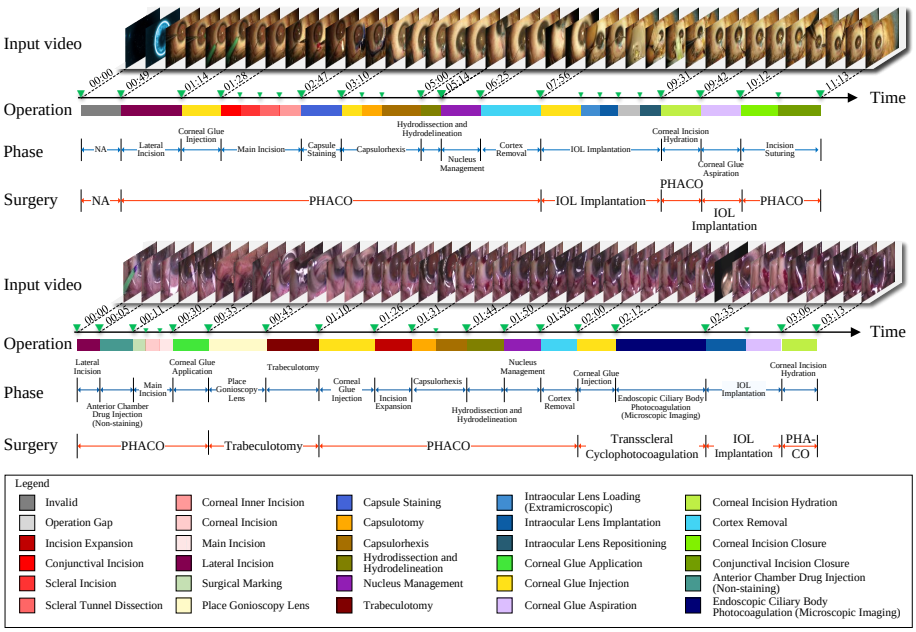



**Fig. 1:** The figure shows two combined surgical videos, *PHACO + IOL implantation* and *PHACO + Trabeculotomy + Transscleral Cyclophotocoagulation + IOL implantation*. For each frame marked in color, we provide time-boundary annotations at surgical, phase and operation levels.

sources, some videos may exhibit poor quality or deviations in surgical representation. We filter out low-resolution videos, black-and-white color schemes, and animated demonstrations. Furthermore, surgical videos featuring non-human eyes, such as pig eyes, rabbit eyes, or pseudo-human eyes, are also outside the scope of our annotation. The data preprocessing was jointly completed by five professionals with experience in ophthalmology for initial screening, followed by a review conducted by six attending ophthalmologists.

### 3.2 Data Annotation

**Classification Annotation.** Different from the single-label classification in most natural video datasets [17, 25, 58, 60, 80], in ophthalmic surgery, physicians typically consider and implement multiple types of surgeries and phases based on the specific conditions and needs of the patient. This multifaceted approach stems from the complexity of the eye as an organ, where numerous diseases often coexist. Consequently, multiple ocular problems may need to be addressed in a single surgery. As shown in Fig. 1, a patient with cataracts may also have glaucoma, necessitating the simultaneous treatment of both conditions within one surgery. Hence, we initially classified all videos based on the primary surgical categories into three principal groups: cataract, glaucoma, and corneal

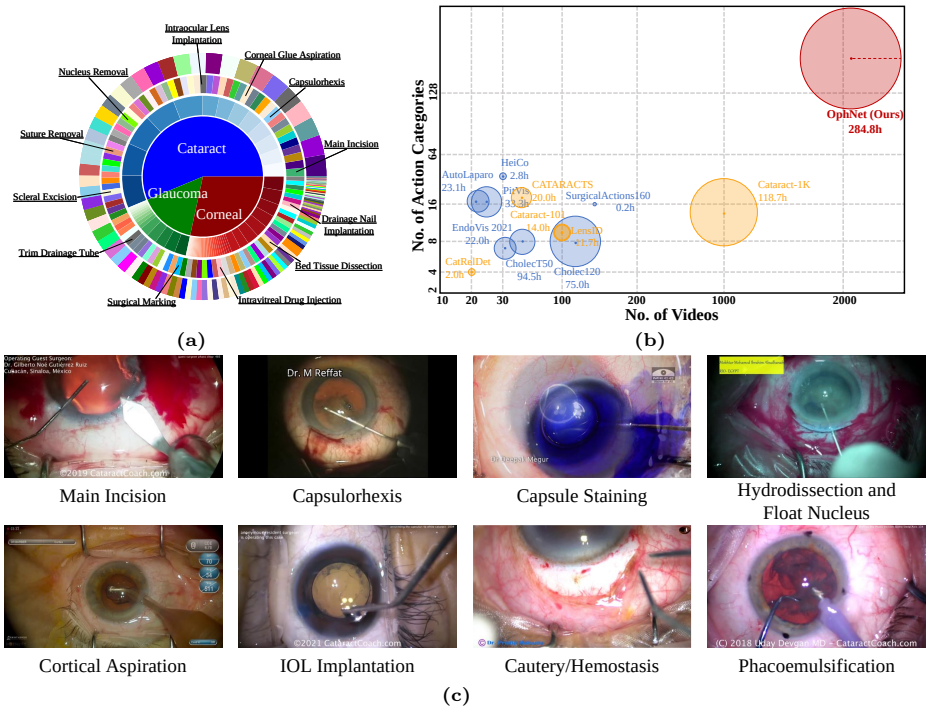
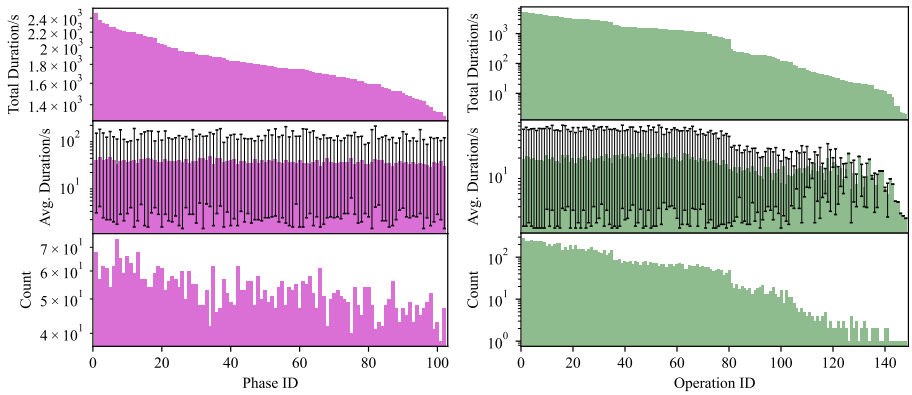



**Fig. 2:** OphNet's composition, comparison with other datasets for the same task, and some phase examples: (a) an overview of the composition ratios at the levels of surgery, phase, and operation; (b) comparison among existing open-source ◯ laparoscopic & endoscopic, and ◯ ophthalmic microscope workflow analysis video datasets and ◯ our OphNet. OphNet stands as the largest real-world video dataset for ophthalmic surgical workflow understanding, featuring the highest number of videos, longest duration, and diverse categories of surgeries and phases; (c) eight phase examples in OphNet.

surgeries. Subsequently, these were further allocated for the detailed annotation of primary and secondary surgery. It is important to note that there exists only a single type of primary surgery, whereas multiple secondary surgery may be present. The categorization and annotation were meticulously completed by a team comprised of 8 experienced ophthalmologists.

**Hierarchical Localization Annotation.** A single surgery is often multi-faceted, involving a series of intricate phases, each requiring distinct techniques and instruments, and the transition between surgeries entails high precision and coordination. Routine cataract surgery involves several steps. It begins with the administration of anesthesia, followed by a small incision in the cornea. The surgeon then creates an opening in the lens capsule and uses ultrasonic vibrations to break up and remove the cataract. Afterward, an artificial intraocular lens (IOL) is inserted into the lens capsule, and the incision is sealed without stitches. We define the phases and operation for various surgeries based on the




**(a)** phase statistics **(b)** operation statistics

**Fig. 3:** We present the data statistics of trimmed videos at the levels of phase and operation, including the number of trimmed videos, average duration, and total duration. The IDs and corresponding names can be found in the appendix.

textbook *Ophthalmic Surgery: Principles and Practice* [61]. To ensure the quality of localization annotations, we assign each annotator videos of 2 different types of primary surgeries. For each action time-boundary annotation, we annotate at three different levels of granularity: surgery, phase, and operation. We also employed the complete linkage algorithm [13] to cluster and merge various temporal boundaries into stable boundaries that received multiple agreements. It's worth noting that an individual video may have multiple separate instances of the same or different phases, thereby leading to multiple boundary definitions. A team of 15 ophthalmologists completes the localization annotation.

### 3.3 Dataset Statistics and Analysis

OphNet includes 2,278 surgical videos (284.8 hours), demonstrating 66 different types of ophthalmic surgeries: 13 types of cataract surgery, 14 types of glaucoma surgery, and 39 types of corneal surgery. There are 102 phases and 150 operations for recognition, detection and prediction tasks, summarized in Tab. 1. Over 77% of videos have high-definition resolutions of 1280 $\times$ 720 pixels or higher. To facilitate algorithm development and evaluation, we selected 523 videos for localization annotations. Additionally, we trim videos according to annotated action boundaries, resulting in 7,320 phase instances and 9,795 operation instances, totaling 51.2 hours. The average duration of trimmed videos is 32 seconds, while untrimmed videos average 337 seconds.

## 4 Experiments

In our study, we explore four potential tasks using the OphNet dataset: 1) primary surgery presence recognition, 2) phase and operation recognition, 3) phase

localization, and 4) phase anticipation. To establish robust baselines, we employ state-of-the-art models known for their effectiveness in human action recognition, detection and anticipation. For each task, we provide a detailed problem formulation and evaluate the baseline models' performance. Our findings offer valuable insights into video understanding within sequences and fine granularity, contributing to the field's knowledge of video tasks in medical contexts.

### 4.1 Surgery Presence, Phase and Operation Recognition

**Task Description.** Surgery presence recognition focuses on identifying various surgical types in untrimmed videos through weakly supervised methods. This requires the model to discern and capture distinct surgical action features within extensive video footage. In OphNet, all types of surgeries occurring in each video are annotated, but only a subset of the videos have time-boundary annotations, as detailed in Sec. 3.2. We identify the primary surgery based on the key objectives and durations of the surgeries. To streamline the process, our experiments are limited to recognizing the presence of these primary surgeries. Additionally, phase recognition segments the surgery into distinct phases using visual cues and movements, such as incision, lens removal, and implantation. Operation recognition involves identifying finer-grained surgical actions.

**Baselins.** We compare the performance of I3D [10], SlowFast [19], X3D [18], and MViT V2 [31] models on this task. These models are evaluated in two versions: 1) random initialization training and 2) pretraining with weights from Kinetics 400 [29], which is a human action recognition dataset. In addition, we also explored the classification performance of X-CLIP [42] and ViFi-CLIP [49], two CLIP [48]-based models. We also compared the effects of different numbers of input frames on the models' performance, where the subscript *_16* represents an input of 16 frames, and *_32* represents an input of 32 frames.

| Baselines | Primary Surgery Classification | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Cataract | | Glaucoma | | Cornea | | All | |
| | Top-1 | Top-5 | Top-1 | Top-5 | Top-1 | Top-5 | Top-1 | Top-5 |
| I3D [10] | 38.9 | 86.3 | 43.6 | 71.7 | 42.7 | 78.2 | 29.8 | 53.2 |
| SlowFast [19] | 45.3 | 82.1 | 44.6 | 72.3 | 45.5 | 77.3 | 27.2 | 54.4 |
| X3D [31] | 42.1 | 87.4 | 44.6 | 74.2 | 43.8 | 76.6 | 28.5 | 62.7 |
| MViT V2 [18] | 43.2 | 84.2 | 45.5 | 81.8 | 45.5 | 75.9 | 29.1 | 60.1 |
| I3D [10] | 36.8 | 84.7 | 48.2 | 81.8 | 48.6 | 76.5 | 27.2 | 50.6 |
| SlowFast* | 49.0 | 83.7 | 47.3 | 80.9 | 49.2 | 75.6 | 27.2 | 50.6 |
| X3D* | 47.4 | 86.3 | 46.4 | 81.5 | 48.3 | 78.2 | 35.4 | 61.4 |
| MViT V2* | 44.2 | 85.3 | 49.1 | 81.8 | 47.7 | 77.3 | 28.5 | 63.3 |
| X-CLIP$_{16}$ [42] | 58.5 | **94.7** | 51.8 | **92.8** | **61.4** | **88.6** | 40.5 | 79.0 |
| X-CLIP$_{32}$ | **60.6** | 92.6 | **53.5** | 83.7 | 56.8 | 84.1 | **58.9** | **81.0** |
| ViFi-CLIP$_{16}$ [49] | 59.6 | 88.3 | 52.4 | 80.2 | 61.4 | 81.8 | 58.9 | 79.8 |
| ViFi-CLIP$_{32}$ | 59.6 | 88.3 | 51.5 | 84.8 | 50.0 | 75.0 | 56.3 | 77.2 |

**Table 2:** Per-class Top-1 and Top-5 accuracy (%) for the primary surgery presence recognition on untrimmed videos. The best performance for each split has been highlighted in **bold**.

**Setup.** The dataset was randomly partitioned to ensure a balanced representation of examples for each surgery category. Specifically, we allocated 70% of the data for training (1,449 surgical videos), 10% for validation (205 surgical videos), and 20% for testing (424 surgical videos). For the phase and operation recognition experiments, we followed the same settings, with 70% of the data used for training (5,024 phase segments, 6,856 operations segments), 10% for

validation (730 phase segments, 975 operations segments), and 20% for testing (1,566 phase segments, 1,964 operations segments). The input for the surgery presence recognition experiment is untrimmed videos, while for the phase and operation recognition experiments, the input consists of trimmed segments. In all classification experiments, we set up the analysis and comparison from four perspectives: cataract surgery, glaucoma surgery, corneal surgery, and all surgical videos.

**Results.** We summarize the results in Tab. 2 and Tab. 3. In primary surgery classification, X-CLIP [42] achieved the highest overall Top-1 accuracy at 58.9%, leading in cataract and glaucoma surgeries with accuracies of 60.6% and 53.5% respectively. For corneal surgeries, X-CLIP also recorded the highest Top-1 and Top-5 accuracies of 61.4% and 88.6%. In phase classification, ViFi-CLIP [49] led with the highest Top-1 accuracy, particularly in cataract surgeries at 75.9%, and exhibited the best performance in both glaucoma and corneal surgeries. Furthermore, in operation classification, ViFi-CLIP outperformed all other models in all categories, especially noted in cataract and corneal surgeries with Top-1 accuracies of 75.1% and 83.7% and Top-5 accuracies of 93.8% and 85.2% respectively. Overall, ViFi-CLIP showed superior performance in both phase and operation classifications across various surgery types. Besides, in the phase and operation classification experiments, a higher number of input frames generally had a positive effect on the model. In Fig. 4, we present the heatmap visualization of four examples from the test set of phase recognition experiments using the ViFi-CLIP model. It can be observed that in a series of frame images, the model focuses on surgical instruments and the operated eye area, which is consistent with human experience. The instruments used in the same phase of ophthalmic surgery are often similar.

### 4.2 Phase Localization

**Task Description.** Phase localization in ophthalmic surgical workflow analysis refers to the task of pinpointing the exact moments or time intervals within a surgical video where specific phases of the surgery begin and end. This involves the detailed temporal segmentation of the entire surgical procedure into its constituent phases based on visual cues, surgeon's actions, and the progression of the surgery. The objective of phase localization is to accurately identify the start and end times of different surgical stages, such as pre-operative preparation, incision, and removal of the lens facilitating a granular and precise understanding of the surgery timeline. This task is crucial for detailed surgical documentation, efficient surgical training, and the development of targeted interventions during specific stages of the surgery, enhancing overall surgical management and post-operative analysis.

**Baselines.** We conducted the experiments using phase-level labels and used two baseline models, ActionFormer [79] and TriDet [56], with backbone networks configured as CSN [64], SwinViviT [34], and SlowFast [19]. Data split follows the setup of the primary surgery classification experiment.

| Baselines | Phase Classification | | | | | | | | Operation Classification | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Cataract | | Glaucoma | | Cornea | | All | | Cataract | | Glaucoma | | Cornea | | All | |
| | Top-1 | Top-5 | Top-1 | Top-5 | Top-1 | Top-5 | Top-1 | Top-5 | Top-1 | Top-5 | Top-1 | Top-5 | Top-1 | Top-5 | Top-1 | Top-5 |
| I3D [10] | 27.2 | 55.7 | 24.1 | 57.5 | 18.9 | 52.1 | 25.7 | 58.2 | 26.8 | 54.9 | 23.5 | 56.0 | 18.0 | 51.2 | 25.0 | 57.1 |
| SlowFast [19] | 26.5 | 56.5 | 23.1 | 56.5 | 24.2 | 49.1 | 26.7 | 60.1 | 25.8 | 55.2 | 22.9 | 45.9 | 23.5 | 48.5 | 26.0 | 59.0 |
| X3D [31] | 27.0 | 58.3 | 21.0 | 55.5 | 21.8 | 28.5 | 26.6 | 62.3 | 26.4 | 47.2 | 20.5 | 44.6 | 21.3 | 27.8 | 26.1 | 61.5 |
| MViT V2 [18] | 26.2 | 54.9 | 21.0 | 53.4 | 26.0 | 46.8 | 27.0 | 59.8 | 25.9 | 43.8 | 20.5 | 42.7 | 25.5 | 45.9 | 26.5 | 58.9 |
| I3D* | 29.5 | 68.9 | 22.1 | 58.6 | 26.0 | 50.9 | 30.2 | 71.2 | 28.8 | 67.5 | 21.8 | 47.9 | 25.7 | 50.3 | 29.5 | 60.0 |
| SlowFast* | 30.6 | 72.3 | 25.2 | 54.7 | 30.7 | 59.8 | 31.7 | 61.8 | 29.9 | 71.1 | 24.8 | 43.9 | 29.5 | 58.7 | 30.5 | 60.9 |
| X3D* | 27.2 | 72.9 | 22.1 | 59.6 | 30.7 | 61.5 | 33.5 | 63.2 | 26.5 | 71.8 | 21.7 | 48.8 | 29.9 | 60.2 | 32.8 | 62.1 |
| MViT V2* | 34.2 | 76.5 | 23.3 | 52.0 | 38.4 | 65.1 | 28.3 | 60.2 | 33.5 | 75.2 | 22.8 | 41.5 | 37.9 | 64.0 | 27.8 | 59.5 |
| X-CLIP$_{16}$ [42] | 68.3 | 92.2 | 47.3 | 89.8 | 53.0 | 77.4 | 63.4 | 85.3 | 67.5 | 91.0 | 46.5 | 78.9 | 82.2 | 76.1 | 62.5 | 84.0 |
| X-CLIP$_{32}$ | 69.1 | **94.0** | 48.7 | 81.7 | 54.8 | 80.4 | 62.7 | 85.8 | 68.0 | 93.0 | **47.9** | **80.5** | 84.0 | 79.5 | 62.0 | 84.7 |
| ViFi-CLIP$_{16}$ [49] | **75.9** | 93.7 | 40.4 | 85.4 | 66.6 | 81.6 | 66.1 | **88.4** | 74.5 | 92.5 | 42.8 | 74.5 | **85.0** | 80.5 | **65.0** | **87.5** |
| ViFi-CLIP$_{32}$ | 73.0 | 92.9 | **49.6** | **82.7** | **57.7** | **81.6** | **68.4** | 87.2 | **75.1** | **93.8** | 43.2 | 80.2 | 83.7 | 85.2 | 64.8 | 86.5 |

**Table 3:** Per-class Top-1 and Top-5 accuracy (%) for the primary surgery presence recognition on untrimmed videos and phase recognition on trimmed videos. * denotes the initialization from the model pre-trained on Kinetics 400 [29]. For the two CLIP models, we chose ViT-B/16 as the backbone and compared the performance of two different input frame numbers, 16 and 32. The best performance for each split has been highlighted in **bold**.

| Baselines | Backbones | mAP (%) | | | | |
|---|---|---|---|---|---|---|
| | | 0.1 | 0.3 | 0.5 | 0.7 | Avg. |
| ActionFormer [79] | CSN [64] | 53.7 | 50.1 | 40.6 | 24.5 | 42.5 |
| | SwinViviT [34] | 59.3 | 54.7 | 43.3 | 26.3 | 46.4 |
| | SlowFast [19] | 60.0 | 55.9 | 45.1 | 26.0 | 47.5 |
| TriDet [56] | CSN | 56.1 | 53.0 | 43.1 | 29.4 | 46.2 |
| | SwinViviT | 61.0 | **57.1** | **47.1** | **33.1** | **50.4** |
| | SlowFast | **61.3** | 56.0 | 45.6 | 30.4 | 48.6 |

**Table 4:** The results for phase detection. ActionFormer and TriDet are state-of-the-art models for human action detection tasks, and we use three different backbones for feature extraction and report mAP at the IoU thresholds of [0.1:0.2:0.9]. Average mAP is computed by averaging different IoU thresholds. The best performance for each split has been highlighted in **bold**.

| Baselines | Top-1 Acc. (%) | | | | |
|---|---|---|---|---|---|
| | 0.1 | 0.3 | 0.5 | 0.7 | Avg. |
| I3D [10] | 26.5 | 42.2 | 49.8 | 51.3 | 47.3 |
| SlowFast [19] | 25.4 | 42.6 | 48.9 | 52.2 | 47.2 |
| MViT V2 [18] | 25.6 | 43.7 | 49.3 | 52.3 | 47.5 |
| I3D* | 27.3 | 43.5 | 50.1 | 51.4 | 47.6 |
| SlowFast* | 27.5 | 43.2 | 49.9 | **52.3** | 47.8 |
| MViT V2* | **27.8** | **43.8** | **50.5** | 51.7 | **48.2** |

**Table 5:** The results for phase anticipation. We report top-1 accuracy at the observation ratios [0.1:0.2:0.9]. Average top-1 accuracy is computed by averaging different observation ratios. The best performance for each split has been highlighted in **bold**.

**Setup.** We excluded *Operation Gap* and *Invalid*, and filtered out tags with fewer than 20 segments. We used two baseline models, ActionFormer [79] and TriDet [56], with backbone networks configured as CSN [64], SwinViviT [34], and SlowFast [19]. To extract features from the videos, we first extracted RGB frames from each video at a rate of 25 frames per second. We also extracted optical flow using the TV-L1 [26, 37] algorithm. We then fine-tuned an I3D [10] model that had been pre-trained on the ImageNet [14] dataset, and used it to generate features for each RGB and optical flow frame. Because each video has a variable duration, we performed uniform interpolation to generate 100 fixed-

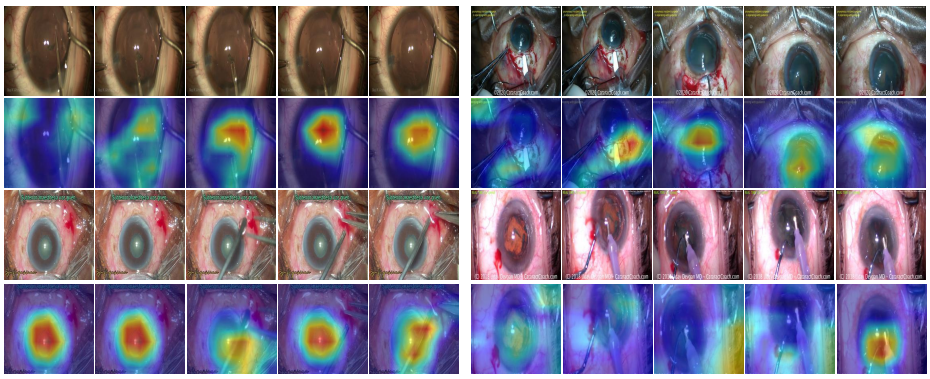


**Fig. 4:** Attention map visualizations of ViFi-CLIP [49] on four examples from OphNet's test set in the phase recognition task.

length features for each video. Finally, we concatenated the RGB and optical flow features into a 2048-dimensional embedding, which served as the input for our model.

**Results.** The experimental results for phase localization are presented in the Tab. 4, showcasing the performance of different baseline models with various backbones in terms of mean Average Precision (mAP) at different Intersection over Union (IoU) thresholds [0.1:0.2:0.9]. The models evaluated include ActionFormer with SwinViviT and SlowFast backbones, and TriDet with CSN, SwinViviT, and SlowFast backbones. The results indicate that the TriDet model with a SwinViviT backbone outperforms other combinations, achieving the highest mAP scores across most IoU thresholds, with notable scores of 61.0% (IoU=0.1), 57.1% (IoU=0.3), 47.1% (IoU=0.5), and 33.1% (IoU=0.7), resulting in an average mAP of 50.4%. This indicates that the TriDet model, especially when combined with the SwinViviT backbone, is particularly effective for phase localization in surgical videos. On the other hand, the TriDet model with a SlowFast backbone shows competitive performance, particularly achieving the highest mAP of 61.3% at the lowest IoU threshold (0.1). However, it falls slightly behind in performance at higher IoU thresholds compared to the SwinViviT backbone.

### 4.3 Phase Anticipation

**Task Description.** This task requires the analysis of real-time or recorded video data to foresee the sequence of events based on current and past surgical activities. By understanding the typical progression of ophthalmic surgeries and recognizing patterns in the surgeon's actions and the use of instruments, the system aims to forecast the next phase of the surgery, allowing for proactive preparation and response. The objective of phase anticipation is to enhance the efficiency and safety of surgical procedures by providing the surgical team with advanced notice of upcoming steps, enabling better resource allocation, timing for critical tasks, and overall coordination within the operating room.

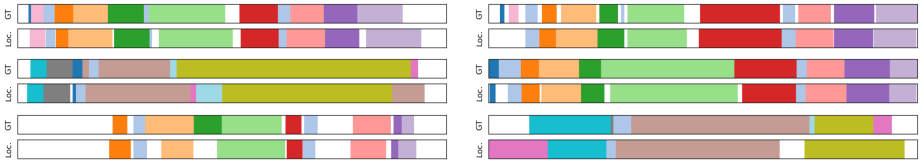


**Fig. 5:** Phase localization visualization of TriDet [56]. *GT* represents the ground truth visualization for phases, while *Loc.* visualizes the model's highest confidence phase category and the time-boundary results. Blank segments denote invalid segments or operation gaps.

**Setup.** Following previous approaches of the primary surgery classification experiment in Sec. 4.1, We randomly mask phase sequences in the test video with different observation ratios.
**Baselines.** We evaluate our datasets with the baseline models such as I3D [10], SlowFast [19], and MViT V2 [31]. For each model, we also adopted two training approaches: random initialization training and using pre-trained weights from Kinetics 400 [29].
**Results.** The phase detection results are illustrated in Tab. 5. The results demonstrate that the baseline models pretrained on Kinetics 400 [29] generally outperform their original counterparts in terms of Top-1 accuracy across different observation ratios for phase anticipation. Specifically, the modified MViT V2* model exhibits the highest improvement, achieving the best average Top-1 accuracy of 48.2%. Moreover, while all models show increased accuracy with higher observation ratios, indicating that more observed data contributes to better performance, the consistent improvement across all ratios for the enhanced models suggests effective modifications.

## 5 Limitations

**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 a little subtitle or watermark shown in Fig. 2. 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, there is a distinct 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 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.

## 6 Conclusion

In response to the current challenges in ophthalmology, a surgical field apt for automation and remote control, we introduce OphNet, a large-scale, diverse, and expert-level video benchmark for understanding ophthalmic surgical workflows. OphNet is the most extensive dataset of its kind, containing a broad range of cataract, glaucoma, and corneal surgeries and detailed annotations for distinct surgical phases. OphNet comprises 2,278 surgical videos (284.8 hours), 7,320 phase segments and 9,795 operation segments (51.2 hours), showcasing 66 different types of ophthalmic surgeries: 13 cataract, 14 glaucoma, and 39 corneal. It is annotated with 102 phases and 150 operations. With OphNet, we explored primary surgery presence recognition, phase localization and phase anticipation on untrimmed videos, phase and operation recognition on trimmed videos. We employed state-of-the-art models to establish robust baselines and provided valuable insights into video understanding within sequences and fine granularity. Our work contributes to the broader understanding of surgical video tasks in medical contexts and promotes the integration of deep learning technologies into ophthalmic surgical procedures.

## References

1. Fair use on youtube. https://support.google.com/youtube/answer/9783148?hl=en#:~:text=If%20the%20use%20of%20copyright,copyright%20removal%20request%20to%20YouTube.
2. Youtube's copyright exception policy. https://www.youtube.com/howyoutubeworks/policies/copyright/#copyright-exceptions
3. Adrito, D., Danyal, Z.K., Dimitrios, P., Yitong, Z., John, G.H., Francisco, V., Sophia, B., Hani, J.M., Danail, S.: Pitvis: Workflow recognition in endoscopic pituitary surgery
4. Al Hajj, H., Lamard, M., Conze, P.H., Roychowdhury, S., Hu, X., Maršalkaitė, G., Zisimopoulos, O., Dedmari, M.A., Zhao, F., Prellberg, J., Sahu, M., Galdran, A., Araújo, T., Vo, D.M., Panda, C., Dahiya, N., Kondo, S., Bian, Z., Vahdat, A., Bialopetravičius, J., Flouty, E., Qiu, C., Dill, S., Mukhopadhyay, A., Costa, P., Aresta, G., Ramamurthy, S., Lee, S.W., Campilho, A., Zachow, S., Xia, S., Conjeti, S., Stoyanov, D., Armaitis, J., Heng, P.A., Macready, W.G., Cochener, B., Quellec, G.: Cataracts: Challenge on automatic tool annotation for cataract

surgery. Medical Image Analysis **52**, 24–41 (2019). https://doi.org/https://doi.org/10.1016/j.media.2018.11.008, https://www.sciencedirect.com/science/article/pii/S136184151830865X
5. Alwassel, H., Giancola, S., Ghanem, B.: TSP: Temporally-sensitive pretraining of video encoders for localization tasks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. pp. 3173–3183 (2021)
6. Bar, O., Neimark, D., Zohar, M., Hager, G.D., Girshick, R., Fried, G.M., Wolf, T., Asselmann, D.: Impact of data on generalization of ai for surgical intelligence applications. Scientific Reports **10**(1), 22208 (Dec 2020). https://doi.org/10.1038/s41598-020-79173-6, https://doi.org/10.1038/s41598-020-79173-6
7. Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., Gil, D., Rodríguez, C., Vilariño, F.: Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. Computerized Medical Imaging and Graphics **43**, 99–111 (2015). https://doi.org/https://doi.org/10.1016/j.compmedimag.2015.02.007, https://www.sciencedirect.com/science/article/pii/S0895611115000567
8. Bodenstedt, S., Wagner, M., Katić, D., Mietkowski, P., Mayer, B., Kenngott, H., Müller-Stich, B., Dillmann, R., Speidel, S.: Unsupervised temporal context learning using convolutional neural networks for laparoscopic workflow analysis. arXiv preprint arXiv:1702.03684 (2017)
9. Borgli, H., Thambawita, V., Smedsrud, P.H., Hicks, S., Jha, D., Eskeland, S.L., Randel, K.R., Pogorelov, K., Lux, M., Nguyen, D.T.D., Johansen, D., Griwodz, C., Stensland, H.K., Garcia-Ceja, E., Schmidt, P.T., Hammer, H.L., Riegler, M.A., Halvorsen, P., de Lange, T.: HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. Scientific Data **7**(1), 283 (2020). https://doi.org/10.1038/s41597-020-00622-y, https://doi.org/10.1038/s41597-020-00622-y
10. 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)
11. Czempiel, T., Paschali, M., Keicher, M., Simson, W., Feussner, H., Kim, S.T., Navab, N.: Tecno: Surgical phase recognition with multi-stage temporal convolutional networks. In: Martel, A.L., Abolmaesumi, P., Stoyanov, D., Mateus, D., Zuluaga, M.A., Zhou, S.K., Racoceanu, D., Joskowicz, L. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. pp. 343–352. Springer International Publishing, Cham (2020)
12. Czempiel, T., Paschali, M., Ostler, D., Kim, S.T., Busam, B., Navab, N.: Opera: Attention-regularized transformers for surgical phase recognition. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part IV 24. pp. 604–614. Springer (2021)
13. Defays, D.: An efficient algorithm for a complete link method. The Computer Journal **20**(4), 364–366 (01 1977). https://doi.org/10.1093/comjnl/20.4.364, https://doi.org/10.1093/comjnl/20.4.364
14. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
15. Dong, S., Hu, H., Lian, D., Luo, W., Qian, Y., Gao, S.: Weakly supervised video representation learning with unaligned text for sequential videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2437–2447 (2023)

16. Duong, H.T., Le, V.T., Hoang, V.T.: Deep learning-based anomaly detection in video surveillance: A survey. Sensors **23**(11) (2023). https://doi.org/10.3390/s23115024, https://www.mdpi.com/1424-8220/23/11/5024
17. Fabian Caba Heilbron, Victor Escorcia, B.G., Niebles, J.C.: Activitynet: A large-scale video benchmark for human activity understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 961–970 (2015)
18. 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)
19. 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)
20. Forslund Jacobsen, M., Konge, L., Alberti, M., la Cour, M., Park, Y.S., Thomsen, A.S.S.: ROBOT-ASSISTED VITREORETINAL SURGERY IMPROVES SURGICAL ACCURACY COMPARED WITH MANUAL SURGERY: A randomized trial in a simulated setting. Retina **40**(11), 2091–2098 (Nov 2020)
21. Ghamsarian, N., El-Shabrawi, Y., Nasirihaghighi, S., Putzgruber-Adamitsch, D., Zinkernagel, M., Wolf, S., Schoeffmann, K., Sznitman, R.: Cataract-1k: Cataract surgery dataset for scene segmentation, phase recognition, and irregularity detection. arXiv preprint arXiv:2312.06295 (2023)
22. Ghamsarian, N., Taschwer, M., Putzgruber-Adamitsch, D., Sarny, S., El-Shabrawi, Y., Schoeffmann, K.: Lensid: A cnn-rnn-based framework towards lens irregularity detection in cataract surgery videos. In: de Bruijne, M., Cattin, P.C., Cotin, S., Padoy, N., Speidel, S., Zheng, Y., Essert, C. (eds.) Medical Image Computing and Computer Assisted Intervention - MICCAI 2021 - 24th International Conference, Strasbourg, France, September 27 - October 1, 2021, Proceedings, Part VIII. Lecture Notes in Computer Science, vol. 12908, pp. 76–86. Springer (2021). https://doi.org/10.1007/978-3-030-87237-3_8, https://doi.org/10.1007/978-3-030-87237-3_8
23. Ghamsarian, N., Taschwer, M., Putzgruber-Adamitsch, D., Sarny, S., Schoeffmann, K.: Relevance detection in cataract surgery videos by spatio- temporal action localization. In: 25th International Conference on Pattern Recognition, ICPR 2020, Virtual Event / Milan, Italy, January 10-15, 2021. pp. 10720–10727. IEEE (2020). https://doi.org/10.1109/ICPR48806.2021.9412525, https://doi.org/10.1109/ICPR48806.2021.9412525
24. Grammatikopoulou, M., Flouty, E., Kadkhodamohammadi, A., Quellec, G., Chow, A., Nehme, J., Luengo, I., Stoyanov, D.: Cadis: Cataract dataset for image segmentation. arXiv preprint arXiv:1906.11586 (2019)
25. Gu, C., Sun, C., Ross, D.A., Vondrick, C., Pantofaru, C., Li, Y., Vijayanarasimhan, S., Toderici, G., Ricco, S., Sukthankar, R., Schmid, C., Malik, J.: Ava: A video dataset of spatio-temporally localized atomic visual actions (2018)
26. Horn, B.K., Schunck, B.G.: Determining optical flow. Artificial intelligence **17**(1-3), 185–203 (1981)
27. Ji, G.P., Xiao, G., Chou, Y.C., Fan, D.P., Zhao, K., Chen, G., Van Gool, L.: Video polyp segmentation: A deep learning perspective. Machine Intelligence Research **19**(6), 531–549 (Dec 2022). https://doi.org/10.1007/s11633-022-1371-y, https://doi.org/10.1007/s11633-022-1371-y
28. Jin, Y., Dou, Q., Chen, H., Yu, L., Qin, J., Fu, C.W., Heng, P.A.: SV-RCNet: Workflow recognition from surgical videos using recurrent convolutional network. IEEE Trans. Med. Imaging **37**(5), 1114–1126 (May 2018)

29. 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)
30. Li, J., Jin, Y., Chen, Y., Yip, H.C., Scheppach, M., Chiu, P.W.Y., Yam, Y., Meng, H.M.L., Dou, Q.: Imitation learning from expert video data for dissection trajectory prediction in endoscopic surgical procedure. In: Greenspan, H., Madabhushi, A., Mousavi, P., Salcudean, S., Duncan, J., Syeda-Mahmood, T., Taylor, R. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2023. pp. 494–504. Springer Nature Switzerland, Cham (2023)
31. 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)
32. Lin, S., Miao, A.J., Lu, J., Yu, S., Chiu, Z.Y., Richter, F., Yip, M.C.: Semantic-super: A semantic-aware surgical perception framework for endoscopic tissue identification, reconstruction, and tracking. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 4739–4746 (2023). https://doi.org/10.1109/ICRA48891.2023.10160746
33. Lin, T., Liu, X., Li, X., Ding, E., Wen, S.: Bmn: Boundary-matching network for temporal action proposal generation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3889–3898 (2019)
34. Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., Hu, H.: Video swin transformer. arXiv preprint arXiv:2106.13230 (2021)
35. Long, F., Yao, T., Qiu, Z., Tian, X., Luo, J., Mei, T.: Bi-calibration networks for weakly-supervised video representation learning. International Journal of Computer Vision **131**(7), 1704–1721 (Jul 2023). https://doi.org/10.1007/s11263-023-01779-w, https://doi.org/10.1007/s11263-023-01779-w
36. Loukas, C.: Video content analysis of surgical procedures. Surgical Endoscopy **32**(2), 553–568 (Feb 2018). https://doi.org/10.1007/s00464-017-5878-1, https://doi.org/10.1007/s00464-017-5878-1
37. Lucas, B.D., Kanade, T., et al.: An iterative image registration technique with an application to stereo vision, vol. 81. Vancouver (1981)
38. Ma, Y., Chen, X., Cheng, K., Li, Y., Sun, B.: Ldpolypvideo benchmark: A large-scale colonoscopy video dataset of diverse polyps. In: de Bruijne, M., Cattin, P.C., Cotin, S., Padoy, N., Speidel, S., Zheng, Y., Essert, C. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. pp. 387–396. Springer International Publishing, Cham (2021)
39. Maier-Hein, L., Wagner, M., Ross, T., Reinke, A., Bodenstedt, S., Full, P.M., Hempe, H., Mindroc-Filimon, D., Scholz, P., Tran, T.N., et al.: Heidelberg colorectal data set for surgical data science in the sensor operating room. Scientific data **8**(1), 101 (2021)
40. Mesejo, P., Pizarro, D., Abergel, A., Rouquette, O., Beorchia, S., Poincloux, L., Bartoli, A.: Computer-aided classification of gastrointestinal lesions in regular colonoscopy. IEEE Transactions on Medical Imaging **35**(9), 2051–2063 (2016). https://doi.org/10.1109/TMI.2016.2547947
41. Ming, H., Lin, W., Siyuan, Y., Don, M., Qingli, R., Peng, X., Wei, F., Peibo, D., Lie, J., Zongyuan, G.: Nurvid: A large expert-level video database for nursing procedure activity understanding. In: Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2023)
42. Ni, B., Peng, H., Chen, M., Zhang, S., Meng, G., Fu, J., Xiang, S., Ling, H.: Expanding language-image pretrained models for general video recognition (2022)

43. Ni, Z.L., Bian, G.B., Zhou, X.H., Hou, Z.G., Xie, X.L., Wang, C., Zhou, Y.J., Li, R.Q., Li, Z.: Raunet: Residual attention u-net for semantic segmentation of cataract surgical instruments. In: International Conference on Neural Information Processing (ICONIP). pp. 139–149. Springer (2019)
44. Nwoye, C.I., Gonzalez, C., Yu, T., Mascagni, P., Mutter, D., Marescaux, J., Padoy, N.: Recognition of instrument-tissue interactions in endoscopic videos via action triplets. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part III 23. pp. 364–374. Springer (2020)
45. Nwoye, C.I., Padoy, N.: Data splits and metrics for benchmarking methods on surgical action triplet datasets. arXiv preprint arXiv:2204.05235 (2022)
46. Nwoye, C.I., Yu, T., Gonzalez, C., Seeliger, B., Mascagni, P., Mutter, D., Marescaux, J., Padoy, N.: Rendezvous: Attention mechanisms for the recognition of surgical action triplets in endoscopic videos. Medical Image Analysis **78**, 102433 (2022)
47. Pan, X., Gao, X., Wang, H., Zhang, W., Mu, Y., He, X.: Temporal-based swin transformer network for workflow recognition of surgical video. Int. J. Comput. Assist. Radiol. Surg. **18**(1), 139–147 (Jan 2023)
48. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
49. Rasheed, H., khattak, M.U., Maaz, M., Khan, S., Khan, F.S.: Finetuned clip models are efficient video learners. In: The IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023)
50. Ross, T., Reinke, A., Full, P.M., Wagner, M., Kenngott, H., Apitz, M., Hempe, H., Mindroc-Filimon, D., Scholz, P., Tran, T.N., et al.: Comparative validation of multi-instance instrument segmentation in endoscopy: results of the robust-mis 2019 challenge. Medical image analysis **70**, 101920 (2021)
51. Ross, T., Zimmerer, D., Vemuri, A., Isensee, F., Wiesenfarth, M., Bodenstedt, S., Both, F., Kessler, P., Wagner, M., Müller, B., et al.: Exploiting the potential of unlabeled endoscopic video data with self-supervised learning. International journal of computer assisted radiology and surgery **13**, 925–933 (2018)
52. Roß, T., Reinke, A., Full, P.M., Wagner, M., Kenngott, H., Apitz, M., Hempe, H., Mindroc-Filimon, D., Scholz, P., Tran, T.N., Bruno, P., Arbeláez, P., Bian, G.B., Bodenstedt, S., Bolmgren, J.L., Bravo-Sánchez, L., Chen, H.B., González, C., Guo, D., Halvorsen, P., Heng, P.A., Hosgor, E., Hou, Z.G., Isensee, F., Jha, D., Jiang, T., Jin, Y., Kirtac, K., Kletz, S., Leger, S., Li, Z., Maier-Hein, K.H., Ni, Z.L., Riegler, M.A., Schoeffmann, K., Shi, R., Speidel, S., Stenzel, M., Twick, I., Wang, G., Wang, J., Wang, L., Wang, L., Zhang, Y., Zhou, Y.J., Zhu, L., Wiesenfarth, M., Kopp-Schneider, A., Müller-Stich, B.P., Maier-Hein, L.: Comparative validation of multi-instance instrument segmentation in endoscopy: Results of the robust-mis 2019 challenge. Medical Image Analysis **70**, 101920 (2021). https://doi.org/https://doi.org/10.1016/j.media.2020.101920, https://www.sciencedirect.com/science/article/pii/S136184152030284X
53. Sato, F., Hachiuma, R., Sekii, T.: Prompt-guided zero-shot anomaly action recognition using pretrained deep skeleton features. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6471–6480 (2023)
54. Schoeffmann, K., Husslein, H., Kletz, S., Petscharnig, S., Münzer, B., Beecks, C.: Video retrieval in laparoscopic video recordings with dynamic content descriptors.

Multim. Tools Appl. **77**(13), 16813–16832 (2018). https://doi.org/10.1007/s11042-017-5252-2, https://doi.org/10.1007/s11042-017-5252-2
55. Schoeffmann, K., Taschwer, M., Sarny, S., Münzer, B., Primus, M.J., Putzgruber, D.: Cataract-101: video dataset of 101 cataract surgeries. In: César, P., Zink, M., Murray, N. (eds.) Proceedings of the 9th ACM Multimedia Systems Conference, MMSys 2018, Amsterdam, The Netherlands, June 12-15, 2018. pp. 421–425. ACM (2018). https://doi.org/10.1145/3204949.3208137, https://doi.org/10.1145/3204949.3208137
56. 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)
57. Shi, X., Jin, Y., Dou, Q., Heng, P.A.: LRTD: long-range temporal dependency based active learning for surgical workflow recognition. Int. J. Comput. Assist. Radiol. Surg. **15**(9), 1573–1584 (Sep 2020)
58. Sigurdsson, G.A., Varol, G., Wang, X., Farhadi, A., Laptev, I., Gupta, A.: Hollywood in homes: Crowdsourcing data collection for activity understanding. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) Computer Vision – ECCV 2016. pp. 510–526. Springer International Publishing, Cham (2016)
59. Smedsrud, P.H., Thambawita, V., Hicks, S.A., Gjestang, H., Nedrejord, O.O., Næss, E., Borgli, H., Jha, D., Berstad, T.J.D., Eskeland, S.L., Lux, M., Espeland, H., Petlund, A., Nguyen, D.T.D., Garcia-Ceja, E., Johansen, D., Schmidt, P.T., Toth, E., Hammer, H.L., de Lange, T., Riegler, M.A., Halvorsen, P.: Kvasir-Capsule, a video capsule endoscopy dataset. Scientific Data **8**(1), 142 (2021). https://doi.org/10.1038/s41597-021-00920-z
60. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild (2012)
61. Spaeth, G., Danesh-Meyer, H., Goldberg, I., Kampik, A.: Ophthalmic Surgery: Principles and Practice E-Book. Elsevier Health Sciences (2011), https://books.google.com.hk/books?id=wHWMUGH-5csC
62. Stauder, R., Ostler, D., Kranzfelder, M., Koller, S., Feußner, H., Navab, N.: The tum lapchole dataset for the m2cai 2016 workflow challenge. arXiv preprint arXiv:1610.09278 (2016)
63. Tian, Y., Pang, G., Liu, F., Liu, Y., Wang, C., Chen, Y., Verjans, J., Carneiro, G.: Contrastive transformer-based multiple instance learning for weakly supervised polyp frame detection. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 88–98. Springer (2022)
64. Tran, D., Wang, H., Torresani, L., Feiszli, M.: Video classification with channel-separated convolutional networks (2019)
65. Twinanda, A.P., Shehata, S., Mutter, D., Marescaux, J., de Mathelin, M., Padoy, N.: Endonet: A deep architecture for recognition tasks on laparoscopic videos. IEEE Transactions on Medical Imaging (TMI) **36**, 86–97 (2016), https://api.semanticscholar.org/CorpusID:5633749
66. Twinanda, A.P., Yengera, G., Mutter, D., Marescaux, J., Padoy, N.: Rsdnet: Learning to predict remaining surgery duration from laparoscopic videos without manual annotations. IEEE Transactions on Medical Imaging **38**(4), 1069–1078 (2019). https://doi.org/10.1109/TMI.2018.2878055
67. Wagner, M., Müller-Stich, B.P., Kisilenko, A., Tran, D., Heger, P., Mündermann, L., Lubotsky, D.M., Müller, B., Davitashvili, T., Capek, M., et al.: Comparative validation of machine learning algorithms for surgical workflow and skill analysis with the heichole benchmark. Medical Image Analysis **86**, 102770 (2023)

68. Wang, T., Li, H., Pu, T., Yang, L.: Microsurgery robots: Applications, design, and development. Sensors **23**(20) (2023). https://doi.org/10.3390/s23208503, https://www.mdpi.com/1424-8220/23/20/8503
69. Wang, X., Zhang, S., Qing, Z., Shao, Y., Gao, C., Sang, N.: Self-supervised learning for semi-supervised temporal action proposal. In: CVPR (2021)
70. Wang, Z., Lu, B., Long, Y., Zhong, F., Cheung, T.H., Dou, Q., Liu, Y.: Autolaparo: A new dataset of integrated multi-tasks for image-guided surgical automation in laparoscopic hysterectomy. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 486–496. Springer (2022)
71. Wu, W., Sun, Z., Ouyang, W.: Revisiting classifier: Transferring vision-language models for video recognition (2023)
72. Wu, W., Wang, X., Luo, H., Wang, J., Yang, Y., Ouyang, W.: Bidirectional cross-modal knowledge exploration for video recognition with pre-trained vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023)
73. Yengera, G., Mutter, D., Marescaux, J., Padoy, N.: Less is more: Surgical phase recognition with less annotations through self-supervised pre-training of cnn-lstm networks. arXiv preprint arXiv:1805.08569 (2018)
74. Yu, F., Silva Croso, G., Kim, T.S., Song, Z., Parker, F., Hager, G.D., Reiter, A., Vedula, S.S., Ali, H., Sikder, S.: Assessment of Automated Identification of Phases in Videos of Cataract Surgery Using Machine Learning and Deep Learning Techniques. JAMA Network Open **2**(4), e191860–e191860 (04 2019). https://doi.org/10.1001/jamanetworkopen.2019.1860, https://doi.org/10.1001/jamanetworkopen.2019.1860
75. Yu, T., Mutter, D., Marescaux, J., Padoy, N.: Learning from a tiny dataset of manual annotations: a teacher/student approach for surgical phase recognition. arXiv preprint arXiv:1812.00033 (2018)
76. Yuan, K., Srivastav, V., Navab, N., Padoy, N.: Hecvl: Hierarchical video-language pretraining for zero-shot surgical phase recognition. arXiv preprint arXiv:2405.10075 (2024)
77. Yuan, K., Srivastav, V., Yu, T., Lavanchy, J., Mascagni, P., Navab, N., Padoy, N.: Learning multi-modal representations by watching hundreds of surgical video lectures (2023)
78. Zha, R., Cheng, X., Li, H., Harandi, M., Ge, Z.: Endosurf: Neural surface reconstruction of deformable tissues with stereo endoscope videos. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 13–23. Springer (2023)
79. Zhang, C.L., Wu, J., Li, Y.: Actionformer: Localizing moments of actions with transformers. In: European Conference on Computer Vision. LNCS, vol. 13664, pp. 492–510 (2022)
80. Zhou, L., Xu, C., Corso, J.J.: Towards automatic learning of procedures from web instructional videos. In: AAAI Conference on Artificial Intelligence. pp. 7590–7598 (2018), https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17344