UNIVERSITÄT KLAGENFURT

Deep-Learning-Assisted Analysis of Cataract Surgery Videos

Negin Ghamsarian

Dissertation

zur Erlangung des akademischen Grades
Doktor der Technischen Wissenschaften

Doktoratsstudium der Technischen Wissenschaften
im Dissertationsgebiet Informatik

Alpen-Adria-Universität Klagenfurt
Fakultät für Technische Wissenschaften

Erstbetreuer
Assoc. Prof. Dr. Klaus Schöffmann
Institut für Informationstechnologie
Alpen-Adria-Universität Klagenfurt

Zweitbetreuer
Assoc. Prof. Dr. Christian Timmerer
Institut für Informationstechnologie
Alpen-Adria-Universität Klagenfurt

Erstgutachter
Prof. Dr. Henning Müller
Institute of Information Systems
University of Applied Sciences Western Switzerland
Department of Radiology and Medical Informatics
University of Geneva

Zweitgutachter
Prof. Dr. Raphael Sznitman
ARTORG Center for Biomedical Engineering
Faculty of Medicine
University of Bern

Klagenfurt am Wörthersee, 27.9.2021

Negin Ghamsarian Klagenfurt am Wörthersee, 27.9.2021

Eidesstattliche Erklärung

Ich versichere an Eides statt, dass ich

- die eingereichte wissenschaftliche Arbeit selbstständig verfasst und keine anderen als die angegebenen Hilfsmittel benutzt habe,
- die während des Arbeitsvorganges von dritter Seite erfahrene Unterstützung, einschließlich signifikanter Betreuungshinweise, vollständig offengelegt habe,
- die Inhalte, die ich aus Werken Dritter oder eigenen Werken wortwörtlich oder sinngemäß übernommen habe, in geeigneter Form gekennzeichnet und den Ursprung der Information durch möglichst exakte Quellenangaben (z.B. in Fußnoten) ersichtlich gemacht habe,
- die eingereichte wissenschaftliche Arbeit bisher weder im Inland noch im Ausland einer Prüfungsbehörde vorgelegt habe und
- bei der Weitergabe jedes Exemplars (z.B. in gebundener, gedruckter oder digitaler Form) der wissenschaftlichen Arbeit sicherstelle, dass diese mit der eingereichten digitalen Version übereinstimmt.

Mir ist bekannt, dass die digitale Version der eingereichten wissenschaftlichen Arbeit zur Plagiatskontrolle herangezogen wird.
Ich bin mir bewusst, dass eine tatsachenwidrige Erklärung rechtliche Folgen haben wird.

Negin Ghamsarian e. h. Klagenfurt am Wörthersee, 27.9.2021

Affidavit

I hereby declare in lieu of an oath that

- the submitted academic paper is entirely my own work and that no auxiliary materials have been used other than those indicated,
- I have fully disclosed all assistance received from third parties during the process of writing the thesis, including any significant advice from supervisors,
- any contents taken from the works of third parties or my own works that have been included either literally or in spirit have been appropriately marked and the respective source of the information has been clearly identified with precise bibliographical references (e.g. in footnotes),
- to date, I have not submitted this paper to an examining authority either in Austria or abroad and that
- when passing on copies of the academic thesis (e.g. in bound, printed or digital form), I will ensure that each copy is fully consistent with the submitted digital version.

I understand that the digital version of the academic thesis submitted will be used for the purpose of conducting a plagiarism assessment.
I am aware that a declaration contrary to the facts will have legal consequences.

Negin Ghamsarian e. h. Klagenfurt am Wörthersee, 27.9.2021

Acknowledgments

I take pride in working under the supervision of professor Klaus Schöffmann and taking advantage of his invaluable guidance during my doctoral studies. I was very fortunate to work with an open-minded, kind-hearted, approachable, and knowledgeable supervisor, who taught me how to develop critical thinking, work independently, and think on my feet. I would like to express my sincere gratitude for his empathy and compassionate leadership, especially during the hard time of the COVID pandemic.

I would like to thank the examiners of my dissertation, professor Henning Müller and professor Raphael Sznitman, for their time and precious feedback.

My heartfelt thanks should also go to the administrative staff of the ITEC department not only for their prompt support but also for their kindness and companionship during my work at Klagenfurt University.

Finally, I would like to thank my mother for her endless moral support and believing in me when no one believes in me, and thank my father for always motivating me to go the extra mile and encouraging me not to give up. No words can convey my profound gratitude to them for consistently supporting me through thick and thin.

Thank you!

Abstract

Following the technological advancements in medicine, the operation rooms are evolving into intelligent environments. The context-aware systems (CAS) can comprehensively interpret the surgical state, enable real-time warning, and support decision-making, especially for novice surgeons. These systems can automatically analyze surgical videos and perform indexing, documentation, and post-operative report generation. The ever-increasing demand for such automatic systems has sparked machine-learning-based approaches for surgical video analysis. This thesis addresses the significant challenges in cataract surgery video analysis to pave the way for building efficient context-aware systems. The main contributions of this thesis are five folds: (1) This thesis demonstrates that spatio-temporal localization of the relevant content can considerably improve phase recognition accuracy. (2) This thesis proposes a novel deep-learning-based framework for relevance-based compression to enable real-time streaming and adaptive storage of cataract surgery videos. (3) Several convolutional modules are proposed to boost the networks' semantic interpretation performance in challenging conditions. These challenges include blur and reflection distortion, transparency, deformability, color and texture variation, blunt edges, and scale variation. (4) This thesis proposes and evaluates the first framework for automatic irregularity detection in cataract surgery videos. (5) To alleviate the requirement for manual pixel-based annotations, this thesis proposes novel strategies for self-supervised representation learning adapted to semantic segmentation.

Contents

List of Figures

List of Tables

Chapter 1 Introduction

Chapter overview — This chapter starts with a brief description of cataract surgery. Afterwards, I delineate the rational behind computerized analysis of cataract surgery, which is the focus of this dissertation. The research questions are then introduced in this chapter. Finally, the last section gives a summary of the research subjects studied in the course of this thesis using an initial pipeline of a cataract surgery exploration system.

This chapter is an adapted version of:

"Ghamsarian, N. Enabling relevance-based exploration of cataract videos. In Proceedings of the 2020 International Conference on Multimedia Retrieval (New York, NY, USA, 2020), ICMR '20, Association for Computing Machinery, p. 378–382."

1.1 Cataract Surgery

Cataract refers to the opacity and occlusion of the eye's natural lens due to the process of aging, eye inflammation, congenital problems, and other issues. (Figure 1.1). This natural lens' cloudiness causes vision deterioration and transparency degradation, but also blindness in the case of dense occlusion. Color perception distortion due to the lens' yellowing, double vision (ghosting), and blurred vision are the most common symptoms of cataracts (Figure 1.2). Cataract surgery is the procedure of returning a clear vision to the eye by removing the occluded lens, followed by implanting an artificial lens named intraocular lens (IoL). Involving over 100 million

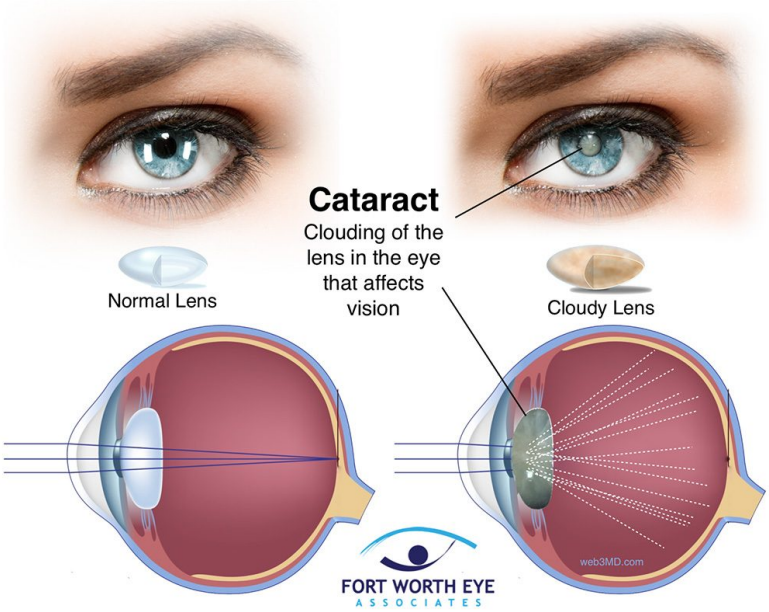


Figure 1.1: Cataract as the eye's natural lens, having become cloudy and causing vision deterioration (picture from https://www.ranelle.com/cataract-surgery/).

people worldwide, cataract surgery is the causative factor of a substantial fraction of worldwide blindness [240].

Cataract surgery is not only the most frequent ophthalmic surgery [31] but also one of the most common surgeries worldwide [231]. This operation is conducted with the aid of a binocular microscope that provides a three-dimensional magnified and illuminated image of the eye for accurately watching the patient's eye (Figure 1.3). The microscope contains a mounted camera, which records and stores the whole surgery for several post-operative objectives.

1.2 Motivation

Due to the growing demand for cataract surgery, this surgery significantly impacts the patient's quality of life. Accordingly, a large body of research is devoted to computerized workflow analysis in this surgery to improve surgical outcomes and diminish potential surgical risks. This objective can be achieved through (1) alleviating the training procedure for junior surgeons via enabling content-aware explorations, (2)

Figure 1.2: Cataract symptoms including color perception distortion, double vision, and blurred vision (picture from `https://www.mathworks.com/company/mathworks-stories/`).

facilitating the storage and streaming of cataract surgery videos, (3) and detecting the unexplored risk factors and irregularities through computerized investigations. In the following, the importance of each contribution is justified.

1.2.1 Training Procedure Alleviation

During the operation of cataract surgery, only two trainees can watch the real-time surgery throughout the teaching binoculars. Thus, the major part of the training procedure is conducted using the videos recordings of cataract surgery. A faster training process results in reducing complications during and after surgeries and diminishing the surgical risks for less-experienced surgeons. As a concrete example,

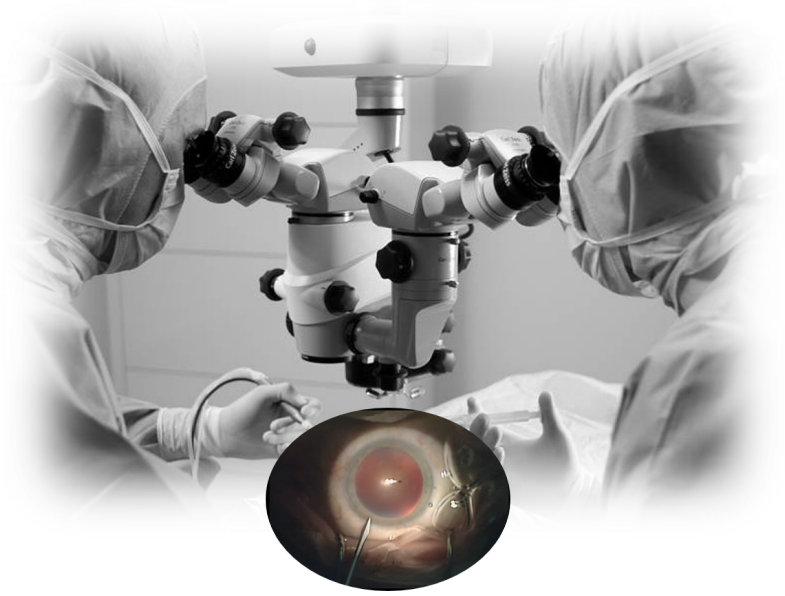


Figure 1.3: The binocular microscope that is used during the surgery to enable accurately watching the patient's eye.

the risk of developing *reactive corneal edema* after surgery for novice surgeons is reported to be 1.6 times higher than that for experienced expert surgeons [147]. Accordingly, it is vital to accelerate the teaching and training process for a surgical technique. Training new surgeons as one of the major duties of experienced expert surgeons demands a considerable supervisory investment. To expedite the training process and subsequently reduce the extra workload on their tight schedule, surgeons are seeking a surgical video retrieval system [146] (Figure 1.4). Automatic workflow analysis approaches can optimize the training procedure by indexing the surgical video segments for online video exploration [157, 204, 162].

1.2.2 Storage and Streaming Facilitation

Recorded cataract surgery videos play a prominent role in training, investigating the surgery, and enhancing surgical outcomes. Due to storage limitations in hospitals, however, the recorded cataract surgeries are typically deleted after a short time, and this precious source of information cannot be fully utilized. Moreover, these videos can be exploited for knowledge exchange among hospitals and surgeons using

remote online exploration. Online exploration usually imposes quality degradation due to limited bandwidth. Lowering the quality for reducing the required storage space or streaming is not advisable since the degraded visual quality results in the loss of relevant information that limits the usage of these videos. Relevance-based compression of cataract surgery videos is the best way to simultaneously provide high quality for the relevant content and low bitrate. This technique entails spatio-temporal relevance detection (*i.e.*, phase detection and semantic segmentation) in cataract surgery videos.

1.2.3 Irregularity Detection

Notwithstanding the numerous advancements in surgical tools and techniques, there are still some intra-operative and post-operative complications in cataract surgery requiring real-time or large-scale computerized workflow analysis.
Irregularity detection in surgical videos can serve two major purposes:

1. By studying the correlations between the intra-operative irregularities and post-operative complications, the surgeons will be able to assess risk factors associated with different complications. Such an advanced knowledge can result in achieving optimal post-operative results.

2. Irregularities in different phases are of great importance for the surgeons in terms of teaching. Indeed, surgical training on irregularities plays a key role in surgical competency enhancement [1].

Pupil reactions and IoL instability, unfolding delay, and rotation are the major unexplored implications in cataract surgery.

[1]Surgical competency is defined as the required skill level to perform a safe surgery indepently. Surgical competency entails not only the technical skill to accurately perform the regular surgical phases, but also the knowledge and judgment required to safely deal with irregularities during surgery [236]

Pupil reactions: During the phacoemulification phase, where the occluded natural lens is corrupted and suctioned, the amount of light received by photoreceptors may suddenly increase. This increase in light reception affects the size of pupil, usually resulting in slow (gradual) pupil contraction. In some cases, however, the pupil unexpectedly reacts to the lighting changes and becomes quickly contracted. This sudden reactions in pupil size can lead to serious intra-operative implications. Especially during phacoemulification phase where the instrument is deeply inserted inside the eye, sudden changes of pupil size may lead to injuries to the eye's tender tissues. Pupil reaction is regarded as a major intra-operative issue, leading to serious implications during surgery. Real-time automatic detection of pupil reactions during phacoemulification phase can highly contribute to a safer surgical procedure as well as providing important insight for further post-operative investigations [216]. Detecting this irregularity involves phacoemulification phase recognition and pupil/iris segmentation.

IoL irregularity: A critical complication after cataract surgery is the dislocation of the lens implant leading to vision deterioration and eye trauma. In order to reduce the risk of this complication, it is vital to discover the risk factors during the surgery. It is argued that there is a direct relationship between the lens irregularities during and after surgery. Especially, the surgeons believe that lens unfolding delay, instability, and rotation during surgery are some of the symptoms of lens dislocation after surgery. Besides, the surgeons claim that these irregularities may bound with some brands of the IoL. However, studying the relationship between lens dislocation and its suspicious risk factors using numerous videos is a time-extensive procedure. Hence, the surgeons demand an automatic approach to enable large-scale investigations of this problem. Such study involves implantation phase detection and IoL/pupil segmentation.

1.3 Research Questions

Considering the major demands for enhancing the recent approaches or investigating unexplored research subjects, this thesis tries to address the following four research questions:

- Can the accuracy of automatic operation phase recognition be improved upon state-of-the-art techniques using deep learning approaches?
- How and to what extent can we reduce the required bitrate for cataract surgery videos while preserving high quality of the relevant content?
- How reliably can unspecified irregularities be automatically detected in ophthalmic surgery videos?
- How can the semantic segmentation performance for the relevant objects in cataract surgery be improved?

1.4 Research Overview

Aiming to enhance the surgical outcomes and diminish the clinical risks, there is a great desire for context-aware retrieval systems. Such a system is particularly advantageous for training amateur and less-skillful surgeons by representing the relevant content from the surgeons' eyes. One of the aims of this doctoral project is to provide the basis for a cataract video exploration system, that is able to automatically analyze and extract the relevant segments of videos from cataract surgery. Indeed, in this doctoral project, content analysis and retrieval methods for recorded cataract surgery videos will be investigated to meet the requirements for building a content exploration system. The system is going to be used by clinicians as a teaching means to communicate operation techniques and complications to trainee surgeons. Hence, such a system should be able to automatically filter the

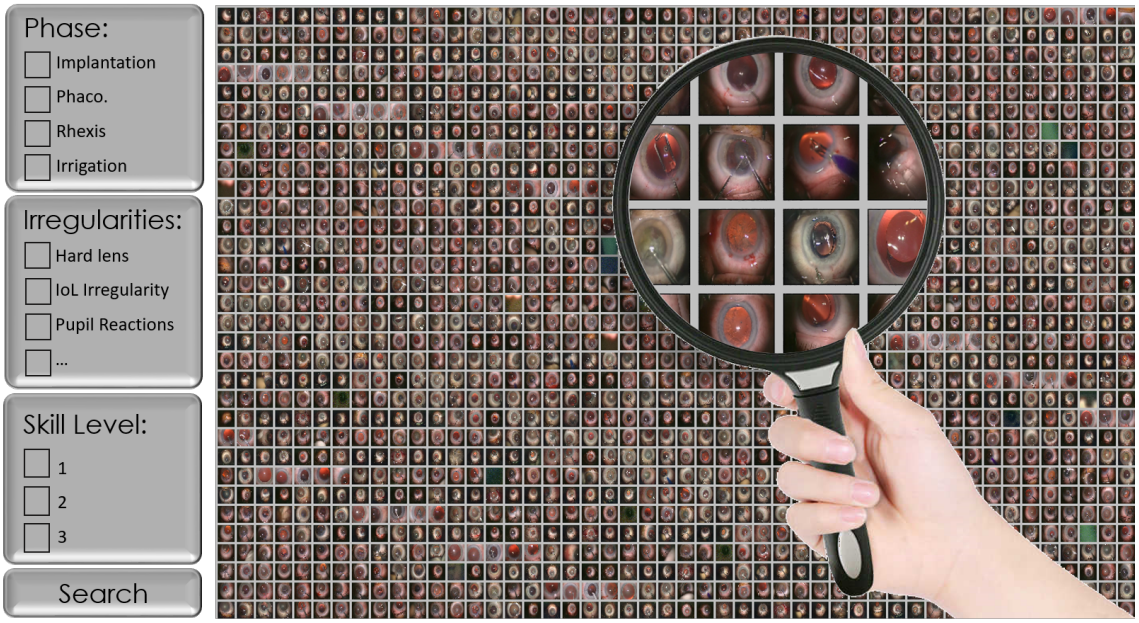


Figure 1.4: A cataract surgery exploration system can substantially optimize the training procedure by enabling relevance-based retrieval.

irrelevant video segments and classify the relevant content (e.g., specific operation tools, operation phases, and surgical actions).

Typically, the recorded cataract videos do not contain any meta-data and are basically one-shot videos, without any editing. Moreover, the content of different operation phases is highly similar and suffers from the following problems:

1. Since the camera focus is manually adjusted by the surgeons as a pre-surgical step, the recorded videos may contain defocus blur.

2. Unconscious eye movements, as well as fast instrument motions, may lead to harsh motion blur in the salient segments of each frame.

3. The operation instruments used in different phases are highly similar in terms of visual appearance, resulting in a narrow inter-class deviation.

Designing an approach to classify the relevant content in such videos, therefore, is quite challenging. Based on the mentioned challenges, we have designed the first pipeline of cataract surgery exploration system as shown in Figure 1.5. The proposed

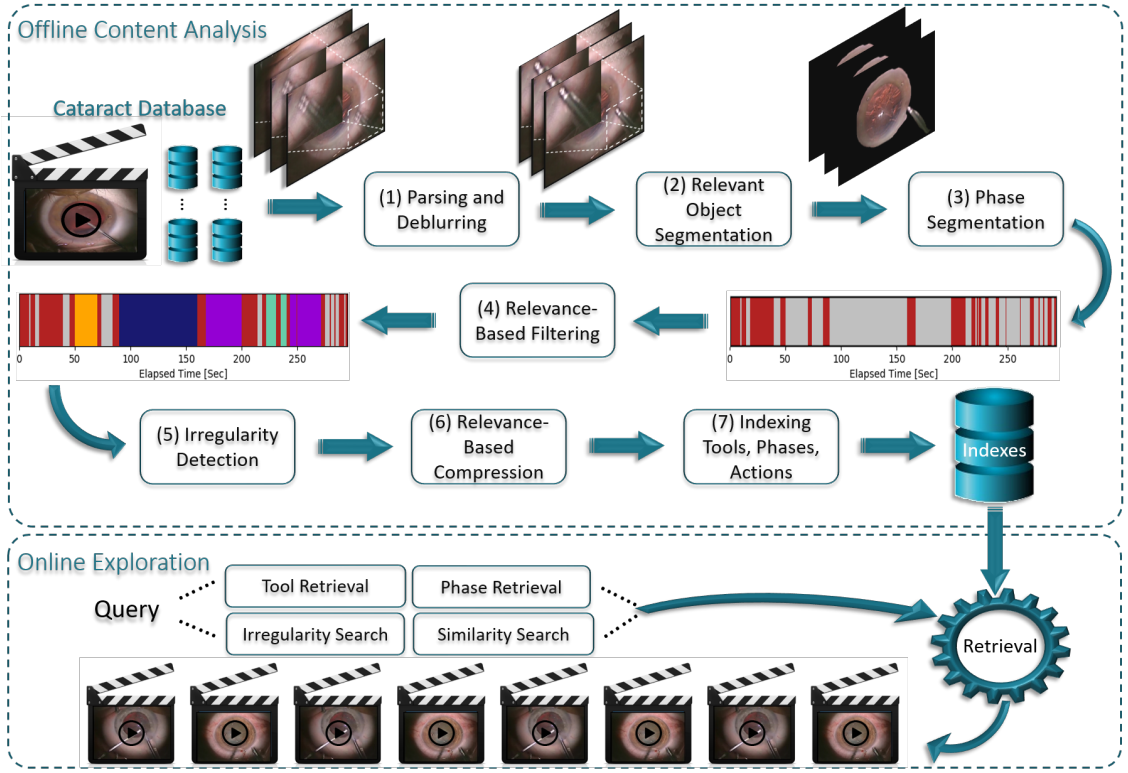


Figure 1.5: The pipeline of the proposed cataract surgery exploration system consisting of an offline content analysis module and an online exploration module.

pipeline consists of an offline content analysis module and an online exploration module. The resulting system allows the user to retrieve tools, phases, and irregulities, as well as performing content-based similarity search. However, the focus of this work is on the content analysis, since another companion doctoral project in our research group will focus on the indexing methods and the interface to use the results achieved herein.

The offline content analysis module consists of several steps. The cataract surgery videos being impaired by defocusing and motion blur are deblurred in the first stage, prior to feeding them into a neural network for our task of interest. The relevant segments of each frame are then extracted using pixel-based or region-based segmentation networks.

In the next step, a CNN for idle frame recognition is employed to retain only the segments showing some actions (i.e., performed with some instruments) and completely ignore the content without any instruments. Afterwards, the segmented

frames are fed into the networks being accounted to detect different relevant phases or irregularities. The outputs of these stages will be further exploited for indexing the segments of cataract surgery videos and provided to the user of the online exploration system. The mentioned steps will be described in details in the following.

1.4.1 Parsing and Deblurring

Neural networks recognize the edges in the first layers to combine them to complicated and meaningful content in the last layers. Thus, neural networks suffer from performance degradation when training on blurry inputs [235]. This is the reason why cataract surgery videos should be deblurred as a pre-processing step to retrieve the diluted high-frequency features and acquire the underlying sharp frames as the inputs for our neural networks.

In Chapter 3, we address the low visual quality of surgical videos using a multi-scale deconvolutional neural network inspired by [179]. The network architecture is rooted in the cross-scale patch recurrence theory: "the small patches in a high-resolution image usually recurr in its ideally downscaled versions", a property that is frequently exploited for super-resolution. Cross-scale patch recurrence is also used for deblurring since the small patches in an ideally downsampled blurry image tend to be similar to the small patches of the original sharp image rather than the blurry one.

The proposed network includes three subgraphs, each undertaking the task of deblurring the downsampled input image and outputting a residual frame. The first subgraph deblurs the input frame downsampled at 1/4 of the initial resolution and outputs a residual frame to counteract the blurriness impairing the input. The sharp downsampled image is then fed to the next subgraph to perform superresolution and retrieve the underlying sharp image at scale 1/2. Similarly, the last subgraph is responsible for outputting a sharp frame with the original resolution.

We prove that using skip connections in each subgraph thanks to defining same number of filters for the consecutive layers can effectively boost deblurring performance.

Besides, we model defocus blur using gaussian filters and show that a model trained to deblur the gaussian-blur-degraded frames can deal with defocus blur. More details of this approach are described in [173].

1.4.2 Relevant Object Segmentation

Semantic segmentation in surgical videos is a prerequisite for a broad range of applications towards improving surgical outcomes and surgical video analysis including but not limited to workflow analysis, compression, and irregularity detection. Besides, neural networks generally suffer from redundant input information. In other words, the less redundant information we provide to the network, the better performance we can expect [92]. Relevant object segmentation offers a solution to this problem. Depending on the task (relevance detection or a particular irregularity detection), the relevant segments of each frame, being informative for classification are extracted using pixel-based segmentation networks. In the case of phase recognition, the background being considered as the redundant part of each frame is then replaced with black pixels (Figure 1.6).

In cataract surgery, various features of the relevant objects such as blunt edges, color and context variation, reflection, transparency, and motion blur pose a challenge for semantic segmentation. In particular, motion blur and reflection distortion in instruments, and color and texture variations in the other relevant objects lead to distant semantic respresentations for the same semantic labels, and close semantic representations for different semantic labels. In Chapter 7, we propose a novel feature-map callibration module to enhance the semantic representation performance. The proposed module callibrates the feature maps considering multi-angle local features centering around each pixel position, and region-channel inter-dependencies. In Chapter 8, we propose a novel semantic segmentation network termed as "Deep-Pyram", which can deal with various semantic segmentation challenges in cataract surgery videos using three proposed modules: (i) Pyramid View Fusion, (ii) De-

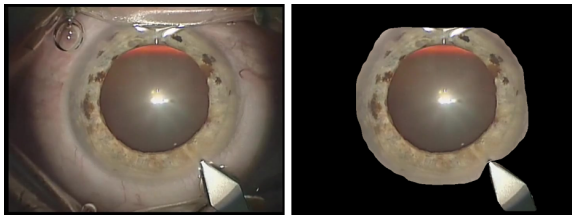


Figure 1.6: Detecting the spatially relevant segments and removing the substantial redundant information using semantic segmentation networks for a representative frame.

formable Pyramid Reception, and (iii) Pyramid Loss. The experimental evaluations confirm the superiority of the DeepPyram over state-of-the-art rival approaches.

Due to many reasons, atificial intelligence in medical domain has lagged behind other domains. First and foremost, the supervised machine learning approaches depend heavily on manual annotations, which requires expert knowledge when it comes to medical images and videos. Thus supervised learning in medical domain is more time extensive and costly. Secondly, the pre-trained weights of convolutional neural networks using larg-scale image datasets are not adaptive initializations due to the large distribution gap between the medical and natural images. On top of that, the rapid advancements in technological tools and instruments inevitably entails expeditious distribution shift in raw data. Self-supervised learning suggests ultimate solutions to alleviate the mentioned problems. In Chapter 9, we propose three self-supervised strategies to encourage representation learning adapted to semantic segmentation in cataract surgery videos. Specifically, we provide a moderate-to-hard representation learning task to bridge the gap between human and neural network semantic interpretation.

1.4.3 Phase Segmentation

In this stage, a CNN networks is employed to perform two-class classification on the input frames and categorize them as *idle* or *action*. Idle frames refer to the frames in which no instrument is visible. This gives us a clue that the surgeon is changing the instrument and accordingly a new phase will begin. In other words, each action

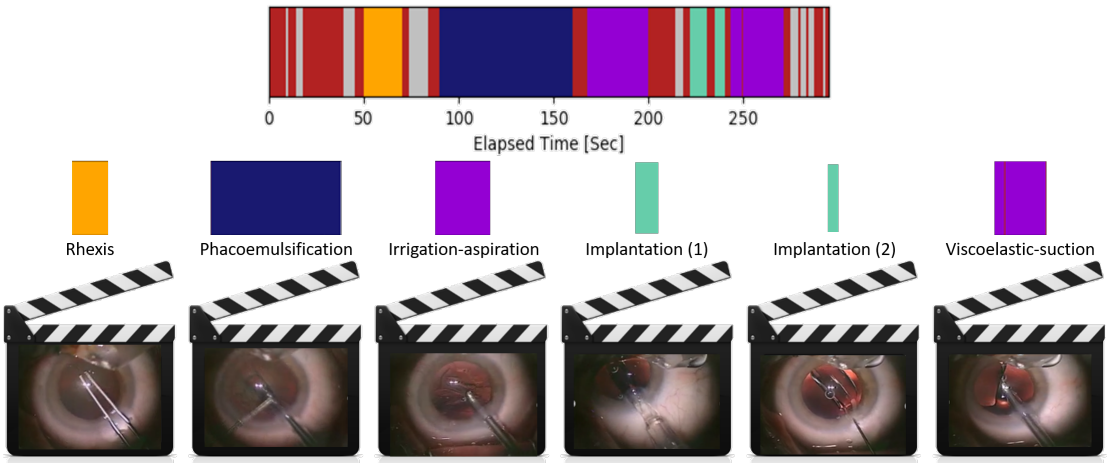


Figure 1.7: Relevence Detection Results of the proposed CNN-RNN framework for a representative cataract surgery video.

phase in a cataract surgery video is delimited by two idle phases. By having a whole segment of a phase, we can exploit and accumulate the features coming from the network for every single frame to boost the accuracy in phase recognition.

1.4.4 Relevance-Based Filtering

A regular cataract surgery video consists of eleven phases: incision, hydrodissection, rhexis, viscoelastic, phacoemulsification, irrigation-aspiration, capsule polishing, lens implantation, viscoelastic-suction, tonifying, and antibiotics. To investigate and teach the cataract surgery, however, not all these phases are relevant (rhexis, phacoemulsification, irrigation-aspiration, and lens implantation are considered relevant by clinicians). Hence, an automatic retrieval approach is required to segment the video and output the relevant segments.

Chapter 4 is devoted to relevance detection in cataract surgery videos using CNN-RNNs. We firstly segment the distinct phases using a static CNN. The label of each segmented phase is then determined independently and disregarding the information from previous and following phases. Such an strategy is particularly useful in the case of irregularities in catarcat surgery videos, where the networks trained on the sequence of consecutive regular phases usually fail to predict the right label. We also

prove that removing the redundant spatial information and increasing the resolution of the relevant spatial content can significantly enhance phase recognition accuracy, especially in the case of visually similar instruments. Our model is able to detect the relevant phases with very high temporal resolution compared to its counterparts.

1.4.5 Irregularity Detection

A surgical video exploration system should be capable of identifying the irregular phases to exclude many redundant regular surgeries, enable irregularity search, and provide scalable statistical analysis. Chapter 6 introduces a framework to detect two suspicious sympthoms of lens relocation after surgery. Specifically, we focus on computing IoL unfolding delay, instability, and rotation. This requires detecting the lens implantation phase and segmenting the pupil and IoL with high accuracy. We, therefore, propose a CNN-RNN to detect the implantation phase, and a novel U-Net-Based architecture termed as Adapt-Net to segment the pupil and IoL. The proposed segmentation network can be adapted to scale and transparency of IoL by fusing the sequential convolutional-block response maps. Moreover, Adapt-Net can deal with shape variations and unpredictable deformations during unfolding by fusing the deformable-filters and structured-filters response maps.

1.4.6 Relevance-Based Compression

In Chapter 5, we introduce a novel framework for relevance-based compression of cataract surgery videos. The proposed framework consists of two modules: (i) a relevance detection module, which is designed to detect the spatio-temporally relevant content based on five differenr scenarios using static and region-based CNNs, and (ii) a compression module that is responsible for assigning low bitrate to irrelevant content to enhance compression ratio, while preserving the high quality of the relevant content.

1.5 Publications in the Context of the Dissertation Topic

In the following, a list of all publications that are directly related to the topic of this thesis is given in chronological order:

- "Ghamsarian, N. Enabling Relevance-Based Exploration of Cataract Videos. In Proceedings of the 2020 International Conference on Multimedia Retrieval (New York, NY, USA, 2020), ICMR '20, Association for Computing Machinery, p. 378–382."
- "Ghamsarian, N., Taschwer, M., and Schoeffmann, K. Deblurring Cataract Surgery Videos Using a Multi-Scale Deconvolutional Neural Network. In 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI) (2020), pp. 872–876."
- "Ghamsarian, N., Amirpourazarian, H., Timmerer, C., Taschwer, M., and Schoffmann, K. ¨ Relevance-Based Compression of Cataract Surgery Videos using Convolutional Neural Networks. In Proceedings of the 28th ACM International Conference on Multimedia (New York, NY, USA, 2020), MM '20, Association for Computing Machinery, p. 3577–3585."
- "Ghamsarian, N., Taschwer, M., Putzgruber-Adamitsch, D., Sarny, S., and Schoeffmann, K. Relevance Detection in Cataract Surgery Videos by Spatio-Temporal Action Localization. In 2020 25th International Conference on Pattern Recognition (ICPR) (2021), pp. 10720–10727."
- "Ghamsarian, N., Taschwer, M., Putzgruber-Adamitsch, D., Sarny, S., El-Shabrawi, Y., and Schoeffmann, K. LensID: A CNN-RNN-Based Framework Towards Lens Irregularity Detection in Cataract Surgery Videos. In Medical Image Computing and Computer Assisted Intervention – MICCAI 2021 (Cham, 2021), M. de Bruijne, P. C. Cattin, S. Cotin, N. Padoy, S. Speidel, Y. Zheng, and C. Essert, Eds., Springer International Publishing, pp. 76–86."

- "Ghamsarian, N., Taschwer, M., Putzgruber-Adamitsch, D., Sarny, S., El-Shabrawi, Y., and Schoeffmann, K. Recal-Net: Joint Region-Channel-Wise Calibrated Network for Semantic Segmentation in Cataract Surgery Videos. In 28th International Conference on Neural Information Processing (ICONIP) (2021), p. To Appear."

- "Ghamsarian, N., Taschwer, and Schoeffmann, K. DeepPyram: Enabling Pyramid View and Deformable Pyramid Reception for Semantic Segmentation in Cataract Surgery Videos. Under Review."

Table 1.1 gives an overview of the papers related to this thesis, and my level of contribution in the proposed approach, implementation, evaluation, and writing in each paper.

Table 1.1: Papers.

Subject	Reference	Proposed Approach	Implementation	Evaluation	Writing
Relevance-Based Exploration	[78]	★★★★★	★★★★★	★★★★★	★★★★★
Deblurring	[83]	★★★★★	★★★★★	★★★★★	★★★★★
Relevance Detection	[82]	★★★★★	★★★★★	★★★★★	★★★★★
Relevance-Based Compression	[79]	★★★★★	★★★★☆	★★★☆☆	★★★★★
Lens Irregularity Detection	[80]	★★★★★	★★★★★	★★★★★	★★★★★
Semantic Segmentation (DeepPyram)	[84]	★★★★★	★★★★★	★★★★★	★★★★★
Semantic Segmentation (ReCal-Net)	[81]	★★★★★	★★★★★	★★★★★	★★★★★
Self-Supervised Learning	To be submitted	★★★★★	★★★★★	★★★★★	★★★★★

CHAPTER

2 A Survey on Deep-Learning-Based Surgical Video Analysis

Chapter overview — This chapter summarizes and compares state-of-the-art machine-learning-based approaches towards computer assisted surgery, with a major focus on deep-learning-based approaches.

Computerized surgical workflow analysis and computer-assisted surgery (CAS) are becoming integral parts of medicine. In particular, state-of-the-art deep-learning-based approaches have impacted scheduling and operation planning, intra-operative assessment and postoperative evaluation, patient briefing [159], and surgical outcome estimation [122]. These approaches can be broadly categorized into workflow analysis, instrument recognition and segmentation, and skill assessment. The following sections briefly review the main approaches related to each of the mentioned categories.

2.1 Workflow Analysis

The applications of automatic workflow analysis approaches include but are not limited to:

- Providing real-time guidance to support decision making in the case of training with surgery simulators,
- Predicting the next required tool,

- Determining remaining surgery duration to optimize staff scheduling and reduce patient waiting time,
- Early warnings through anomaly or deviation detection [107],
- Determining the amount of required anesthesia

The existing approaches belonging to workflow analysis can be categorized into two groups: phase recognition and remaining surgery duration.

2.1.1 Phase Recognition

Phase recognition in surgical videos can be broadly divided into classical (feature-extraction-based) and deep-learning-based methods. The first-generation approaches extract hand-engineered features to be further used as the inputs to the classical machine learning methods. Some of these methods exploit hand-crafted features such as texture information, color, and shape [141]. The major discriminative features in different phases of surgical videos are the specific tools used in these phases. Accordingly, many approaches exploit binary instrument usage information [181], RFID tags [20], tool tracking equipment, or built-in sensors [102]. Some methods based on conditional random fields [196], random forests [217], or Hidden Markov Models (HMMs) [34], have taken advantage of tool presence information for phase recognition. Another method exploits Dynamic Time Warping (DTW), Hidden Markov Models (HMMs), and Conditional Random Fields (CRFs) along with the tool presence signals for surgical video analysis [35]. Regarding gesture and action recognition, Zappella *et al.* [260] use linear dynamic systems, bag of features (BoF), and multiple kernel learning (MKL) to classify the pre-segmented video clips.
Hand-crafted features may provide sub-optimal classification results or fail to provide robust classification. Besides, tool usage information relies on additional sensors and devices, which are not ubiquitous, or manual annotations being expensive and time-consuming. Automated feature extraction thanks to deep neural networks have

proved to provide more optimal and robust classifiers for complicated problems. Hence, deep-learning-based approaches have drawn much attention from the medical imaging community in recent years, and many studies in the next generation have been dedicated to automatic workflow recognition using deep-learning-based approaches. Authors of [119] use an integrated CNN-RNN (end to end ResNet+LSTM) network to fully utilize the complementary spatio-temporal features for phase recognition in cholecystectomy videos. To reduce the demand for annotations or compensate for the lack of enough labeled data, another work [72] proposes three self-supervised pre-training methods using temporal coherence: (1) contrastive loss, (2) ranking loss, and (3) first and second-order contrastive loss. Their results provide evidence that a pre-trained model fine-tuned on fewer videos outperforms the baseline trained on more videos. Using Gated Recurrent Unit (GRU) as an alternative for LSTM is also suggested to improve the classification performance [25]. The authors of [251] proposed a "semi-supervised" approach based on "self-supervised pre-training" to predict Remaining Surgery Duration (RSD). EndoNet [232] employs a CNN architecture to carry out phase recognition and tool presence detection in a multi-task manner. It leverages AlexNet trained for tool recognition as the feature extractor for Hierarchical Hidden Markov Model (HHMM) to classify the phases in laparoscopic surgeries. EndoRCN [120] uses the same technique with ResNet50 as the backbone network and recurrent layer instead of HHMM for phase recognition. DeepPhase [280] exploits tool features as well as tool classification results of ResNet152 as the input for RNNs to perform phase recognition. A comparison between the performance of stacked GRU layers [46] and LSTM [101] in phase recognition revealed that GRU is more successful in inferring the phase based on binary tool presence. On the other hand, LSTM performs better when trained on tool features. Authors of [119] propose an end-to-end CNN-LSTM to exploit the correlated spatio-temporal features for phase recognition in cholecystectomy videos. To address transition-sensitivity in predictions, they propose the concept of Prior Knowledge Inference (PKI). In

another study, the most informative region of each frame in laparoscopic videos is extracted using the static version of "Adaptive Whitening Saliency" (AWS) to be used as input to the CNN [160].

Besides the mentioned deep-learning-based approaches for phase recognition in surgical videos, there are many research efforts focusing on action recognition in regular videos, being customizable to the surgical videos. Regarding action recognition in regular videos, DevNet [75] has achieved promising results by adopting a spatio-temporal saliency map. LSTA [220] proposes an attention mechanism to smoothly track the spatially relevant segments in egocentric activities. R(2+1)D [229] introduces a novel spatiotemporal convolutional block to boost the performance in action recognition. To address 3D CNNs' under-performing in the case of insufficient training examples, a gate-shift module is introduced to turn a static lightweight CNN into a spatiotemporal feature extractor [221]. While the aforementioned approaches provide outstanding results, they are specifically designed for action recognition in case of static backgrounds.

2.1.2 Remaining Surgery Duration

High expenditure in the surgical departments of hospitals primarily comes from two problems: (i) underutilization of operation room (OR) resources due to overestimation of surgery duration, and (ii) high patient waiting time due to underestimation of surgery duration [124]. In addition to the financial turnover, imprecise OR occupancy estimation adversely affects the patient's comfort and surgical safety due to increasing the awaiting time or anesthesia and ventilation duration. Accurate estimation of surgery duration is a crucial factor towards precise OR occupancy estimation for optimizing OR scheduling and enabling full utilization of OR capacity.

Surgery duration estimation approaches can be broadly categorized into pre-operative and intra-operative approaches. The first-category approaches use operational factors such as assigned surgical team [124, 125], temporal factors [125], and patient identity

Table 2.1: Comparisons among the deep-learning-based workflow recognition approaches. In the "Dataset" column, "NP" refers to Nonpublic.

Reference	Target Surgery	Proposed Approach	Learning Method	Target Result	Features	Dataset	Year
[120]	Cholecystectomy	Evaluation	Supervised	Phase classification	CNN-LSTM	Cholec80 [232]	2016
[232]	Cholecystectomy	Framework	Supervised	Phase Classification	CNN & SVM & HMM	Cholec80 [232]	2017
[25]	Laparoscopy	Framework	Self-Supervised	Pre-Training with Frame Ordering for Phase Recognition	CNN	NP	2017
[8]	Cholecystectomy	Network	Self-Supervised	Remaining Surgery Duration	CNN-LSTM	Cholec80 [232]	2017
[188]	Gynecology	Evaluation	Supervised	Shot Classification	CNN-LSTM	NP	2017
[119]	Cholecystectomy	Framework	Supervised	Phase Classification	CNN-LSTM	Cholec80 [232]	2018
[160]	Cholecystectomy	Framework	Supervised	Phase Classification	CNN-LSTM	Cholec80 [232]	2018
[58]	Robotic Surgery (da Vinci Surgical System)	Framework	Self-Supervised	Action Classification	LSTM Autoencoder	MISTIC-SL (NP)	2018
[251]	Cholecystectomy	Framework	Self-Supervised	Pre-training with Remaining Surgery Duration for Phase Classification	CNN-LSTM	Cholec120 [1]	2018
[280]	Cataract	Network	Supervised	Joint Instrument Classification & Phase Recognition	CNN-LTSM & CNN-GRU	[9]	2018
[73]	Cholecystectomy	Evaluation	Self-Supervised	Pre-training with Contrastive and Ranking Loss for Phase Classification	CNN-LSTM	Cholec80 [232]	2018
[233]	Cholecystectomy & Bypass	Network	Self-Supervised	Remaining Surgery Duration	CNN-LSTM	Cholec120 [1] & Bypass170 (NP)	2018
[171]	Cataract	Evaluation	Supervised	Phase Classification	CNN	NP	2019
[24]	Cholecystectomy	Framework	Active Learning	Instrument & Phase Classification	Bayesian CNN-LSTM	Cholec80 [232]	2019
[252]	Cholecystectomy & Laparoscopy	Framework	Semi-Supervised	Hard Frame Detection for Phase Classification	CNN	m2cai16-tool [116] & Cholec80 [232]	2019
[255]	Cataract	Evaluation	Supervised	Phase Classification	CNN-RNN	NP	2019

Table 2.2: Comparisons among the deep-learning-based workflow recognition approaches. In the "Dataset" column, "NP" refers to Nonpublic.

Reference	Target Surgery	Proposed Approach	Learning Method	Target Result	Features	Dataset	Year
[123]	Laparoscopy	Network	Supervised	Early Surgery Type Recognition	CNN-LSTM	Laparo425 (NP)	2019
[59]	Robotic Surgery (da Vinci Surgical System)	Framework	Self-Supervised	Action Classification	RNN-Based Generative Model	JIGSAWS [10, 4, 5] & MISTIC-SL (NP)	2019
[133]	Laparoscopic Sigmoidectomy	Evaluation	Supervised	Phase Classification	CNN	NP	2020
[52]	Cholecystectomy	Network	Supervised	Phase Classification	Multi-Stage Temporal CNN	Cholec80 [232] & Cholec51 (NP)	2020
[256]	Cholecystectomy	Framework	Semi-Supervised	Teacher/Student approach for Phase Classification	CNN-biLSTM-CRF & CNN-LSTM	Cholec120 [1]	2020
[121]	Cholecystectomy	Network	Supervised	Joint Phase Classification & Instrument Presence Detection	CNN-LSTM & Relatedness Loss Base on Kullback-Leibler (KL) Divergence	Cholec80 [232]	2020
[178]	Orthopedic Surgery	Framework	Supervised	Phase Classification	CNN-LSTM & Multimodal Training using Elapsed Time	NP	2020
[132]	Laparoscopy	Evaluation	Supervised	Phase Classification & Instrument Segmentation	CNN & U-Net	NP	2020
[57]	Laparoscopy	Network	Supervised	Phase Classification	Tow-Stream Mixed CNN (2D-3D CNN)	m2cai16-workflow [218]	2020
[76]	Robotic Surgery (da Vinci Surgical System)	Framework	Supervised	Action Classification	Reinforcement Learning	JIGSAWS [10, 4, 5]	2020
[194]	Robotic Surgery	Network	Supervised	Action Segmentation	CNN-RNN	JIGSAWS [10, 4, 5] & RIOUS [194] (NP)	2020
[164]	Cataract	Framework	Supervised	Remaining Surgery Duration jointly with Phase Classification and Surgeon Experience Classification	CNN-LSTM	NP	2021

information [257].

Since surgery duration depends on many factors, including surgeon skill level, patient conditions, and intra-operative irregularities, accurate pre-operative surgery duration estimation is impossible. As a concrete example, the duration of cataract surgery depends on many factors, including the surgeon's experience, patient's eye irregularity, surgical risks and complications [3, 142], and cataract hardness. It is reported that even the surgeons and anesthesiologists generally underestimate the surgery duration [230]. Therefore, a possible way to enhance surgery duration estimation is through verbal communications with the surgeons [56]. However, intra-operative communications among the surgeons and clinicians for optimizing OR scheduling is not feasible due to distracting the surgeons, impacting surgical workflow smoothness, and consequently exposing the patient's health to risk [245]. Hence, modern operation rooms demand an automatic and real-time remaining surgery duration (RSD) estimation approach for OR scheduling and full utilization.

The intra-operative RSD estimation approaches can be split into two groups. The approaches belonging to the first group use sensor information such as right-hand signals [163]. As acquiring particular signals in every hospital is impossible, the second group's approaches estimate the RSD purely based on visual information acquired from the recorded videos. Since different surgical tasks usually occur in different stages of surgery, many approaches opt for RSD estimation based on phase recognition. Besides, different surgical phases usually need particular instruments. Hence, some approaches use instrument presence information as cues for RSD estimation.

Tables 2.1 and 2.2 summarize the properties of state-of-the-art approaches for workflow recognition in surgical videos. We can infer from the tables that recurrent neural networks are commonly used for workflow analysis in surgical videos. Besides, majority of the approaches have focused on Laparoscopy and Cholecystectomy videos, whereas scant attention has been devoted to the challenges in cataract

surgery workflow analysis.

2.2 Instrument Recognition and Segmentation

Recognition, localization, pose-estimation, tracking, and segmentation of surgical instruments are the intermediate steps in many applications of computer-assisted surgery (CAS), ranging from surgical skill assessment and surgical phase estimation to automatic guidance for workflow optimization and decision making [122]. In particular, motion analysis of instruments is a prerequisite for many surgical-skill-assessment approaches [44]. The classical instrument recognition and segmentation approaches use the histogram and rotation-invariant hough transform [64].
Tables 2.3 and 2.4 include the configurations of state-of-the-art instrument recognition and segmentation approaches. As listed in the tables, many approaches exploit region-based CNNs such as Faster R-CNN [201] and Mask R-CNN [97] for instrument localization, segmentation, and tracking. Some recent approaches opt for pixel-level recognition using U-Net-based architectures and propose various convolutional modules to deal with different instrument segmentation challenges [177, 176, 175].

2.3 Skill Assessment

High-quality surgical procedure is a determining factor in global public health and a contributing factor in healthcare cost reduction. Computer-aided surgical skill assessment can effectively contribute to the quality of surgical procedures through individualized feedback and automatic coaching. The computer-aided surgical skill assessment approaches can be split into two sub-problems: (1) how to extract the skill-related features, and (2) how to classify the skills based on the extracted features. Accordingly, skill assessment approaches can be broadly categorized into three groups: (i) hand-engineered features plus classical machine-learning-based classification, (ii) hand-engineered features plus deep-learning-based classification,

Table 2.3: Comparisons among the deep-learning-based instrument recognition and skill assessment approaches. In the "Dataset" column, "NP" refers to Nonpublic.

Reference	Target Surgery	Proposed Approach	Learning Method	Target Result	Features	Dataset	Year
[48]	Cholecystectomy	Evaluation	Supervised	Instrument Localization	YOLO	Cholec80 [232]	2017
[270]	Robotic Surgery	Network	Supervised	Instrument Detection & Tracking	CNN for instrument-tip detection	[49]	2017
[207]	Robotic Surgery	Evaluation	Supervised	Instrument Localization	Region Proposal Network	[207]	2017
[65, 139]	Endoscopy & Retinal Microscopy	Network	Supervised	Articulated Pose Estimation	CNN + Bipartite Graph Matching	RMIT [224], Endovis 2015 [23]	2018
[22]	Endoscopy	Challenge Results	Supervised	Instrument Detection	Static CNN	[23]	2018
[214]	Robotic Surgery	Comparative Evaluation	Supervised	Instrument Segmentation	U-Net-Based	[14]	2018
[117]	Robotic Surgery	Framework	Supervised	Instrument Localization & Skill assessment	Faster R-CNN	Modified [207]	2018
[117]	Cholecystectomy	Evaluation	Supervised	Instrument Localization & Skill Assessment	Faster R-CNN	[116]	2018
[11]	Cholecystectomy	Network	Supervised	Instrument Detection	Boosted NN-RNN	Cholec80 [232] & CATARACTS [12]	2018
[243]	Robotic Surgery	Framework	Supervised	Skill Assessment	CNN	JIGSAWS [10, 4, 5]	2018
[111]	Robotic Surgery	Network	Supervised	Skill Assessment	CNN	JIGSAWS [10, 4, 5]	2018
[182]	Robotic Surgery	Network	Supervised	Instrument Segmentation	ResNet18-Based	[14]	2019
[24]	Cholecystectomy	Network	Active Learning	Instrument Detection	Bayesian CNN-LSTM	Cholec80 [232]	2019
[269]	Robotic Surgery	Network	Supervised	Localization	Heatmap-Based Bounding-Box Regression	Endovis 2015 [23] & ATLAS Dione [207, 208]	2019
[12]	Cataract Surgery	Challenge Results	Supervised	Instrument Classification	Static CNNs	[12]	2019
[110]	Robotic Surgery	Network	Supervised	Instrument Segmentation & Saliency Prediction	Multi-Branch CNN	[14]	2019
[47]	Biportal Endoscopic Spine Surgery	Framework	Supervised	Instrument's Tip Detection	RetinaNet [154, 200]	NP	2019
[259]	Cataract Surgery	Network	Supervised	Instrument Classification and Localization	YOLOv2 Inspired	CaSToL [269]	2019

Table 2.4: Comparisons among the deep-learning-based instrument recognition and skill assessment approaches. In the "Dataset" column, "NP" refers to Nonpublic.

Reference	Target Surgery	Proposed Approach	Learning Method	Target Result	Features	Dataset	Year
[177]	Cataract Surgery	Network	Supervised	Instrument Segmentation	Dealing with Specular Reflection	Cata7 (NP)	2019
[118]	Robotic Surgery	Network	Supervised	Instrument Segmentation	Temporal-Prior Guided U-Net	[14]	2019
[74]	Robotic Surgery	Evaluation	Supervised	Skill Assessment	3D-CNN [29] & Temporal Segment Network [239]	JIGSAWS [10, 4, 5]	2019
[144]	Robotic Surgery	Evaluation	Supervised	Instrument Segmentation/Tracking & Skill Assessment	Mask R-CNN	[14]	2020
[152]	Endoscopy	Evaluation	Supervised	Instrument Segmentation/Tracking & Skill Assessment	NP	FCN	2020
[253]	Gastrectomy	Framework	Semi-Supervised	Localization	Addressing Class-Imbalance	NP	2020
[88]	Endoscopy	Framework	Supervised	Instance-Based Segmentation	Mask R-CNN [97] + FlowNet2 [109]	[14, 13]	2020
[183]	Endoscopy	Network	Unsupervised	Semantic Segmentation	Cycle GAN	[14]	2020
[192]	Robotic Surgery	Framework	Supervised	Instrument Trajectory & Surgical State	+LSTM	JIGSAWS [77] and RIOUS [194] (NP)	2020
[144]	Robotic Surgery	Framework	Supervised	Instrument Trajectory & Skill Assessment	Mask R-CNN [97] + deep SORT [246]	[14]	2020
[134]	Endoscopy	Framework	Supervised	Skill Assessment	Mask R-CNN [97]	[23, 14]	2020
[176]	Endoscopy, Cataract Surgery	Network	Supervised	Instrument Segmentation	Dealing with Illumination and Scale Variation	[14], Cata7 (NP)	2020
[175]	Endoscopy, Cataract Surgery	Network	Supervised	Instrument Segmentation	Dealing with Specular Reflection and Scale Variation	[14], Cata7 (NP)	2020
[215]	Cataract Surgery	Framework	Supervised	Instrument Classification & Generalization Evaluation	Static CNNs	[209, 12]	2020
[242]	Robotic Surgery	Network	Supervised	Instrument Segmentation	U-Net-Based	[14]	2021

and (iii) deep-learning-based feature extraction and classification.

The approaches belonging to the first category utilize the collected sensor data such as robot kinematics and tool motions with some hand-engineered features for skill assessment. Fard *et al.* [169] extract several features such as motion smoothness from the motion trajectories of the instruments collected from sensors. These features are used for support vector machine (SVM) classification, k-nearest-neighbors, and logistic regression to finally classify the surgical skill level as expert or novice. Zia *et al.* [278] exploit three different types of features: (1) frequency-based features using discrete Cosine transform (DCT) and discrete Fourier transform (DFT), (2) entropy-based features using approximate entropy (ApEn), and (3) texture features using sequential motion texture (SMT). The dimensionality of these features is then reduced using principal component analysis before being fed into a nearest-neighbor classifier. Tao *et al.* [227] propose to model the surgical skills as time-series using Sparse Hidden Markov Models (SHMM). In the training stage, one SHMM is trained per each skill level. In the testing stage, the corresponding class to the generated model with the highest likelihood is selected as the class of the surgeon's skill. Zia *et al.* [279] exploit the combination of accelerometer data and spatio-temporal interest points extracted from the video frames for skill assessment. In another study [258], the duration of different phases manually annotated by a surgeon is used to predict the surgeon's skill.

In the second group, Kim *et al.* [129] utilize the motion trajectories computed from the manual annotations of the tip of instruments in the *Rhexis* phase. These trajectories are fed into a two-dimensional CNN to assess the intra-operative skills in this phase. Some recent methods suggest using instance detection and segmentation approaches to extract the position of the instrument using raw video [134, 117, 259]. Jin *et al.* [117] propose to use a region-based CNN to track the instruments in laparoscopic videos. The bounding box information is then used to compute motion trajectories, heat maps, and a timeline of the instrument used for skill assessment.

Hand-engineered feature extraction as a preprocessing step is not only a burdensome process but also may lead to suboptimal results when it comes to complex problems. In contrast, convolutional neural networks, which can provide hierarchical representation from the input data, can perform feature extraction and classification simultaneously. Hence, CNNs are regarded as superior alternatives to hand-engineered-features-based methods. On the other hand, these approaches require a very big dataset which may not be usually available. As an example of the methods belonging to the third group, Fawaz *et al.* [111] utilize a one-dimensional CNN containing three hidden layers to directly transfer 76-D kinematic data to low-dimensional latent variables. Ziheng *et al.* [243] exploit a deeper architecture consisting of five hidden layers to classify the surgeons' skill level in minimally invasive surgery. Besides, Funke *et al.* [74] propose a fully automatic skill assessment approach using a 3-D CNN [30] for action recognition from video snippets and a temporal segment network [239] for optimal aggregation of the spatio-temporal features of several snippets.

Tables 2.3 and 2.4 include the configurations of state-of-the-art skill assessment approaches. We can deduce from the table that a majority of instrument recognition and skill assessment approaches are based on supervised learning. There are, however, few approaches focus on active [24], semi-supervised [253], and unsupervised [183] learning. We can also infer from the tables that skill assessment is mainly performed using instrument recognition. Thus enhancing skill assessment accuracy necessitates accurate pixel-wise instrument recognition.

CHAPTER

3 Deblurring Using a Multi-Scale Deconvolutional Neural Network

Chapter overview — A common quality impairment observed in surgery videos is blur, caused by object motion or a defocused camera. Degraded image quality hampers the progress of machine-learning-based approaches in learning and recognizing semantic information in surgical video frames like instruments, phases, and surgical actions. This problem can be mitigated by automatically deblurring video frames as a preprocessing method for any subsequent video analysis task. In this chapter, we propose and evaluate a multi-scale deconvolutional neural network to deblur cataract surgery videos. Experimental results confirm the effectiveness of the proposed approach in terms of the visual quality of frames as well as PSNR improvement.

This chapter is an adapted version of:

"Ghamsarian, N., Taschwer, M., and Schoeffmann, K. Deblurring cataract surgery videos using a multi-scale deconvolutional neural network. In 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI) (2020), pp. 872–876. "

3.1 Introduction

Machine learning algorithms generally suffer from inadequate quality of training data [92]. Comparing the performance of various object detectors trained on quality degraded images via JPEG compression, motion blur, and defocus blur reveals that quality degradation negatively affects object detection performance [225]. Besides, it is proved that increasing image or video compression ratio (in particular, JPEG

and H264) for the test images decreases the semantic segmentation performance of both networks trained with low quality and networks trained with high-quality images [190]. A study on the effect of different quality degradation types on OCT image boundary segmentation suggests that U-Nets are vulnerable to noise, contrast reduction, and gamma correction [137]. In particular, neural networks' performance degrades when trained on blurry images [235] due to the fact that high-frequency components of visual signals are diluted[1] and the network has to undertake two tasks simultaneously: deblurring and the target task. Moreover, it is known that a CNN trained on high-quality images is often not generalizable to blurry images [234]. Hence, we desire sharp images for both training and evaluation purposes.

Taking into account that acquiring surgical videos is a cost-intensive procedure that typically interferes with the main medical purpose of surgery, high-quality video acquisition is often not possible and making use of already captured videos with degraded visual quality is a necessity in practice. We therefore propose to use a separate neural network trained to deblur surgery videos as a preprocessing step, prior to feeding video frames to a machine learning algorithm for the target task. In this study, we focus on deblurring cataract surgery videos to further leverage them for video analysis tasks like instrument detection, operation phase recognition, and surgical action recognition.

In cataract surgery, the video signal captured by the microscope is recorded and stored for postoperative analysis. Since the focus of the video camera and of the ocular tube used by surgeons need to be adjusted separately, the resulting videos are often very blurry.

In this study, we aim to address quality degradation in cataract surgery videos resulting from a defocused camera. Since we need sharp ground-truth video frames to measure the differences between blurred and deblurred frames on a meaningful scale, we use Gaussian convolutional filters to simulate focus degradation. Our proposed

[1] CNNs extract high-frequency features in their first layers such as sharp edges, and combine them to more complicated patterns in subsequent layers.

multi-scale deblurring network is composed of convolutional, deconvolutional, and residual layers and termed as Defocused image Restoration Network (*DRNet*).
In the following section, we review the state-of-the-art image and video deblurring approaches. Section 3.3 describes the proposed neural network architecture and implementation details. In Section 3.4, the experimental setup is explained and experimental results are presented. The paper concludes with a short discussion in Section 3.5.

3.2 Related Work

Image and video deblurring approaches can be divided into two classes: geometry-based approaches and deep-learning-based approaches [263]. Following the revolution of deep-learning-based image processing approaches, several neural network architectures have been proposed to address various blur-related tasks such as blur type deduction (e.g. linear motions, Gaussian blur, out-of-focused, etc. [7]), blur modeling [45], blur kernel estimation [114], and specific deblurring [210]). With respect to deblurring types, deblurring schemes can be divided into two categories: specific (targeted) deblurring, and blind (universal) deblurring. Targeted methods have been designed to deblur images blurred by a specific kernel, while blind methods assume that no information about the causes of blurring is available. Since obtaining a dataset containing blurry images and their corresponding natural sharp images is costly, a lot of synthetic blur creation functions have been proposed [219, 179, 128]. Synthetic blurry images can be obtained in two ways: convolution of sharp images with a blur kernel, or averaging a set of consecutive sharp frames of a given video (motion blur simulation).
In [138], a deblurring technique based on generative adversarial networks (GANs) [198] is proposed. The authors have suggested a novel blurring kernel based on different types of shaking, making the blurred version of each sample more complex and

realistic. In [199], a global skip connection in the convolutional neural network (CNN) is proposed to make the network more compatible with the deblurring task. Not only helps this skip connection mitigate the vanishing gradient problem when training the CNN, but it also feeds gradients from the last layer directly to the first convolutional layer during backpropagation, enabling the network to learn more informative features and converge faster [261]. It has been shown empirically [85] that most of the small patches (namely 5×5 or 7×7) in a naturally sharp image recur in its *ideally* downsampled versions. This property is frequently exploited for blind super-resolution. In [166], cross-scale recurrence is employed for deblurring since the small patches in a downsampled version of a blurry image tend to be similar to that of the sharp image. Also, it has been demonstrated in [179] that a multi-scale scheme facilitates the convergence of deblurring networks. Their proposed neural network architecture for deblurring consists of three subgraphs designed to ease the deblurring procedure. Since the amount of blur decreases with downsampling, the first subgraph undertakes the easiest task of deblurring the input image downsampled by a factor of 4 and outputs a residual frame. The second graph uses this deblurred image along with the input image downsampled at 1/2 of its original scale to produce a deblurred half-size frame. Likewise, using the deblurred output frame at scale 1/2, the last subgraph outputs the deblurred frame with the same resolution as the input.

3.3 Proposed Method

Overview. Inspired by the method suggested in [179], we propose a multi-scale network in which we use the same number of filter response maps in consecutive layers, enabling the use of residual layers. However, our network has much fewer parameters compared to the network suggested in [179] (5.25 million vs. more than 24.5 million trainable parameters) and takes advantage of four residual layers. Besides the above differences, *DRNet* (the proposed network) has been exploited to address

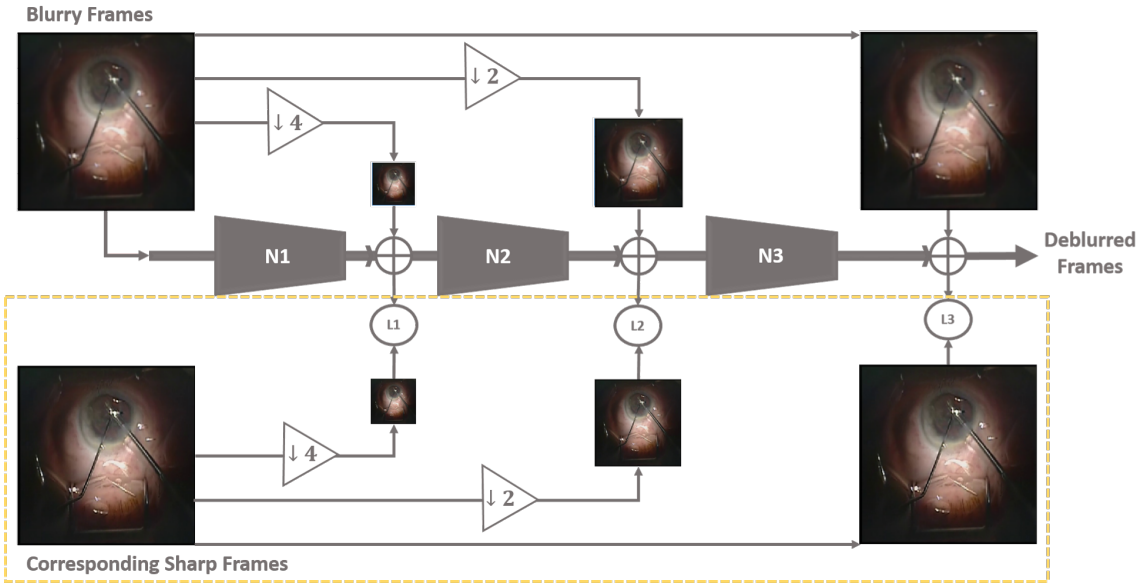


Figure 3.1: The general architecture of the deblurring network, which is inspired by [179].

defocus and Gaussian blur, while *DeblurNet* [179] is designed to deal with motion blur.

Network Architecture. Our proposed deconvolutional network is illustrated in Figure 3.1. It is comprised of three sorts of convolutional layers: down-sampling convolutional layers which abstract the non-informative spatial information and subsequently reduce the features' spatial resolution, while providing more spatial support for each pixel in the down-sampled image; flattening convolutional layers, being responsible for extracting more complicated features from input features; and deconvolutional layers, which account for performing super-resolution on input features. Each convolutional layer is followed by a batch normalization and a ReLU layer, except for the convolutional layer before the output layer of *N1* (in which we concatenate the blurry image with the residual image). Padding the images by one pixel in all dimensions ensures to avoid the reduction of dimensions after applying flattening convolutional layers (convolutions with stride 1). The detailed specification of the proposed network is given in Table 3.1.

Table 3.1: Configuration of the proposed deblurring network. The network consists of three sub-networks (SubNet). It includes some down-sampling (stride=2) and flattening (stride=1) convolutional (conv) layers as well as deconvolutional (deconv) layers. Except for the output layers, layers are followed by batch normalization (BN) and ReLU activation layers as operations (Ops). Furthermore, there are two skip connections in sub-network *N1*, and one skip connection in each of the other sub-networks.

SubNet	**N1**							**N2**					**N3**				
Layer	conv1	conv2	conv3	conv4	conv5	conv6	conv7	conv8	conv9	conv10	conv11	**deconv1**	conv12	conv13	conv14	conv15	**deconv2**
Kernel	11	7	7	7	3	3	3	5	5	5	5	5	5	5	5	5	5
OutCh	128	128	128	128	128	128	3	128	128	128	128	3	128	128	128	128	3
Stride	2	1	1	2	1	1	1	1	1	1	1	2	1	1	1	1	2
Ops	BN	BN	BN	BN	BN	BN	–	BN	BN	BN	BN	–	BN	BN	BN	BN	–
	ReLU	ReLU	ReLU	ReLU	ReLU	ReLU		ReLU	ReLU	ReLU	ReLU		ReLU	ReLU	ReLU	ReLU	
Skip	conv4	–	–	conv7	–	–	–	deconv1	–	–	–	–	deconv2	–	–	–	–

Implementation Details. We experimentally found that the random normal initialization method of network parameters leads to the fastest convergence and is the most sophisticated one for the deblurring task. Hence, random normal initialization with a mean of zero and standard variation of 0.01 is applied to all parameters. Adam activation [130] with default values of $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$ and $decay = 0.0$ is applied as optimization. The learning rate is fixed to 0.0001 for the first 20 epochs and halved after each subsequent 20 epochs. The network is trained for 60 epochs. Finally, due to limitations of GPU memory, the batch size is set to 4.

3.4 Experiments

Dataset. For evaluations, we use the Cataract-101 dataset [209], consisting of 101 cataract surgery videos with a resolution of 720×540 pixels (PAL standard). The videos are compressed using H.264/AVC with a frame rate of $25 fps$.

Training Details. Our proposed neural network is implemented using Keras deep learning framework. We have manually annotated the videos of our dataset and selected only naturally sharp frames. In order to provide the training set, a Gaussian filter with a randomly selected window size from $w \in \{1, 3, 5, 7\}$ is applied to each sharp frame. In the case of $w = 1$, no distortion will be applied. In fact, we feed the network with both blurry and sharp images in order to encourage it not to change the content of sharp images as far as possible.

Experimental Results. To evaluate the achievable deblurring performance, we compute the PSNR values between naturally sharp frames and their artificially blurred and deblurred representatives. We also evaluate the performance of the *DeblurNet* network [179] trained using the same configuration as our proposed method. The results are given in Figure 3.2 and reveal the following insights: (1) the deblurred versions of blurry frames have clearly better visual quality than their input (higher

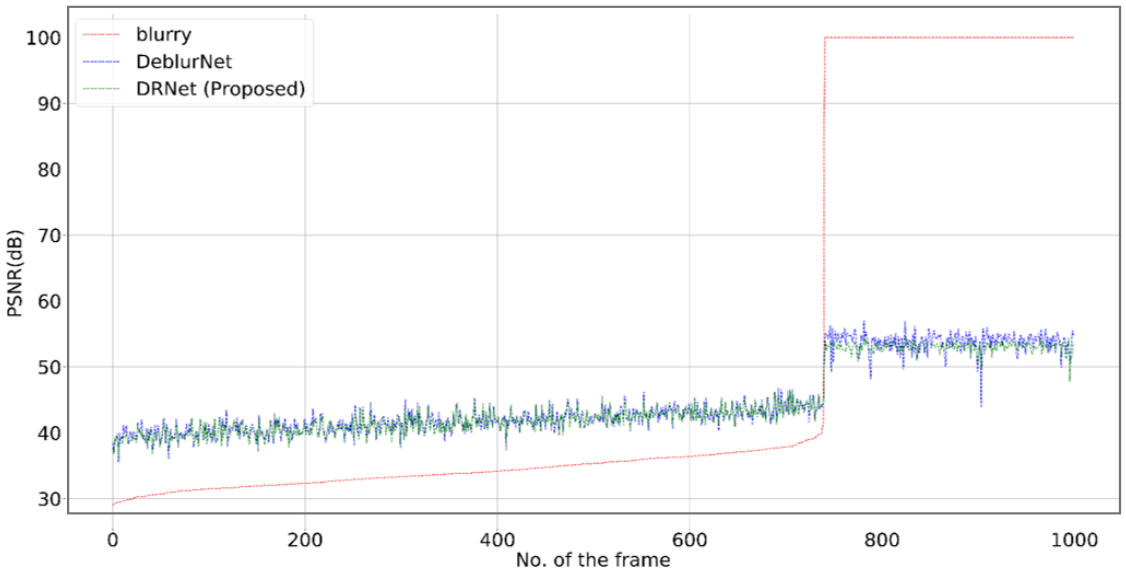


Figure 3.2: Comparative results of the proposed network (*DRNet*) and rival (*DeblurNet* [179]) for 1000 test frames (the frames are sorted by PSNR of input frames)

Table 3.2: Mean PSNR of deblurred video frames (w denotes the size of the Gaussian filter used to blur input frames).

Method	$w = 1$	$w = 3$	$w = 5$	$w = 7$
DRNet	53.06	43.05	41.70	40.11
DeblurNet [179]	53.78	43.06	41.86	40.47

PSNR values); (2) deblurred versions of naturally sharp frames (right part in the figure) have high PSNR values, meaning that they are not significantly distorted[2]; (3) our proposed model with less than a quarter of parameters provides smoother fluctuations and more stable results in comparison with the one proposed in [179]. Backed by the last observation, we argue that since our task is deblurring a dataset with very similar content in different frames, using a network with a large receptive field as in *DeblurNet* [179] can lead to overfitting.

As perceived from Table 3.2, the differences between mean PSNRs of our *DRNet* network and *DeblurNet* [179] are not significant. Figure 3.3 shows visual examples that further confirm the correct operation of the proposed network: the deblurred versions (c) of the input images (a) are perceived as almost identical to the naturally sharp images (d). In addition, norms of residual frames (b) are directly related to

[2]Images with $PSNR \geq 50$ dB are perceived as visually not distinguishable from their originals.

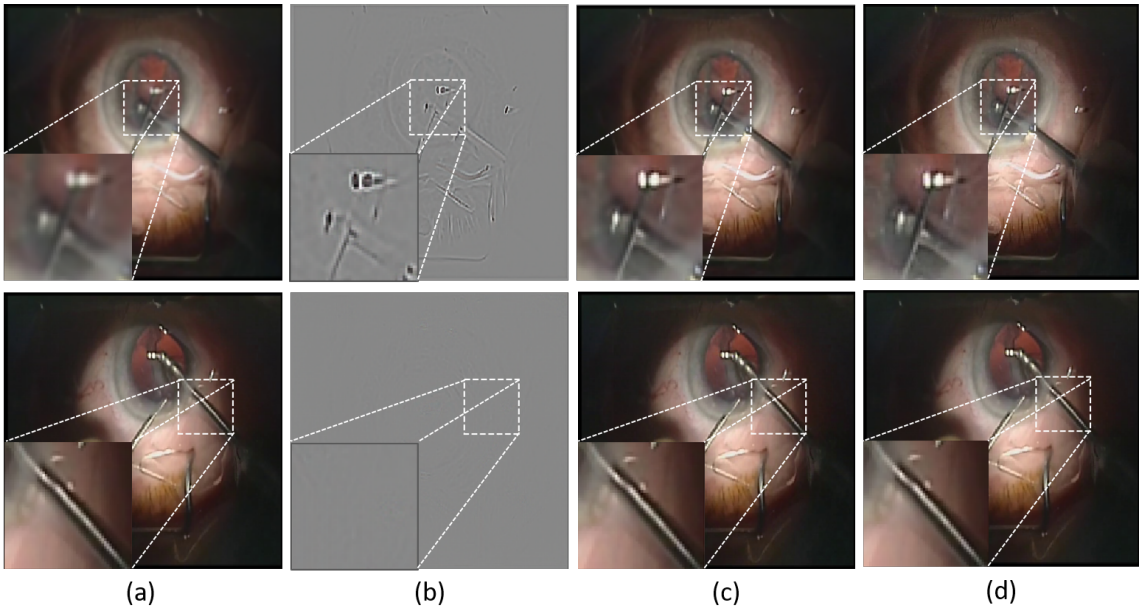


Figure 3.3: Comparative results of *DRNet*. (a) the input of the network (b) the residual frame estimated by the network (c) the output of the network (d) the naturally sharp image. The first row represents a blurry input frame, while the second row corresponds to a sharp image being fed to the network.

the amount of blur in input images.

Since we have trained the network with artificially blurred frames (random selection of Gaussian convolutions), we also want to evaluate whether the proposed network is able to deblur naturally blurred frames. For this purpose, we have manually selected blurry frames from the dataset and created visual examples that we qualitatively assess in Figure 3.4. We can conclude from the figure (which is representative for many examples we have inspected), that the network is also able to deblur such naturally blurred frames (caused by an unfocused camera), validating our assumption that defocus degradation can be simulated by Gaussian filtering.

3.5 Discussion

The distortion resulting from undesired blur in cataract surgery videos can lead to a substantial drop in the effectiveness of recognition tasks, such as instrument or phase recognition. This may be explained by the fact that noisy data will make it

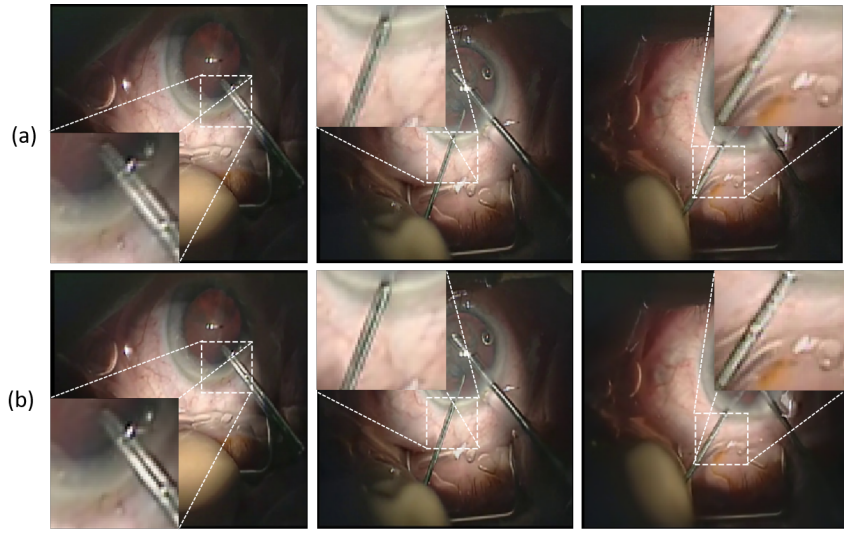


Figure 3.4: Performance of *DRNet* in case of naturally blurry frames. (a) Naturally blurry images (b) The corresponding output of the proposed network.

harder for machine learning algorithms to learn the underlying patterns. Cataract surgery video deblurring can, therefore, boost the accuracy of machine-learning-based recognition approaches [235]. Moreover, deblurring will enhance the visual quality of such videos and make them more favorable for clinical documentation or teaching. In this chapter, we have proposed a novel multi-scale residual deconvolutional network to reduce distortion in cataract surgery videos caused by a defocused camera. The proposed network is trained to estimate a residual frame which after adding to the blurry frame will result in reducing Gaussian or defocus blur being inflicted on the underlying sharp frame. Experimental results verify the capability of our proposed *DRNet* network in handling defocus blur. Even though *DRNet* originates from *DeblurNet*, it has much lower complexity in terms of parameters, leading to dependency to less annotations and less tendency to overfit when it comes to surgical videos. Moreover, it has demonstrated more reliable results thanks to residual layers. Future work could include generalizing the proposed method to other types of surgical videos or even to general deblurring problems.

Chapter

4 Relevance Detection via Spatio-Temporal Action Localization

Chapter overview — To optimize the training procedure with the video content, the surgeons require an automatic relevance detection approach. In addition to relevance-based retrieval, these results can be further used for skill assessment and irregularity detection in cataract surgery videos. In this chapter, a three-module framework is proposed to detect and classify the relevant phase segments in cataract videos. Taking advantage of an idle frame recognition network, the video is divided into idle and action segments. To boost the performance in relevance detection, the cornea where the relevant surgical actions are conducted is detected in all frames using Mask R-CNN. The spatiotemporally localized segments containing higher-resolution information about the pupil texture and actions, and complementary temporal information from the same phase are fed into the relevance detection module. This module consists of four parallel recurrent CNNs being responsible to detect four relevant phases that have been defined with medical experts. The results will then be integrated to classify the action phases as irrelevant or one of four relevant phases. Experimental results reveal that the proposed approach outperforms static CNNs and different configurations of feature-based and end-to-end recurrent networks.

This chapter is an adapted version of:
" Ghamsarian, N., Taschwer, M., Putzgruber-Adamitsch, D., Sarny, S., and Schoeffmann, K. Relevance detection in cataract surgery videos by spatio- temporal action localization. In 2020 25th International Conference on Pattern Recognition (ICPR) (2021), pp. 10720–10727."

4.1 Introduction

To systematically study the relevant surgical phases and investigate the irregularities in cataract surgeries, the sole videos are not sufficient. Dealing with tens of thousands of videos to find particular relevant phases and irregularities is burdensome and time-consuming. Hence, the surgeons require a surgical video exploration system which can shorten the surgical training curve (*i.e.*, reducing the training time by enabling fast search) and subsequently result in improved overall surgical outcomes. One of the fundamental components of such a system is an automatic phase segmentation and classification tool [78].

A cataract surgery video regularly consists of eleven action phases: incision, hydrodissection, rhexis, viscoelastic, phacoemulsification, irrigation-aspiration, capsule polishing, lens implantation, viscoelastic-suction, tonification, and antibiotics. However, not all of the aforementioned phases are equally relevant to clinicians. They consider only *rhexis*, *phacoemulsification*, *irrigation-aspiration* with *viscoelastic-suction*, and *lens implantation* as important from a medical perspective. The intraoperative complications resulting from these phases are reported to have a higher rate compared to that of the irrelevant phases [161]. Hence, detecting the aforementioned relevant phases in cataract surgery videos is of prime concern.

Figure 4.1 displays sample frames from relevant and irrelevant phases in cataract

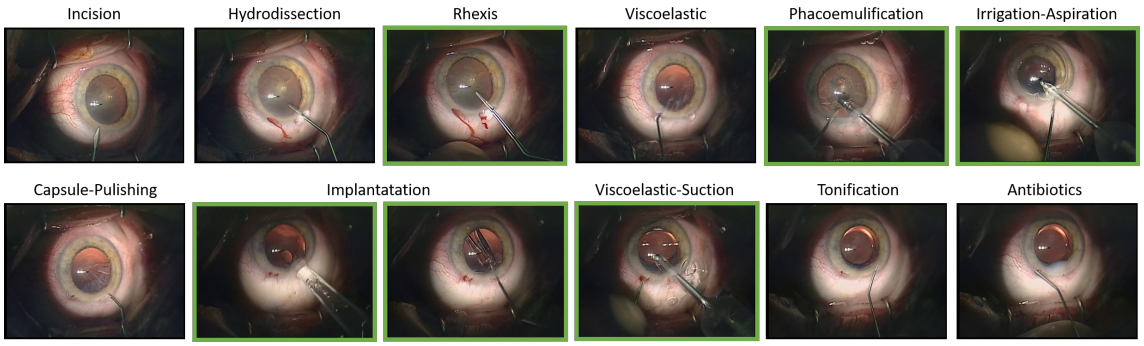


Figure 4.1: Sample frames of action phases in cataract surgery. Medically relevant phases are illustrated with green borders.

surgery videos. Designing an approach to detect and classify the relevant phases in these videos with the frame-wise temporal resolution is quite challenging due to several reasons: (i) These videos may contain defocus blur due to manual adjustment of the camera focus [83]. (ii) Unconscious eye movements and fast motion of the instrument lead to motion blur and subsequently dilution of the discriminative and salient spatial segments. (iii) As shown in Figure 4.1, the instruments that are regarded as the major difference between relevant phases are highly similar in some phases. This similarity can result in a narrow inter-class deviation in a trained classifier. (iv) The stored videos do not contain any metadata to be used as side information.

In this chapter, we propose a novel deep-learning-based approach to detect the relevant phases in cataract surgery videos. Our main contributions are listed as follows:

1. We present a broad comparison between many different neural network architectures for relevant phase detection in cataract surgery videos, including static CNNs, feature-based CNN-RNNs, and end-to-end CNN-RNNs. This comparison enables confidently scaling up the best approach to various types of surgeries and different datasets.

2. We propose a novel framework for relevance detection in cataract surgery videos using cooperative localized spatio-temporal features.

3. To enable utilizing complementary temporal information for relevance detection, idle frame recognition is proposed to temporally localize the distinct action phases in cataract surgery videos. Besides, using a state-of-the-art semantic segmentation approach, the cornea region in each frame is extracted to localize the spatial content in each action phase. In this way, we avoid inputting the substantial redundant and misleading information to the network as well as providing higher resolution for the relevant spatial content.

4. Together with this work, we publish a dataset containing the training and testing videos with their corresponding annotations. This public dataset will allow direct comparison to our results.

5. The experimental results confirm the superiority of the proposed approach over static, feature-based recurrent, and end-to-end recurrent CNNs.

In Section 4.2, we describe the shortcomings of existing approaches and give a brief explanation of *Mask R-CNN*. We then delineate the proposed relevance detection framework based on action localization termed as *LocalPhase*. The experimental settings are explained in Section 4.3 and experimental results are presented in section 4.4. The paper is finally concluded in Section 4.5.

4.2 Methodology

4.2.1 Shortcomings of Existing Approaches

Despite using the state-of-the-art baselines and showing good performance in phase recognition, the existing approaches suffer from several flaws, which are discussed as follows:

(I) Cataract surgery videos which contain irregularities in the succession of phases are of major importance for clinicians. Hence, it is expected that the trained model can recognize the phases in case of irregularities in the order and duration of them on a frame-level basis. Such a network should not be trained on the time-related and neighboring-phase-related information. Otherwise, the network memorizes the succession of phases and the relative time index of each phase in regular surgeries. Consequently, the network will fail to accurately infer the phases

in irregular surgeries [1].

(II) Previous methods in phase recognition either suppose that each video is initially segmented into different actions [260] or perform action recognition with low temporal resolution [280]. However, these approaches are incapable of providing automatic frame-wise labeling. Similarly, some methods assume that particular side information, such as tool presence signals, is available during training [35]. However, relying purely on surgical videos for workflow analysis is preferred, since (1) providing side information by the surgeons will impose burdens on their constraint schedule, and (2) RFID data or tool usage signals are not ubiquitous in regular hospitals.

(III) The existing RNN architectures such as DeepPhase [280] are not end-to-end approaches and this may result in suboptimal performance. An end-to-end approach can enable correlated spatio-temporal feature learning between recurrent and spatial convolutional layers. DeepPhase also exploits the information from the previous states to infer the corresponding phase to the current state. Due to the high computational complexity of recurrent neural networks, however, the recurrent layer can be unrolled over a limited number of frames. Since some phases may span for several seconds, in the mentioned schemes, the input frames are sampled with a low frame-rate per second (3fps in DeepPhase). Consequently, these schemes are unable to infer the phases with high temporal resolution.

(IV) To perform classification on barely separable data, more complicated features, and therefore deeper neural network architectures are required. However, a deep neural network entails more annotations and is more vulnerable to overfitting. To obtain higher accuracy with fewer image annotations, the networks are trained on low-resolution inputs. A serious defect of these approaches is the distortion of

[1] As a concrete example, a rarely occurred irregularity in cataract surgery videos is *hard-lens* condition where the surgeons have to perform *phacoemulification* phase in two stages. In the second stage, *phacoemulsification* is performed after *irrigation-aspiration*, while in a regular cataract surgery, the *irrigation-aspiration* phase is always followed by *capsule-polishing* and *viscuelastic*.

the relevant content during image downsampling. Regarding phase recognition in cataract surgery videos, these relevant contents include the instruments and cornea. The distortion inflicted by downsampling can negatively affect the classification performance.

(V) Finally, due to class imbalance in some datasets [280], the classification results cannot accurately reflect the performance of the trained networks.

4.2.2 Mask R-CNN

Mask Region-based CNN [97] (Mask R-CNN) is designed to perform instance segmentation by dividing it into two sub-problems: (1) object detection that is the process of localizing and classifying each object of interest, and (2) semantic segmentation that handles the delineation of the detected object at pixel-level. In the object detection module, a region proposal network (RPN) uses the low-resolution convolutional feature map (CFM) coming from a backbone network. RPN attaches nine different anchors centered around each feature vector in CFM. These anchors have three different aspect ratios to deal with different object types (horizontal, vertical, or squared) and three different scales to deal with scale variance (small, medium, and large-sized objects). The anchor properties along with the computed features corresponding to each anchor are used to decide the most fitted bounding box for each object of interest. In the semantic segmentation module, a fully convolutional network (FCN) is utilized to output instance segmentation results for each object of interest. Using the low-resolution detection of the object detection module, the FCN produces the masks with the same resolution as the original input of the network.

4.2.3 Proposed Approach

Figure 4.2 demonstrates the overview of the proposed framework which consists of three modules:

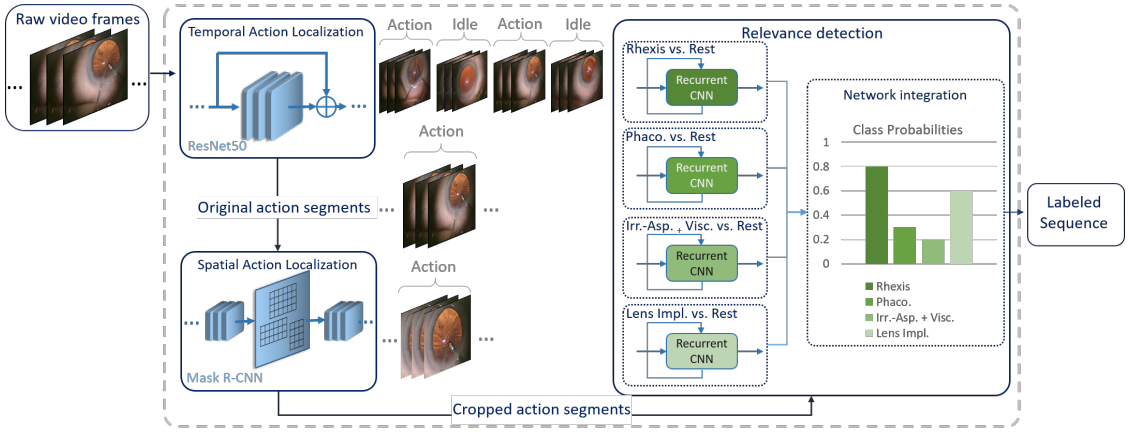


Figure 4.2: Block diagram of the proposed approach.

Temporal action localization: Each action phase in cataract surgery is always delimited by two idle phases. An *idle phase* refers to a temporal segment in a cataract surgery video where no instrument is visible inside the frame – and accordingly, no surgical action is performed. Detecting the idle phases can enhance relevance detection results by enabling the use of complementary spatio-temporal information from the same action phase. We propose to use a static residual network to categorize the frames of cataract surgery into *action* or *idle* frames. This pre-processing step plays a crucial role in alleviating phase classification in the *relevance detection* module.

Spatial action localization: The rationale behind spatial action localization is to mitigate the effect of the low-resolution input image on classification performance while retaining all discriminative and informative content for training. Since all the relevant phases in cataract surgery occur inside the cornea, higher resolution of the cornea can significantly boost the classification results. One way to provide a higher-resolution cornea for a network with a particular input size is to detect the cornea, crop the bounding box of the cornea, and use the cropped version instead of the original frame as the input of the network. In addition to providing high-resolution relevant content, this localization approach results in eliminating the redundant information that can cause network overfitting during training. We suggest using

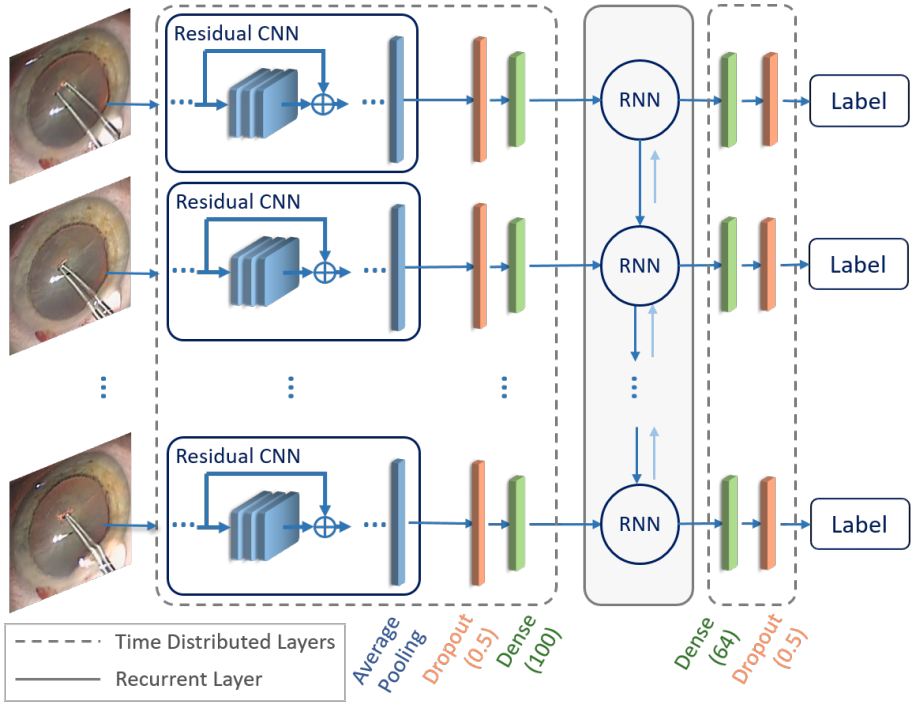


Figure 4.3: Schematic of the proposed CNN-RNNs for relevance detection.

Mask R-CNN as the state-of-the-art approach in instance-segmentation to detect the cornea and then resize the bounding box of the cornea to fit the input size of the relevance detection module.

Relevance Detection: For this module, we propose a recurrent CNN to be trained on the spatiotemporally localized action segments of cataract videos. The network is provided with sequences of the relevance-based cropped frames that contain higher-resolution information for the cornea region and complementary temporal information from the neighboring frames. In an end-to-end training manner, the network can benefit from the seamless integration of spatio-temporal features. Moreover, using back-propagation through time, the recurrent layer encourages the spatial descriptors to learn the shared representations among the input frames while preventing them from learning exceptional irrelevant features. We have experimentally found out that integrated *one-vs-rest* networks provide higher accuracy compared to multi-class classification networks. Thus we propose to train four parallel recurrent networks

(Figure 4.2), each one being responsible for detecting one particular relevant phase. Fig 4.3 demonstrates the configuration of the proposed network for relevance detection. The network contains a time distributed residual CNN (ResNet101 in our experiments) that outputs a sequence of spatial feature maps for the sequence of input frames. The network is trained one time per each relevant phase (to detect the relevant phase versus rest). The results of the four networks trained to detect four relevant phases are further integrated by using their output class probabilities in addition to the classification results. If an input is classified as the relevant phase in more than one network, the relevant phase with the highest probability is chosen as the corresponding class to that input.

4.3 Experimental Settings

4.3.1 Alternative approaches

For relevance detection, we have implemented and compared many approaches that can be categorized into (i) *static convolutional networks*, (ii) *feature-based CNN-RNN* – here, we first train the CNN independently, then replace the output layer with recurrent layers (all the layers of CNN are frozen during training the RNN layers), and (iii) *end-to-end CNN-RNN* – in contrast with feature-based CNN-RNN, the CNN is not frozen in end-to-end training manner.

To provide a fair comparison, all the networks are trained and tested on the same dataset being created based on the results of the *temporal action localization* module. In the proposed approach, however, we further pass the dataset through the *spatial action localization* module to prove the effectiveness of this module in enhancing the model accuracy.

4.3.2 Dataset

Together with clinicians from *Klinikum Klagenfurt* (Austria), we recorded videos from 22 cataract surgeries and annotated medically relevant phases. The dataset (with all videos and annotations) is publicly released with this work in the following link: https://ftp.itec.aau.at/datasets/ovid/relevant-cat-actions/.

Temporal action localization: For this step, all frames of 22 videos from the dataset are annotated and categorized as *idle* or *action* frames. We have used 18 randomly selected videos from the annotations for training and the remaining videos are used for testing. To prepare a balanced dataset for both training and testing stages, 500 idle and 500 action frames are uniformly sampled from each video, composing 9000 frames per class in the training set and 2000 frames per class in the testing set.

Spatial action localization: The area of the *cornea* in 262 frames from 11 cataract surgery videos is annotated for the cornea detection network. The network is trained using 90% of the annotations and tested on the remaining 10%.

Relevance detection: For this module, all the action segments of 10 videos are annotated and categorized as *rhexis*, *phacoemulsification* (termed as *Phaco.*), *irrigation-aspiration* with *viscoelastic-suction* (*Irr.-Asp.+Visc.*), *lens implantation* (*Lens Impl.*), and the remaining content (*Rest*). From these annotations, eight videos are randomly selected for training and two other videos are used for testing. Since recurrent CNNs require a sequence of images as input, we have created a video dataset using the annotated segments. Each annotated segment is decoded and 75 successive frames (three seconds) with a particular overlapping step are losslessly encoded as one input video. Due to different average duration of different relevant phases, we use a different overlapping step for short relevant phases to yield a balanced dataset. This overlapping step is one frame for the rhexis and implantation phase, and four

frames for other phases. Afterward for each network, 2000 clips per class from the training videos and 400 clips per class from the test videos are uniformly sampled. This amounts to 4000 clips from eight videos as the training set and 800 clips from two other videos as the testing set.

4.3.3 Neural Network Models

Temporal action localization: For idle-frame recognition, we have exploited ResNet50 and ResNet101 [98] pre-trained on ImageNet [55]. Excluding the top of these networks, the average pooling layer is followed by a *dropout* layer with its dropping probability being equal to 0.5. Next, a *dense* layer with two output neurons and *softmax* activation is added to the network to form the output layer. The classification performances of these networks are compared and the network with the best performance is used for later experiments.

Spatial action localization: For cornea detection, we utilize the *Mask R-CNN* network [97, 2]. We train the network on two different backbones (ResNet50 and ResNet101) and use the backbone with the best results to produce the cropped input for the relevance detection module.

Relevance detection: For static CNNs, we have used ResNet50 and ResNet101 with the same configuration as for the *temporal action localization* module. The network with the best results is then used as the baseline for both feature-based and end-to-end recurrent networks. Figure 4.3 shows the shared schematics of CNN-RNNs. In feature-based models, the average-pooling layers of the trained static models are used as a backbone by freezing all of the CNN layers. In contrary to feature-based models, we train the static and recurrent layers of the end-to-end models simultaneously and by starting from the weights initialized from ImageNet. We have compared four different recurrent models: (1) CNN+LSTM in which the recurrent layer includes one LSTM layer, (2) CNN+GRU which contains a GRU layer, (3) CNN+BiLSTM that

utilizes a bidirectional LSTM layer, and (4) CNN+BiGRU containing a bidirectional GRU layer. All of the recurrent layers contain five units.

4.3.4 Neural Network Settings

Temporal action localization: The SGD optimizer with $decay = 1e-6$ and $momentum = 0.9$ is set as the optimization function during training. The temporal action localization network is trained for 10 epochs with the initial learning rate lr_1 being set to 0.0005. The network is then fine-tuned for 10 epochs with $lr_2 = lr_1/5$, and 10 other epochs with $lr_3 = lr_1/10$. To avoid network overfitting, all the layers except for the last 20 layers are frozen during training. Also, *categorical cross-entropy* is used as the loss function.

Spatial action localization: The *Mask R-CNN* network pre-trained on the COCO dataset [153] is fine-tuned in an end-to-end manner starting with learning rate being equal to 0.001. The network is trained for 50 epochs; the initial learning rate is divided by 2, 10, 20 and 100 after epochs 10, 20, 30, and 40, respectively.

Relevance detection: Table 4.1 details the settings of hyper-parameters for the proposed relevance detection approach and the rival neural networks simulated to evaluate the effectiveness of the proposed approach. We have performed extensive hyperparameter optimizations to achieve a feasible speed in training while preventing overfitting during training. We came up with different numbers of frozen layers and different learning rates. For instance, the initial learning rate for static networks is set to 10^{-4}. This learning rate is divided by 2, 10, and 20 after 2, 20, and 30 epochs respectively. Besides, we use the same settings for the SGD optimizer in this module as for the *temporal action localization* module.

Due to the high computational complexity of the end-to-end training approaches, the RNN layer should be just unrolled over a short segment (clip) instead of a complete video. In this study, we assume that the RNN layers can access up to five frames

Table 4.1: Training hyperparameters and specification of the proposed and alternative *relevance detection* approaches.

Model	specification	Name	Hyper parameters				
			Frozen-Layers	Optimizer	Epochs	Learning-Rate	Batch-Size
Static	ResNet50	CNN50	[1:-20]	SGD	40	$lr = 10^{-4}, \frac{lr}{2} \mid_2, \frac{lr}{10} \mid_{20}, \frac{lr}{20} \mid_{30}$	16
	ResNet101	CNN101	[1:-10]				
	ResNet152	CNN152	[1:-10]				
Recurrent (feature-based) ResNet50 backbone	GRU	GRU-FB	CNN	SGD	15	$lr = 10^{-4}, \frac{lr}{2} \mid_5$	16
	LSTM	LSTM-FB					
	Bidirectional GRU	BiGRU-FB					
	Bidirectional LSTM	BiLSTM-FB					
Recurrent (end to end) ResNet50 backbone	GRU	GRU-E2E	[1:-10]	SGD	15	$lr = 10^{-4}$	16
	LSTM	LSTM-E2E					
	Bidirectional GRU	BiGRU-E2E					
	Bidirectional LSTM	BiLSTM-E2E					
Recurrent (end to end)	Bidirectional GRU	LocalPhase-G	[1:-10]	SGD	15	$lr = 10^{-4}$	16
	Bidirectional LSTM	LocalPhase-L					

for decision. The sequence generator divides each input video into five temporal segments and randomly chooses one frame from each temporal segment. This random choosing of the frames encourages the network to learn the relationships in diverse temporal distances. It can also act as a temporal data augmentation technique.

4.3.5 Data Augmentation Methods

Data augmentation during training plays a vital role in preventing network overfitting as well as boosting the network performance in case of unseen data. Accordingly, the input frames to all networks are augmented during training using offline and online transformations. Table 4.2 lists the detailed descriptions of augmentation methods utilized for all networks. The listed transformations are selected based on either inherent or statistical features in the dataset. For instance, motion blur and Gaussian blur augmentations are chosen due to having harsh motion blur and defocus blur in our dataset.

Table 4.2: Data augmentation methods applied to the classification and segmentation networks.

Augmentation Method	Property	Value
Brightness	Value range	[-50,50]
Gamma contrast	Gamma coefficient	[0.5,2]
Gaussian blur	Sigma	[0.0, 5.0]
Motion blur	Kernel size	9
Crop and pad	Percentage	[-0.25,0.25]
Affine	Scaling percentage	[0.5,1.5]

4.3.6 Evaluation Metrics

We report the performances of *temporal action localization* and *relevance detection* networks using the common classification metrics namely precision, recall, accuracy, and F1-score. For the *spatial action localization* network, we evaluate the performance using average precision over recall values with different thresholds for Intersection-

over-Union (IoU), as well as mean average precision (mAP) over IoU in the range of 0.5 to 0.95.

Table 4.3: Instance detection and segmentation results of *spatial action localization* module.

Backbone	Mask Segmentation		
	mAP_{80}	mAP_{85}	mAP
ResNet101	1.00	0.92	0.89
ResNet50	1.00	1.00	0.88
	Bounding-Box Segmentation		
	mAP_{80}	mAP_{85}	mAP
ResNet101	1.00	1.00	0.95
ResNet50	1.00	1.00	0.94

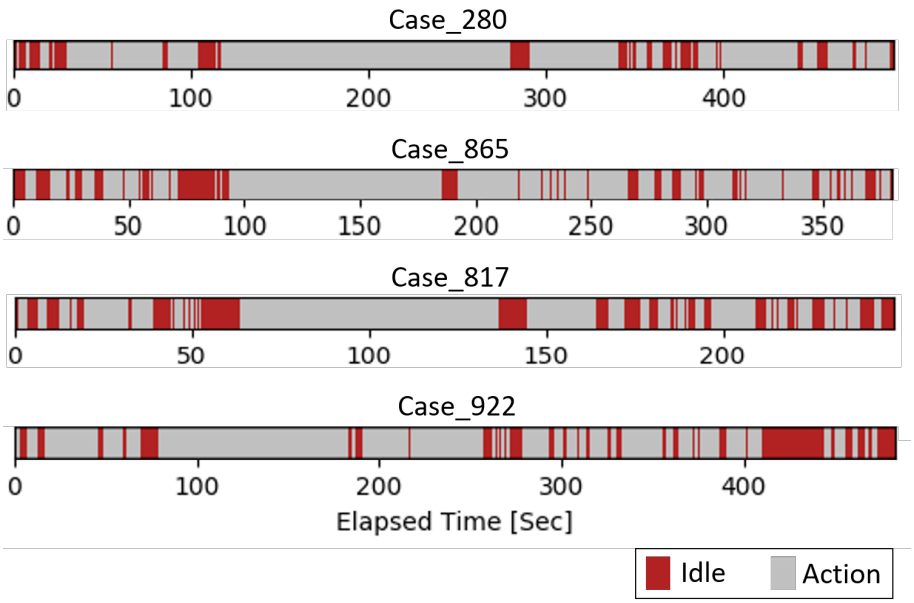


Figure 4.4: Pattern of *temporal action localization* for four representative videos.

4.4 Experimental Results and discussion

4.4.1 Temporal Action Localization

Table 4.4 reports the main classification metrics for *temporal action localization* using ResNet50 and ResNet101. As can be perceived, both models are fairly accurate and

Table 4.4: Precision, Recall, F1-Score, and accuracy of *temporal action localization* module.

Network	Class	Precision	Recall	F1-Score	Accuracy
ResNet50	Action	1.00	0.85	0.92	
	Idle	0.87	1.00	0.93	
	Macro avg	0.93	0.92	0.92	0.92
ResNet101	Action	0.99	0.88	0.93	
	Idle	0.89	0.99	0.94	
	Macro avg	0.94	0.94	0.94	0.94
ResNet152	Action	0.73	0.78	0.76	
	Idle	0.77	0.72	0.74	
	Macro avg	0.75	0.75	0.75	0.75
DenseNet121	Action	0.99	0.87	0.93	
	Idle	0.89	0.99	0.93	
	Macro avg	0.94	0.93	0.93	0.93
DenseNet169	Action	0.96	0.90	0.93	
	Idle	0.90	0.96	0.93	
	Macro avg	0.93	0.93	0.93	0.93
DenseNet201	Action	0.98	0.88	0.93	
	Idle	0.89	0.99	0.93	
	Macro avg	0.94	0.93	0.93	0.93
VGG16	Action	0.98	0.95	0.96	
	Idle	0.95	0.99	0.97	
	Macro avg	0.97	0.97	0.97	0.97
VGG19	Action	0.99	0.96	0.97	
	Idle	0.96	0.99	0.97	
	Macro avg	0.97	0.97	0.97	0.97

show a close performance, with the F1-score of ResNet101 being 1% better than ResNet50. Thus we use ResNet101 for further experiments. The reason why we have a lower rate of recall for action phases is rooted in the inherent problems in the dataset. The harsh motion blur in some action frames distorts the instruments' spatial content and makes them even invisible for the human eye. To retrieve these wrong predictions, we use a temporal mean filter with a window size of 15 (around 0.5 seconds) as a post-processing step. Figure 4.4 illustrates the filtered *temporal action localization* results for four representative cataract surgery videos.

4.4.2 Spatial Action Localization

The mask segmentation and bounding-box detection results for cornea tracking are presented in Table 4.3. It should be noted that the bounding-box segmentation results based on instance segmentation networks are much more accurate compared to that of the object detectors. This is the reason why we use *Mask R-CNN*, although we just need the bounding-box of the cornea. The figures for bounding-box segmentation affirm that both networks can detect the cornea with at least 0.85% IoU. Since the network trained with the ResNet101 backbone shows 1% higher mAP compared to that with ResNet50 backbone, this trained network will be used for further experiments.

4.4.3 Relevance Detection

The classification reports of the different static, feature-based recurrent, and end-to-end recurrent neural networks are listed in Table 4.5 and Table 4.6. Considering the static CNNs (namely ResNet50, ResNet101, and ResNet152), we can see different behaviors of a network for different phases. This difference lies in the level of similarity between each target phase and other phases. For instance, the *Irr.-Asp.+Visc.* phase shares a lot of statistics and visual similarities with the *Phaco.* phase. Since the frames corresponding to *Phaco.* contain more feature variations, all static CNNs tend to classify *Irr.-Asp.+Visc.* frames as *Phaco.* This tension decreases by increasing the number of layers in the network, as it contributes to discriminating more complicated features. On the other hand, networks with more parameters are more prone to overfit during training on small datasets. In summary, ResNet50 and ResNet101 have shown the same level of accuracy on average. Thus we choose ResNet50 having fewer parameters as the baseline for the recurrent networks.

Thanks to the *temporal action localization* module, the feature-based recurrent neural networks have shown noticeable enhancement in classification results, specifically for

Table 4.5: Precision, Recall, and F1-Score of the proposed and alternative *relevance detection* approaches.

Network	Rhexis			Phaco.			Lens Impl.			Irr.-Asp.+Visc.		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
CNN50	0.81	0.81	0.81	0.88	0.85	0.85	0.83	0.82	0.82	0.74	0.55	0.45
CNN101	0.82	0.73	0.71	0.86	0.83	0.83	0.77	0.58	0.49	0.72	0.72	0.72
CNN152	0.95	0.95	0.95	0.76	0.71	0.69	0.81	0.74	0.72	0.69	0.63	0.60
GRU-FB	0.99	0.99	0.99	0.93	0.92	0.92	**0.88**	0.85	0.84	0.80	0.73	0.71
LSTM-FB	0.95	0.95	0.95	0.90	0.90	0.90	0.85	0.80	0.79	0.77	0.77	0.77
BiGRU-FB	0.98	0.98	0.98	0.93	0.92	0.92	0.87	0.82	0.79	0.78	0.79	0.67
BiLSTM-FB	**1.00**	**1.00**	**1.00**	0.92	0.91	0.91	**0.88**	0.84	0.83	0.79	0.76	0.76
GRU-E2E	0.98	0.98	0.98	0.92	0.91	0.91	0.83	0.75	0.74	0.78	0.66	0.63
LSTM-E2E	0.95	0.94	0.94	0.83	0.83	0.83	0.84	0.79	0.78	0.69	0.66	0.64
BiGRU-E2E	0.99	0.99	0.99	0.92	0.91	0.91	0.84	0.76	0.75	0.80	0.76	0.75
BiLSTM-E2E	**1.00**	**1.00**	**1.00**	0.93	0.93	0.93	0.80	0.67	0.63	0.74	0.71	0.70
LocalPhase-G	0.99	0.98	0.98	0.95	0.94	0.94	0.86	0.85	0.85	**0.85**	0.80	0.80
LocalPhase-L	0.99	0.99	0.99	**0.96**	**0.96**	**0.96**	0.87	**0.86**	**0.86**	0.84	**0.83**	**0.83**

Table 4.6: Accuracy of the proposed and alternative *relevance detection* approaches.

Network	Rhexis	Phaco.	Lens Impl.	Irr.-Asp.+Visc.
CNN50	0.81	0.85	0.82	0.55
CNN101	0.73	0.83	0.58	0.72
CNN152	0.95	0.71	0.74	0.63
GRU-FB	0.99	0.92	0.85	0.73
LSTM-FB	0.95	0.90	0.80	0.77
BiGRU-FB	0.98	0.92	0.82	0.69
BiLSTM-FB	**1.00**	0.91	0.84	0.76
GRU-E2E	0.98	0.91	0.75	0.66
LSTM-E2E	0.94	0.83	0.79	0.66
BiGRU-E2E	0.99	0.91	0.76	0.76
BiLSTM-E2E	**1.00**	0.93	0.67	0.71
LocalPhase-G	0.98	0.94	0.85	0.81
LocalPhase-L	0.99	**0.96**	**0.86**	**0.83**

rhexis and *Phaco.* phase. Interestingly, the bidirectional LSTM network can retrieve 100% of the frames corresponding to the *rhexis* phase. In summation, it can be observed that all the different configurations of the feature-based recurrent networks outperform the static CNNs. Regarding the end-to-end training approaches, we can notice some drops in the classification results for *Irr.-Asp.+Visc.* phase and *Lens Impl.* phase. This drop can occur due to an insufficient number of training examples. The end-to-end training approaches are more vulnerable to overfitting due to their high degree of freedom.

Both configurations of the proposed approach (namely bidirectional GRU and bidirectional LSTM) have achieved superior performance compared to the alternative approaches. Also, it can be perceived from the Table 4.6 that our models have the best accuracy in detecting the *Irr.-Asp.+Visc.* that is the most challenging phase to retrieve. With completely identical configuration to BiGRU-E2E, LocalPhase-G which takes advantage of the *spatial action localization* module has achieved more reliable results (3% gain in F1-score for *Phaco.* phase and 5% percent gain in F1-score for *Irr.-Asp.+Visc.* phase). Likewise, LocalPhase-L has achieved 3% and 13% higher

F1-score for the *Phaco.* and *Irr.-Asp.+Visc.* phase, respectively. These results reveal the influence of high-resolution relevant content on network training as well as the effect of redundant-information elimination on preventing network overfitting.

4.5 Conclusion

Today, considerable attention from the computer-assisted intervention community (CAI) is focused on enhancing the surgical outcomes and diminishing the potential clinical risks through context-aware assistant or training systems. The primary requirement of such a system is a surgical phase segmentation and recognition tool. In this chapter, we have proposed a novel framework for relevance detection in cataract surgery videos to address the shortcomings of the existing phase recognition approaches. Indeed, the proposed approach is designed to (i) work independently of any metadata or side information, (ii) provide relevance detection with a high temporal resolution, (iii) be able to detect the relevant phases notwithstanding the irregularities in the order or duration of the phases, and (iv) be less prone to overfitting in case of the small non-diverse training sets. To alleviate the network convergence and avoid network overfit on small training sets, we have proposed to localize the spatio-temporal segments of each action phase. A recurrent CNN is then utilized to take advantage of this complementary spatio-temporal information by simultaneous training of static and recurrent layers. Experimental results confirm that the networks trained on the relevant spatial regions are more robust against overfitting due to substantially less misleading content. Besides, we have presented the first systematic analysis of recurrent CNN frameworks for phase recognition in cataract surgeries that further confirms the superiority of the proposed approach.

CHAPTER

5 Relevance-Based Compression

Chapter overview — In this chapter, to facilitate cataract surgery video streaming and address storage limitation, we propose a relevance-based compression technique consisting of two modules: (*i*) relevance detection, which uses neural networks for semantic segmentation and classification of the videos to detect relevant spatio-temporal information, and (*ii*) content-adaptive compression, which restricts the amount of distortion applied to the relevant content while allocating less bitrate to irrelevant content. The proposed relevance-based compression framework is implemented considering five scenarios based on the definition of relevant information from the target audience's perspective. Experimental results demonstrate the capability of the proposed approach in relevance detection. We further show that the proposed approach can achieve high compression efficiency by abstracting substantial redundant information while retaining the high quality of the relevant content.

This chapter is an adapted version of:

"Ghamsarian, N., Amirpourazarian, H., Timmerer, C., Taschwer, M., and Schoffmann, K. ¨ Relevance-based compression of cataract surgery videos using convolutional neural networks. In Proceedings of the 28th ACM International Conference on Multimedia (New York, NY, USA, 2020), MM '20, Association for Computing Machinery, p. 3577–3585."

5.1 Introduction

In the field of ophthalmology, there is an ever-increasing demand to record videos from cataract surgeries. These videos are urgently required for teaching purposes – a fact that is specifically important in ophthalmology, where operations are performed with a microscope, typically allowing for only one additional viewer (*e.g.*, a trainee surgeon). Due to documenting every single moment of the surgery, much better than any textual report would do, these videos are also used for other purposes such as forensics, surgical quality assessment, and post-operative case investigations.

Recording videos of cataract surgery, however, requires huge storage space being generally not available in hospitals [172]. Assuming an exemplary mid-sized hospital with two operating rooms for ophthalmology, 20 surgeries per room with their duration being around seven minutes, we end up with more than 4.5 hours of videos per day. These videos are often stored with at least 720p resolution at 60 frames per second (fps) and 12 MBit/s [1], amounting to a storage space of 630 MB per operation, *i.e.*, 12.3 GB per day for that hospital. As this can sum up to more than 3 TB per year, the recorded ophthalmic videos are often deleted after a few days or weeks. This is especially unfortunate as the recorded videos document every subtle detail, including irregularities and complications, which rarely happen but are particularly important for teaching. On top of that, these videos can be used for knowledge exchange among hospitals through a remote online exploration system to accelerate the training process and improve the surgical outcomes [78]. The remote exploration entails degraded visual quality resulting from the limited transmission bandwidth. The quality degradation in the relevant regions of video can distort the information being crucial to the surgeons. Hence, a domain-specific compression technique is needed to penalize the distortion in relevant regions of these videos to allow efficient utilization of this great source of information.

[1] As we experienced in the hospital of our research partner

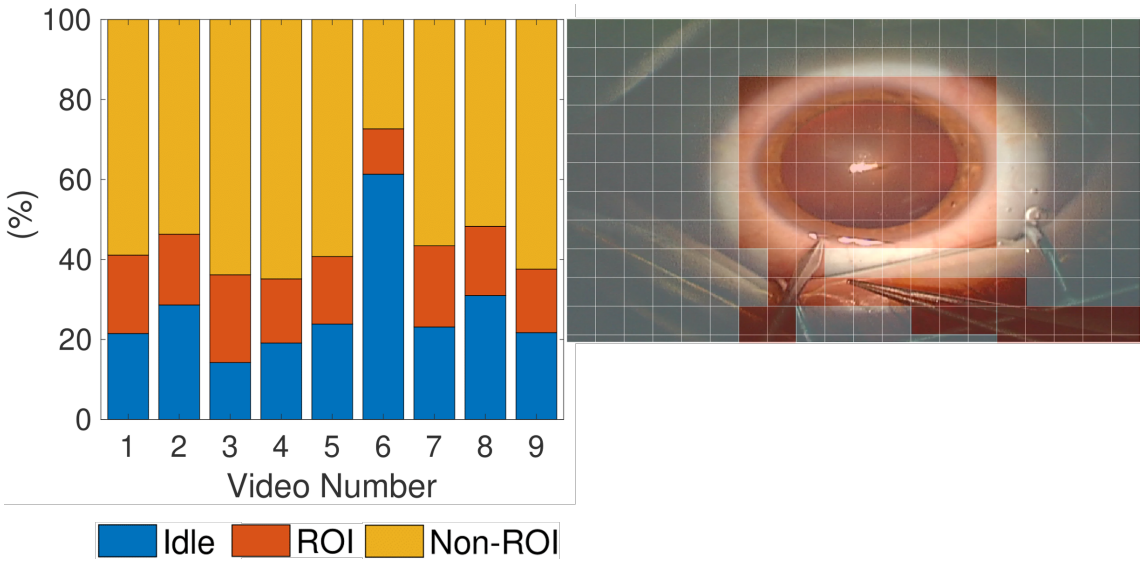


Figure 5.1: Left: the percentage of CTUs corresponding to relevant and irrelevant content in nine representative cataract surgery videos. Right: HEVC Coding Tree Units (CTUs) aligned with relevant (red) and irrelevant content.

In this work, we investigate the achievable storage space gain with content-adaptive compression of cataract surgery videos using special domain knowledge and varying quantization parameters (QP) for different coding tree units (CTU) in HEVC. We utilize the fact that *(I) idle phases* (*i.e.*, temporal regions of the video where no instrument is visible) and *(II) all spatial content except the inner part of the eye (i.e., inner circle of the cornea) and visible instruments* in *action phases* (*i.e.*, phases in which a particular surgical action is conducted using at least one instrument) are not relevant for the target audience. Accordingly, we train two different kinds of CNNs, namely *(a)* a static frame-based CNN to automatically detect *idle frames*, and *(b)* a region-based convolutional neural network (Mask R-CNN) to automatically locate the *relevant spatial regions* (cornea and surgical instruments). The results from both CNNs are further utilized to penalize the distortion of relevant content by relevance-based encoding with *High Efficiency Video Coding* (HEVC) [222]. That is, the relevant content or region of interest (ROI) is compressed yielding high quality by using a default QP, while the irrelevant content is compressed with lower quality by using a higher QP. Figure 5.1 (right) visualizes the 64×64 pixels CTU structure

of HEVC for an action frame, which contains two instruments (*primary incision knife* on the left, and *stabilizing forceps* on the right). Only these two instruments and the inner part of the eye (*i.e.*, everything inside the cornea) is relevant for users. The percentage of CTUs corresponding to idle frames, non-ROI, and ROI regions in action frames for nine cataract surgery videos are represented in Figure 5.1 (left). It can be noticed that the percentage of the relevant content in cataract surgery videos hovers around 20%. The rest of the frame can be encoded with stronger quantization, or even be removed entirely (*e.g.*, by pure black inpainting). Our main contributions are summarized as follows:

1. We propose a relevance-based compression approach using surgical domain knowledge to compress cataract surgery videos. The proposed method integrates neural networks and HEVC to achieve high compression efficiency by applying more distortion to irrelevant content, while providing high quality for the relevant content.

2. Considering the target audience of these videos – which can be expert surgeons, trainees, or computer scientists (for the sake of computerized surgical workflow analysis) – several definitions for relevant content are possible. Accordingly, we introduce a set of novel scenarios to compress cataract surgery videos to be proportionate to the users' demands.

3. The compression results show a storage space gain of up to **63%** when only slightly reducing the visual quality of the irrelevant content, and storage gain of up to **68%** by removing some parts of the irrelevant content.

4. The proposed relevance-based compression approach is generalizable to all sorts of surgical videos where the domain knowledge is required for the extraction of relevant content.

To ensure reproducibility, the selected dataset (*i.e.*, video files with annotations) has

been publicly released in the project website:
http://ftp.itec.aau.at/datasets/ovid/CatRelevanceCompression/.
The remainder of the paper is organized as follows. In Section 5.2, we position our approach in the literature by reviewing the related work on ROI-based compression and instance segmentation. We detail our method termed as RECOV (RElevance-based Compression of Ophthalmic Videos) in Section 5.3. The experimental setup is presented in Section 5.4. The relevance detection module of the proposed method is evaluated in Section 5.5. We then evaluate the achievable compression efficiency using videos from nine cataract surgeries adopting content-adaptive HEVC in Section 5.6. We finally conclude the paper in Section 7.

5.2 Related Work

5.2.1 ROI-Based Compression

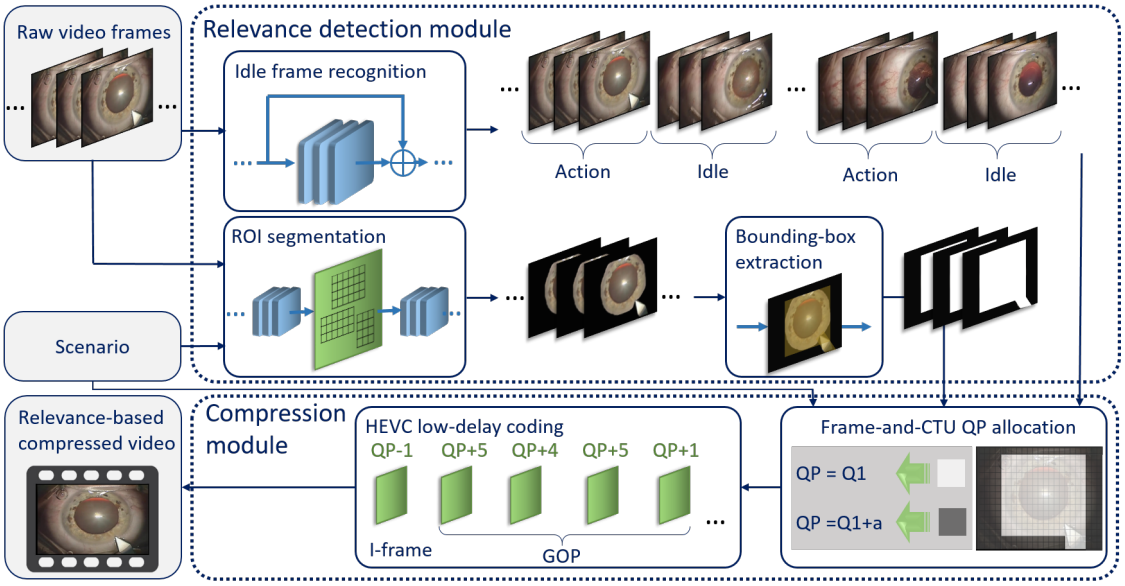


Figure 5.2: Overview of the proposed content-adaptive cataract surgery video compression framework.

ROI-based compression can be modeled as a rate-distortion problem where ROI quality is of prime concern. It can be generally split into two problems: *(i)* how

to efficiently segment the ROI (*i.e.*, ROI detection), and *(ii)* how to compress the ROI segment and background (*i.e.*, ROI encoding). ROI detection methods have experienced two stages of development. In the first generation, ROI prediction methods exploit hand-crafted features including *(a) local features* such as relation to the border regions [104] and color information [156, 155], *(b) global features* such as histogram and contrast, and *(c) dynamic features* such as motion-based saliency estimation [91]. Some methods have used a combination of low-level hand-crafted features with CNN's outputs to improve the saliency detection results [148]. For ROI encoding, traditional methods use entropy and Huffman coding [86]. Some methods adapted compression standards to be compatible with different ROI encoding strategies. The authors of [151] adapted Advanced Video Coding (AVC) to be capable of using different coding parameters assigned to different macroblocks. In [91], the rate-distortion optimization in AVC is further improved to avoid quality degradation in salient regions in the motion compensation phase. A more practical approach is to utilize image or video compression standards that support ROI encoding, such as JPEG2000 [197, 275] and HEVC [222]. In [249, 149], an ROI encoding scheme in HEVC is proposed for a hierarchical perception-based model of faces. Some recent works propose joint ROI detection-encoding approaches. In [104], the ROI is extracted during the inter-prediction phase of HEVC, using motion estimation results. Using encoder-decoder networks, the authors of [28] propose an ROI compression framework that performs image ROI detection and compression simultaneously.

In the aforementioned methods, ROI prediction is generally performed using salient-object-detection approaches. Saliency detection approaches attempt to identify the regions in images or videos that grab the human visual system's attention. Due to the characteristics of cataract surgery videos such as moving background with the same motion properties as ROI, saliency extraction methods are inefficient. In fact, the relevant segments in cataract surgery videos do not contain the saliency characteristics. Moreover, because of fairly similar visual content, hand-engineered

features are incapable of achieving high accuracy in relevance-base segmentation. The neural-network-based instance segmentation methods have superseded the traditional approaches in object segmentation. Hence, we focus on instance segmentation using neural networks in the rest of this section to be further used in the proposed relevance-based compression approach. Besides, HEVC/H265 [222] is set to be the state-of-the-art in video coding, yielding superior compression efficiency compared to its predecessors. Thus, we utilize this codec to compress the cataract surgery videos using the detected relevant content.

5.2.2 Instance Segmentation

Since the rise of the Convolutional Neural Networks (CNNs), the ability of machines in recognition tasks is improving from the course level of image classification to the precise level of pixel classification. Before the rise of deep-learning-based semantic segmentation, successful methods relied on hand-crafted features. These features were used in traditional methods such as Boosting or Random Forests to classify the central pixel of each input patch, independently and regardless of the patch's relative spatial information [213]. Conditional Random Fields were then employed to smooth these noisy predictions and improve the classification accuracy [140].

The deep-learning-based instance segmentation approaches can be grouped into two categories: *(i)* pixel-based and *(ii)* region-based approaches. Pixel-based approaches aim at performing semantic segmentation by merging the pixel-wise predictions using clustering methods, whereas region-based methods aim at jointly solving object detection and segmentation. The methods in the first category use metric learning [94, 68, 174], boundary clustering [131], conditional random fields [15], watershed transform [18], *etc.* Regarding the region-based approaches, Long *et al.* [158] proposed a method to convert classifier networks to Fully Convolutional Networks (FCN). The resulting networks apply a series of convolutional layers to the input image of arbitrary size and output a pixel-wise probability map of the same

size per each semantic category. The major flaw of FCNs is that these networks are unable to distinguish individual object instances and consequently fail to address *instance-aware semantic segmentation*. To provide instance-aware segmentation, some methods belonging to the second category apply object detectors such as *R-FCN* [53] and *Faster R-CNN* [202] to detect the bounding box of the target object and predict the mask using the detected region. Building on top of *Faster R-CNN*, *Mask R-CNN* [97] performs object detection and instance-aware semantic segmentation simultaneously. For object detection, a feature proposal network (FPN) uses the convolutional feature map (CFM) from a backbone network. FPN attaches nine different anchors centered around each feature vector in CFM with different aspect ratios to deal with different object types and different scales to deal with scale variance. The anchor properties along with the computed features corresponding to each anchor are used to decide the most fitted bounding box for each object. For semantic segmentation, a fully convolutional network uses the low resolution detected objects to finally output instance segmentation results for each object with the same resolution as the original input of the network. Due to fast training and exceptional performance, the Mask R-CNN network has become a solid baseline in instance segmentation [106, 71, 37].

5.3 Proposed Approach

Figure 5.2 shows the block diagram of the proposed cataract surgery video compression approach. Our relevance-based compression approach consists of two modules: (1) *relevance detection* that accounts for classifying the video frames into idle and action frames and classifying the pixels in action frames into ROI and non-ROI pixels, (2) a *compression module* being responsible for compressing the video content using different QPs allocated to different CTUs and different frames. We generally consider two different QPs: one for the CTUs related to the relevant content (OP_r),

and another one for CTUs related to the irrelevant content (OP_i).

Cataract surgery video normally consists of eleven action phases. Each action phase in cataract surgery is delimited by two idle phases, which do not contain any informative content and are regarded as irrelevant phases. To investigate this surgery, besides, not all the action phases are important [2]. Since the actions and reactions in important phases do not occur outside of it, the cornea (*i.e.*, iris and pupil) is considered as the region-of-interest for cataract surgery. In case that video and image processing approaches such as instrument, phase, and action recognition are to be conducted on the cataract surgery videos, distortion on instruments should also be penalized. Everything outside of these regions is not relevant and can be heavily quantized or even be removed. Accordingly, five scenarios for relevant detection based on user preference (cataract surgeons or computer scientist) are proposed in this study:

- **Scenario I:** *Action*

  In this scenario, it is supposed that idle frames are the only irrelevant parts. Therefore, the whole action frames are considered as ROI.

- **Scenario II:** $Action \cap (cornea \cup instrument)$

  A primary goal of recording cataract surgeries is to employ advanced technology to analyze the surgeries to achieve superior surgical outputs. This profile is based on the assumption that the resulting videos might be exploited in computerized image processing, where the quality of frames in eye and instrument positions play a major role in recognition tasks, such as phase and action recognition. Thus, cornea and instruments in action frames are regarded as the ROI.

- **Scenario III:** $Action \cap cornea$ (simple)

  In exploration, teaching, and surgical documentation, just the cornea in action frames is regarded as the relevant content. Therefore, the smaller QP (QP_r) is

[2] Rhexis, phacoemulsification, irrigation-aspiration, and lens implantation are the only phases being important for clinicians.

allocated to the CTUs belonging to the cornea in action frames, whereas CTUs belonging to the non-ROI regions and idle frames are quantized with the higher QP (QP_i).

- **Scenario IV:** *Action* $\cap$ *cornea* (Luma preference)

  As chroma channels convey less important information, in this scenario we aim to compress C_b and C_r channels of irrelevant content with more distortion compared to the Y channel. Therefore, the luma components of CTUs belonging to irrelevant content will be compressed by the same QP_i as the previous scenario, while the chroma components will be compressed using $QP_i + \alpha$.

- **Scenario V:** *Action* $\cap$ *cornea* (removed background)

  In this scenario, we attempt to further reduce the required storage by completely removing the content of CTUs not belonging to the cornea segments in both idle and action frames. The cornea-related content in idle frames is kept (and compressed with QP_i) to avoid attention distraction during watching the videos.

5.3.1 Relevance Detection Module

Idle frame recognition: As mentioned in Section 5.1, cataract surgery video can be divided into idle and action phases. An idle phase refers to a temporal segment in which no instrument is visible inside the frame. It is typically the moment where surgeons are changing the instruments. As shown in Figure 5.1, usually around 20% of the frames in a cataract surgery video belongs to the idle frames. As these frames are not relevant from a surgeons' point of view, such frames can be compressed with a higher quantization parameter (*i.e.*, resulting in lower quality). To achieve this goal, we suggest employing a classifier network to categorize the cataract surgery frames into idle and action frames. Due to the fast convergence of ResNet50 and ResNet101 [98], thanks to the residual layers, these networks will be trained and exploited for idle frame recognition. A mean filter of size 15 is then applied to the

labels of consecutive frames to reduce the wrong predictions.

ROI segmentation: Depending on the selected scenario, the region-of-interest in action frames is extracted using trained Mask R-CNN networks. For the first scenario, the whole action frame is regarded as the ROI. Therefore, ROI segmentation will not be applied in this scenario. In the other scenarios except for II, just the cornea is considered as the ROI.

We experimentally discovered that the accuracy of mask segmentation drops when trained to jointly segment the instruments and eye. This is due to the fact that in some phases (such as *phacoemulsification*), corresponding regions to the corrupted lens have visually similar properties to the instruments. Accordingly, we suggest training two separate networks to segment the cornea and instrument and combine the results for scenario II.

ROI segmentation: Bounding-box extraction: Since the surgical videos usually undergo particular machine-learning-based image and video processing approaches for effective exploration and investigation, it is crucial to preserve the high quality of relevant content[3]. To make sure that the relevant content is far less likely to be distorted, we suggest using bounding boxes instead of the masks. For typical bounding box extraction, the top-left and bottom-right coordinates of the mask are extracted and used as the input for the compression module. Since the instruments are long and angled in relation to the bounding boxes, typical bounding box extraction may lead to bitrate wastage. A more optimized method to accommodate both demands (*i.e.*, less distortion for ROI and less bitrate wastage) is to extract an oriented bounding box using the segmentation results [71]. The instrument is first rotated to be positioned vertically, the bounding box is extracted using the rotated

[3]The distortion resulting from low-quality compression of the *false negative* detections (*i.e.*, the relevant regions which are wrongly classified as irrelevant) can extremely degrade the performance of machine learning approaches in recognition tasks (*e.g.*, phase and action recognition). Notably, the diluted high-frequency features in relevant regions hamper the learning process in neural networks.

instrument, and the extracted bounding box is reversely rotated to fit the actual position of the instrument.

5.3.2 Compression Module

Frame-and-CTU QP allocation: In this stage, two QPs will be defined for the input video: the smaller one for the relevant content based on the selected scenario, and the larger one for the irrelevant content. For the first scenario, only a frame-based QP allocation will be conducted. The smaller QP (Q_r) is allocated to the action frames, while the greater QP ($QP_i = QP_r + \Delta Q$) will be assigned to the idle frames. For the other profiles, the smaller QP (QP_r) will be just allocated to the CTUs of the action frames in which there exists at least one pixel from the ROI. In scenario IV, the irrelevant content will be further distorted by assigning $QP_i + \alpha$ to the chroma channels of irrelevant content. These QP masks along with the raw video frames will be fed into the HEVC encoder for compression.

HEVC encoding: In this stage, using the raw video frames and initially allocated QPs, the whole cataract surgery video will be encoded. Since the idle frames are always encoded with a higher QP compared to the action frames, the idle frames should not be exploited as the reference frame for the action frames. Otherwise, the distortion inflicted by the higher QP in idle frames can propagate in subsequent action frames and adversely affect their output quality. Thus, in addition to the first frame of the video, the first frame of each action phase is considered as I-frame. The remaining frames will be set to P-frames according to the HEVC Low-Delay-P (IPPP) configuration [26]. The reason why we encode the remaining frames as P-frame is the inherent temporal properties of the cataract surgery videos. The succession of idle and action phases with different durations makes it difficult to use an encoding profile including bidirectional frames without negatively affecting the visual quality in action frames.

According to the Low-Delay-P configuration [26], the initially assigned QPs are further adjusted to provide better compression efficiency as well as retaining the perceived visual quality. Except for the I-frames, the consecutive frames are split into GOPs (Group Of Pictures) consisting of four frames. As shown in Figure 5.2, the assigned QP to each CTU in I-frames is subtracted by one, and in successive frames of each GOP is added by 5, 4, 5, and 1, respectively. The video is then encoded using the offset QPs.

5.4 Experimental Setup

5.4.1 Dataset

Without loss of generality, the released Cataract-101 dataset [209] being collected in 2017-2018 at Klinikum Klagenfurt (Austria) is used to evaluate the proposed method.

Classification Networks: For idle frame recognition, all frames of 22 videos from the dataset are annotated and categorized as idle or action frame. From these annotations, 18 videos are randomly selected for training and remaining videos are used for testing. Subsequently, 500 idle and 500 action frames are uniformly sampled from each video, composing 9000 frames per class in the training set and 2000 frames per class in the testing set.

Segmentation Network: We have annotated the cornea of 262 frames from 11 cataract surgery videos for the eye segmentation network, and the instruments of 216 frames from the same videos for the instrument segmentation network. We trained each network using 90% of annotations and tested them using the remaining 10%.

Relevance-Based Compression: The complete sequences of nine videos (excluding the annotated videos for classification) are selected as representative videos (Fig-

ure 5.1) to test the compression efficiency of the proposed relevance-based compression approach.

5.4.2 Neural Network Models and Settings:

Classification Networks: For idle-frame recognition networks, we have utilized ResNet50 and ResNet101 [98] pre-trained on ImageNet [55]. We use the average pooling layer of these networks. This layer is followed by a *dropout* layer with its dropping probability being equal to 0.5, before feeding to a *Dense* layer with two output neurons and *Softmax* activation. As we experimentally found out that the networks perform better with Stochastic Gradient Decent (SGD) rather than Adam optimizer, SGD optimizer with $decay = 1e-6$ and $momentum = 0.9$ is used for training. We train the classification networks for 30 epochs with the initial learning rate being set to 0.0005. The learning rate is divided by 5 after 10 epochs, and by 10 after 20 epochs. To avoid overfitting, the layers before the last 20 layers are frozen during training. Also, *categorical cross-entropy* is used as the loss function. The classification performance of these two networks will be compared and the best network will be used for later experiments.

Segmentation Network: The Mask R-CNN network [97, 2] as the state-of-the-art method in image segmentation is exploited for ROI segmentation. The network is trained on two different backbones (ResNet50 and ResNet101) and the backbone with the best results is employed in the relevant detection module. The pre-trained network on the COCO dataset [153] is fine-tuned in an end-to-end manner starting from $learning - rate = 0.001$. The network is trained for 50 epochs; the initial learning rate is divided by 2, 10, 20 and 100 after epochs 10, 20, 30, and 40 respectively.

Table 5.1: Data augmentation methods applied to the classification and segmentation networks.

Network	Augmentation Method	Property	Value
Classification	brightness	percentage	[0.5,1.5]
	rotation	degree	[-20,20]
	width shift	fraction of width	±0.1
	height shift	fraction of height	±0.1
	zoom	scaling percentage	[0.8,1.2]
	shear	intensity	0.15
Segmentation	brightness	value range	[-50,50]
	Gamma contrast	Gamma coefficient	[0.5,2]
	Gaussian blur	sigma	[0.0, 5.0]
	motion blur	kernel size	9
	crop and pad	percentage	[-0.25,0.25]
	affine	scaling percentage	[0.5,1.5]

5.4.3 Data Augmentation Methods

To optimize the training procedure and subsequently boost the performance of the network as well as avoiding overfitting, the input images of both networks are augmented applying offline and online transformations. The detailed descriptions of augmentation methods exploited for both types of networks are listed in Table 5.1. These augmentation methods are selected based on the inherent features in the dataset (and verified by several additional experiments). As a concrete example, the cataract surgery videos usually suffer from defocus blur due to manually adjusted focus [173]. Moreover, unconscious eye movements and fast instrument motions result in severe motion blur in these videos. Some other data augmentation methods such as mirroring are not employed because of the statistical features of the videos in our specific medical domain (*e.g.*, the surgeons always insert the instruments from the same side).

Table 5.2: Classification report of *Idle frame recognition*

Network	Class	Precision	Recall	f1-score
ResNet50	Action	1.00	0.85	0.92
	Idle	0.87	1.00	0.93
	Macro avg	0.93	0.92	0.92
ResNet101	Action	0.99	0.88	0.93
	Idle	0.89	0.99	0.94
	Macro avg	0.94	0.94	0.94

5.4.4 Evaluation Metrics

The performances of *classification networks* are compared using the common classification metrics namely precision, recall, and f1-score. For the *segmentation networks*, we report the performance using average precision over recall values with different Intersection-over-Union (IoU) thresholds, as well as mean average precision (mAP) with specific levels of IoU in the range of 0.5 to 0.95.

To evaluate the compression efficiency, we compute the percentage of storage gain compared to regular compression (considering the whole content of the video as the relevant content). Moreover, we demonstrate the capability of the proposed method in yielding high-quality relevant content using Peak Signal to Noise Ratio (PSNR).

5.5 Relevance Detection Results

Table 5.3: Instance detection and segmentation results of Mask R-CNN.

Target	Backbone	Mask Segmentation			Bounding-Box Segmentation		
		mAP_{80}	mAP_{85}	mAP	mAP_{80}	mAP_{85}	mAP
Cornea	ResNet101	1.00	0.92	0.89	1.00	1.00	0.95
	ResNet 50	1.00	1.00	0.88	1.00	1.00	0.94
		mAP_{60}	mAP_{65}	mAP	mAP_{80}	mAP_{85}	mAP
Instrument	ResNet 101	0.77	0.65	0.41	1.00	1.00	0.89
	ResNet 50	0.58	0.49	0.29	0.64	0.26	0.65

Table 5.2 reports the main classification metrics for idle frame recognition using ResNet50 and ResNet101. Overall, it is evident that both networks are capable of

discriminating idle frames from action frames with high accuracy. Additionally, both networks have shown quite similar results, with ResNet101 being a little bit more accurate. Looking first at the *precision* results, ResNet50 is perfectly accurate in predicting the action frames, while having 13% error in detecting the idle frames. ResNet101 has shown more balanced results in terms of precision per action and idle class. Comparing the *recall* values, it can be perceived that ResNet101 is more successful in retrieving the action frames. The *f1-score* is reported to establish a balance between *precision* and *recall* scores. Based on the *f1-score*, ResNet101 shows 2% higher performance on average. It is worth mentioning that the wrong prediction in the idle class usually occurs at the beginning and last frames of each action phase (*i.e.*, during insertion and removal of an instrument). This is because in these frames just a small part of instruments is visible and strong motion blur in some frames contributes to the network confusion between the lid holders and instruments. Since in this work the predicted idle frames are regarded as irrelevant and compressed using a larger QP, it is important to have as a few false positives as possible in the idle class. Accordingly, the trained ResNet101 is used in the idle-frame-recognition stage of the relevance detection module.

Figure 5.3 illustrates the predicted patterns of idle frames in four cataract surgery videos out of the nine representative videos (described in Figure 5.1) with different durations. These patterns are obtained after applying a mean filter of size 15 to the predictions of ResNet101. It can be seen that around 20% of frames in cataract surgery videos are idle and consequently irrelevant from a medical perspective.

The cornea and instrument segmentation results of Mask R-CNN using two different backbones are listed in Table 5.3. Considering cornea segmentation, we can observe fairly similar results from the networks with different backbones. Although, ResNet50 has shown better performance in mean average precision for IoU$\geq$80 compared to ResNet101 (1.00 versus 0.92). For the bounding boxes extracted using the mask

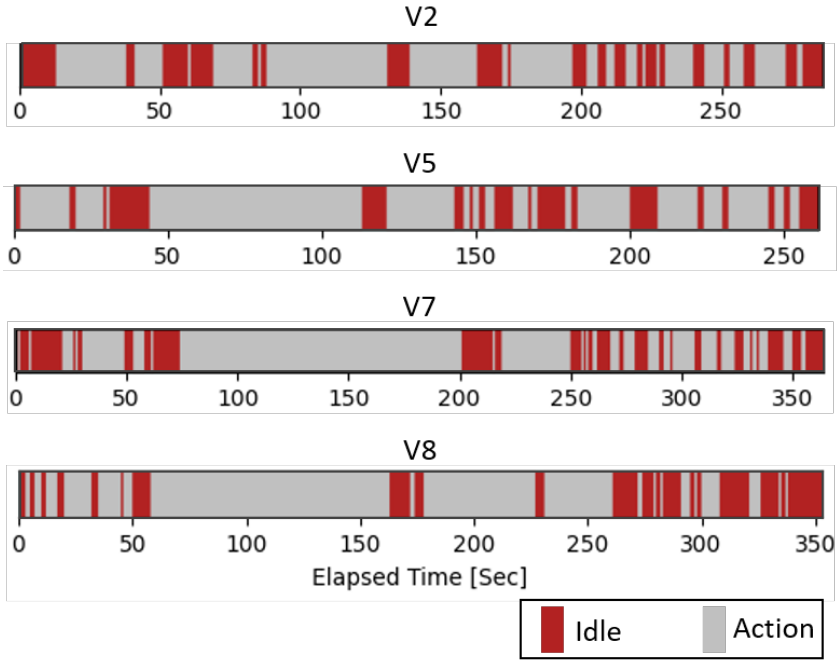


Figure 5.3: Pattern of idle frames for four videos out of nine representative videos.

segmentation results [4], the figures affirm that both networks are capable of detecting the cornea with IoU being at least equal to 0.85. Besides, the mean average precision over different IoU thresholds equals to 0.95 and 0.94 for ResNet101 and ResNet50 backbones, respectively. This affirms the reliability of our trained Mask R-CNN network in segmenting the cornea, which is the most relevant region in cataract surgery videos.

Regarding the instrument segmentation results, we can see lower accuracy in both mask and bounding box segmentation results compared to eye segmentation networks. This is based on the fact that the cornea is regularly located in the center of the frames, specifically in the action frames, where illumination is focused, while a part of the instruments is always located in the corners being usually dark. In fact, it is even hard for human eyes to distinguish the instruments' edges in dark and low-contrast regions. Fortunately, wrong ground-truth segmentations in these areas are not important and do not affect any machine-learning-based image processing

[4]It should be noted that these bounding boxes are different from the output bounding boxes of object detection networks and much more accurate.

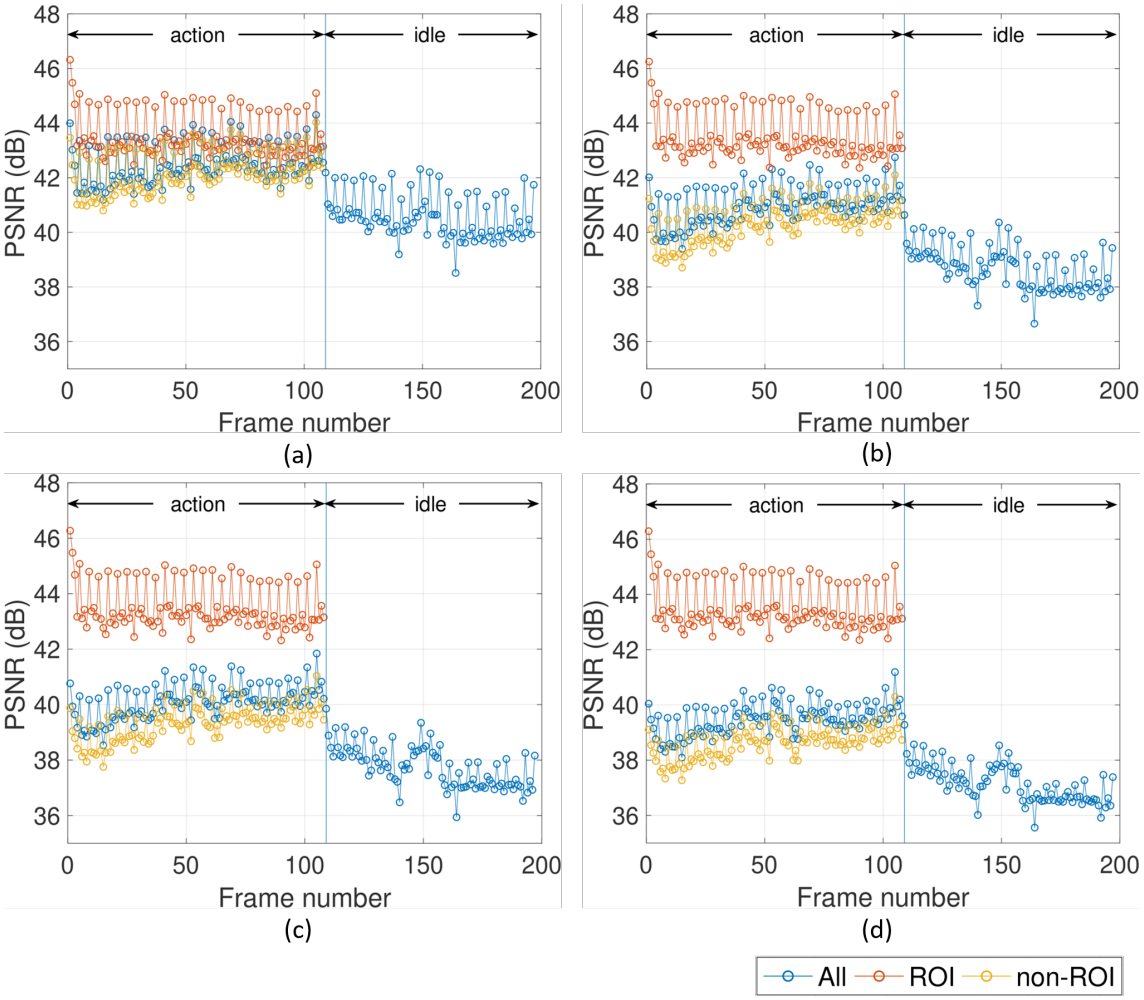


Figure 5.4: PSNR values for an exemplary segment of a cataract surgery video compressed using Scenario II with different QP differences. (a) $\Delta Q = 5$, (b) $\Delta Q = 10$, (c) $\Delta Q = 13$, (d) $\Delta Q = 15$.

approaches.

Since Mask R-CNN trained with ResNet101 as the backbone network provides much better results for both bounding-box and mask segmentation of instruments, this trained network will be used in the ROI segmentation stage. Also, due to slightly better performance on average, the trained network for cornea detection using ResNet101 as a backbone will be utilized for later experiments.

Finally, it should be noted that we are the first to address the problem of content-adaptive compression for videos in ophthalmology. This does not allow us to compare our results to any other work. However, since we release our dataset, any further work on this subject can be directly compared to our results.

5.6 Compression Results

Figure 5.4 demonstrates the output PSNR for a segment of a cataract surgery video including four seconds of an action phase followed by four seconds of an idle phase. This segment is compressed using scenario III in which the whole idle frames, as well as outside of cornea in action frames, are regarded as irrelevant content. The CTUs belonging to the relevant content (cornea in action frames) are compressed using a small QP (QP_r), whereas the other CTUs are compressed using a larger QP ($QP_i = QP_r + \Delta Q$). In this study, QP_r is fixed to 22, and QP_i is increased from $QP_i = QP_r + 5$ in Figure 5.4 (a) to $QP_i = QP_r + 15$ in Figure 5.4 (d). As can be seen in the figures, the PSNR of ROI is equivalent in all four situations – hence, the relevant content always keeps high quality. Depending on the QP assigned to irrelevant content, the distortion in idle frames and non-ROI regions of action frames is gradually increased. The fluctuations in PSNR of consecutive frames are due to the HEVC low-delay encoding configuration that adapts the quantization parameter based on perceptive visual quality. These figures confirm the effectiveness of the proposed approach in preserving the high quality of the relevant content during the compression procedure.

The achievable bitrate reduction of our proposed method for nine representative cataract surgery videos is shown in Figure 5.5. For all scenarios excluding Scenario IV, we have considered four different QPs for the irrelevant content: (1) $QP_i = QP_r + 5$, (2) $QP_i = QP_r + 10$, (3) $QP_i = QP_r + 13$, (4) $QP_i = QP_r + 15$.

To avoid attention-grabbing artifacts (*i.e.*, the blocky content in non-ROI regions resulting from a large QP that may distract the viewer's attention from the relevant content), we do not go further and choose $QP_r + 15$ as the maximum QP for non-ROI. In Scenario IV, QP for the luma channel of the irrelevant content is fixed ($QP_i^Y = QP_r + 13$), and a different quantization parameter is applied to the chroma channels of irrelevant content: (1) $QP_i^{C_bC_r} = QP_i^Y + 3$, (2) $QP_i^{C_bC_r} = QP_i^Y + 5$, (3)

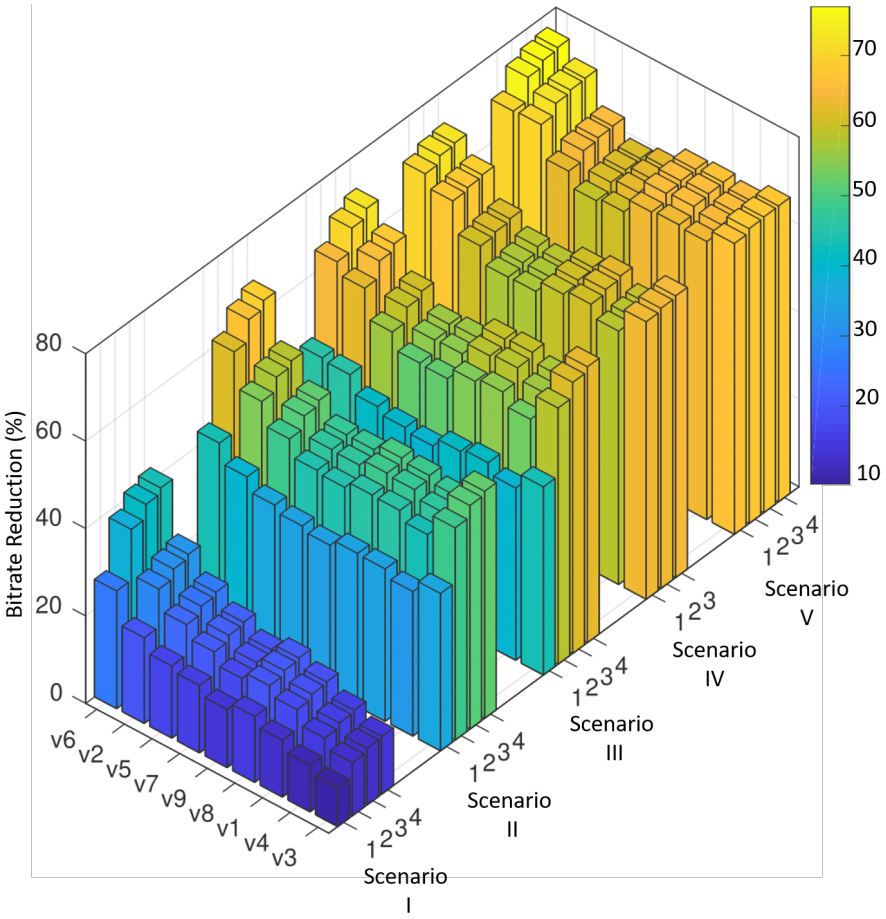


Figure 5.5: The percentage of bitrate reduction resulting from different scenarios and different QP differences for nine representative videos.

$QP_i^{C_bC_r} = QP_i^Y + 7.$

It can be perceived from Figure 5.5 that the storage space gain in each scenario is fairly close for different input videos. These results show that regardless of the input video, we can gain up to 23% storage space for the first scenario to 68% storage space for Scenario V, compared to the output bitrate of regular compression. Furthermore, Figure 5.6 shows the output size of a representative video as well as the PSNR of ROI after compression using different scenarios (the darker bars correspond to larger QPs for irrelevant content). It is evident that the proposed approach is capable of preserving the high quality of relevant regions while reducing the overall output bitrate.

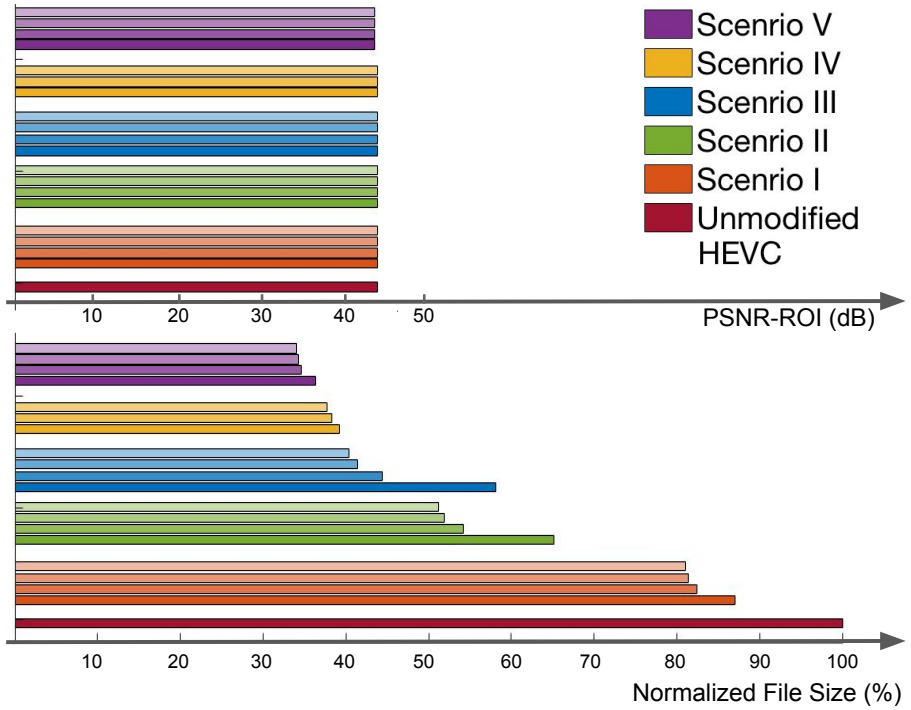


Figure 5.6: The PSNR of ROI and output size corresponding to an exemplary video compressed in different scenarios.

Expert Review. We have conducted a qualitative study to verify the suitability of the relevance-based compressed videos with six cataract experts for the five defined scenarios. The surgeons are asked to rate the visual quality of the relevant content in compressed videos in comparison with the results of unmodified HEVC on a Likert-scale from 5 (strongly agree – very suitable) to 1 (strongly disagree – not suitable). For scenarios I, II, and IV, the average ratings (ave) were 5.0 with its standard deviation (std) being equal to 0.0. In scenario III, ave = 3.66 and std = 0.81; and in scenario V, ave = 4.66 and std = 0.51. In can be inferred that for the first four scenarios, the clinicians consider the visual quality resulting from the proposed approach as almost equal to the original video, while they seem to consider the complete removal of irrelevant content in action frames as a little distracting.

5.7 Conclusion

To date, numerous investigations have sought efficient ROI-based compression. While these research efforts provide precious insights, scant attention has been devoted to ROI-based video compression. In particular, ROI-based video compression in case of dynamic backgrounds using domain knowledge has been undervalued. In this chapter, a relevance-based compression approach for cataract surgery videos (where relevant content is regarded as ROI) is proposed. The goal is to achieve high compression efficiency for irrelevant content as well as preserving the high quality of relevant content considering the target audience. To achieve this goal, the amount of distortion in different regions of videos should be proportionate to their level of relevance. Due to the inherent features of cataract surgery videos (dynamic background and close statistical properties of relevant and irrelevant content), common ROI detection approaches (*e.g.*, saliency detection approaches) are incompetent. Hence, we have proposed to use classification and semantic segmentation convolutional neural networks to detect the relevant content. These results are exploited to penalize the distortion in relevant content using CTU-wise QP allocation in HEVC. Experimental results confirm the capability of the proposed method in detecting and subsequently retaining the high quality of the relevant content, while gaining high compression efficiency by imposing more distortion on irrelevant content. The compression results show up to 63% storage-saving by applying stronger quantization on spatio-temporal irrelevant content, and up to 68% gain in storage space by removing the spatially irrelevant content. In contrast to the existing ROI-based video compression frameworks, the proposed framework can be generalized to all sorts of videos, specifically surgical videos, notwithstanding the presence of dynamic backgrounds. By automatically segmenting out the irrelevant part of videos, high fidelity for ROI as well as high video compression efficiency is possible.

CHAPTER

6 Lens Irregularity Detection

Chapter overview — In this chapter, we propose a novel framework as the major step towards lens irregularity detection. In particular, we propose (I) an end-to-end recurrent neural network to recognize the lens-implantation phase and (II) a novel semantic segmentation network to segment the lens and pupil after the implantation phase. The phase recognition results reveal the effectiveness of the proposed surgical phase recognition approach. Moreover, the segmentation results confirm the proposed segmentation network's effectiveness compared to state-of-the-art rival approaches.

This chapter is an adapted version of:

"Ghamsarian, N., Taschwer, M., Putzgruber-Adamitsch, D., Sarny, S., El-Shabrawi, Y., and Schoeffmann, K. LensID: A CNN-RNN-based framework towards lens irregularity detection. In 24th International Conference on Medical Image Computing & Computer Assisted Interventions (MICCAI 2021) (2021), p. to appear."

6.1 Introduction

Over several years, there have been numerous advances in surgical techniques, tools, and instruments in ophthalmic surgeries. Such advances resulted in decreasing the risk of severe intraoperative and postoperative complications. Still, there are many ongoing research efforts to prevent the current implications during and after surgery. A critical issue in cataract surgery that has not yet been addressed is

intraocular lens (IOL) dislocation. This complication leads to various human sight issues such as vision blur, double vision, or vision inference as observing the lens implant edges. Intraocular inflammation, corneal edema, and retinal detachment are some other consequences of lens relocation. Since patient monitoring after the surgery or discharge is not always possible, the surgeons seek ways to diagnose evidence of potential irregularities that can be investigated during the surgery.

Recent studies show that particular intraocular lens characteristics can contribute to lens dislocation after the surgery [165]. Moreover, the expert surgeons argue that there can be a direct relationship between the overall time of lens unfolding and the risk of lens relocation after the surgery. Some surgeons also hypothesize that severe lens instability during the surgery is a symptom of lens relocation. To discover the potential correlations between lens relocation and its possibly contributing factors, surgeons require a tool for systematic feature extraction. Indeed, an automatic approach is required for (i) detecting the lens implantation phase to determine the starting time for lens statistics' computation and (ii) segmenting the lens and pupil to compute the lens statistics over time. The irregularity-related statistics can afterward be extracted by tracking the lens's relative size (normalized by the pupil's size) and relative movements (by calibrating the pupil). Due to the dearth of computing power in the operation rooms, automatic phase detection and lens/pupil segmentation on the fly is not currently achievable. Alternatively, this analysis can be performed in a post hoc manner using recorded cataract surgery videos. The contributions of this work are:

1. We propose a novel CNN-RNN-based framework for evaluating lens unfolding delay and lens instability in cataract surgery videos.

2. We propose and evaluate a recurrent convolutional neural network architecture to detect the "implantation phase" in cataract surgery videos.

3. We further propose a novel semantic segmentation network architecture termed

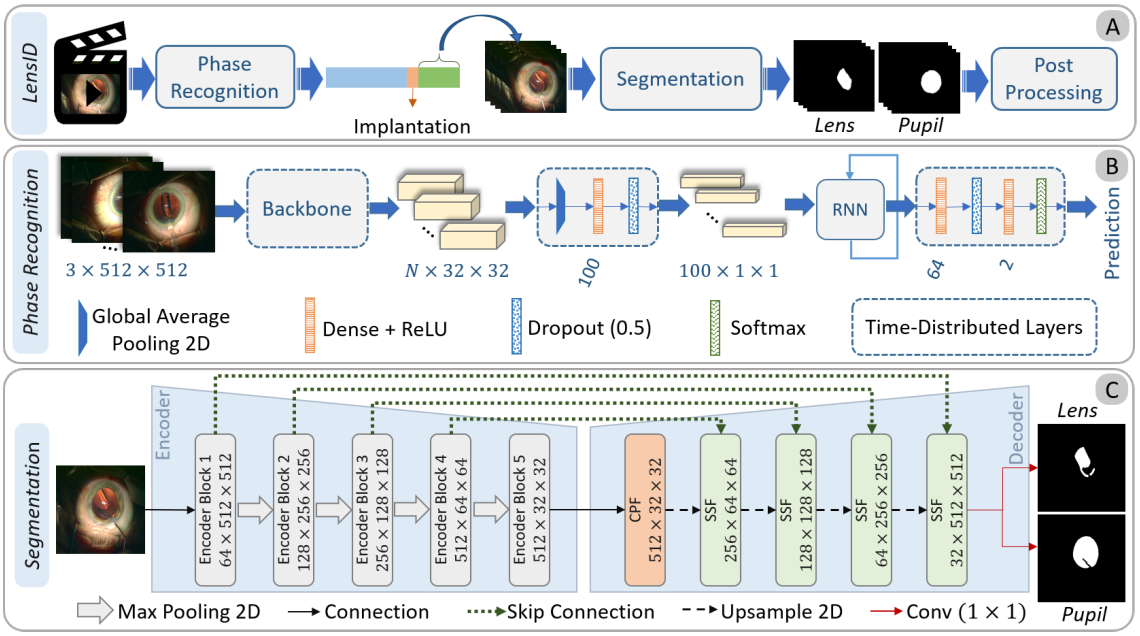


Figure 6.1: The block diagram of *LensID* and the architecture of *Phase Recognition* and *Semantic Segmentation* networks.

as *AdaptNet*[1], that can considerably improve the segmentation performance for the intraocular lens (and pupil) compared to ten rival state-of-the-art approaches.

4. We introduce three datasets for phase recognition, pupil segmentation, and lens segmentation that are publicly released to support reproducibility and allow further investigations for lens irregularity detection[2].

6.2 Methodology

Figure 6.1 demonstrates the block diagram of *LensID* and the network architecture of the phase recognition and segmentation steps. As the first step towards lens irregularity detection, we adopt a recurrent convolutional network (Figure 6.1-B) to detect the lens implantation phase (the temporal segment in which the lens implantation instrument is visible). We start segmenting the artificial lens and pupil

[1]The PyTorch implementation of AdaptNet is publicly available at https://github.com/Negin-Ghamsarian/AdaptNet-MICCAI2021.
[2]http://ftp.itec.aau.at/datasets/ovid/LensID/

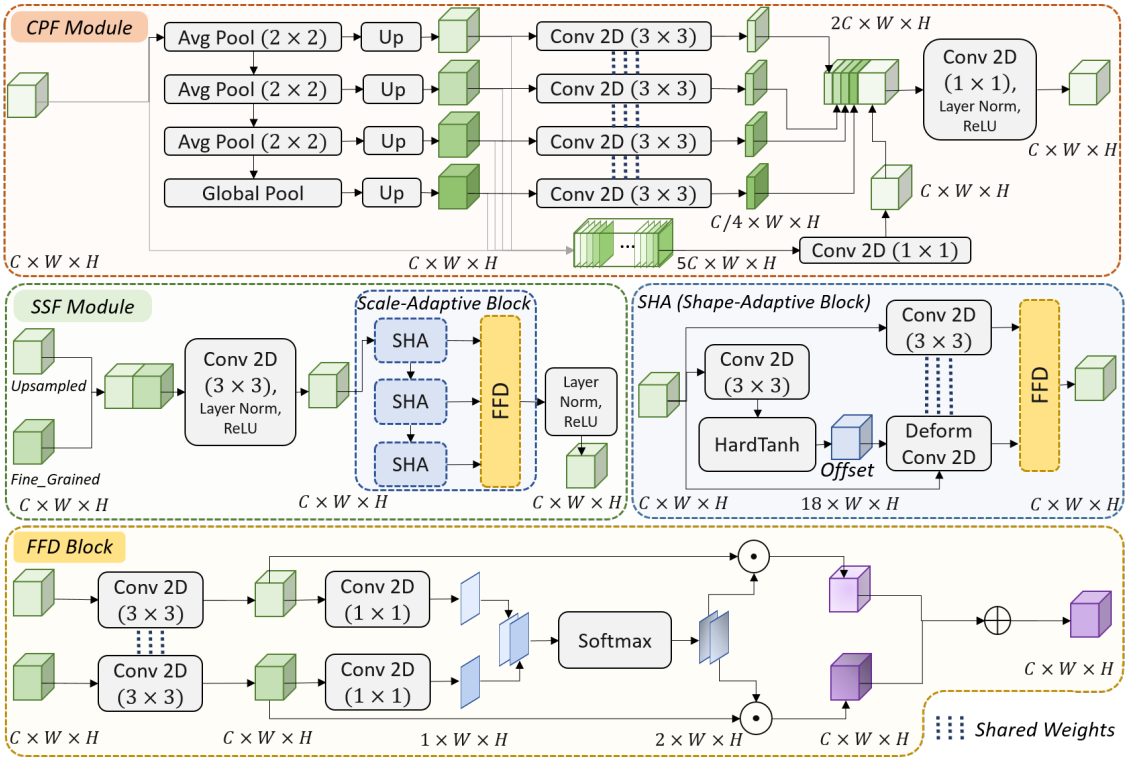


Figure 6.2: The detailed architecture of the *CPF* and *SFF* modules of AdaptNet.

exactly after the lens implantation phase using the proposed semantic segmentation network (Figure 6.1-C). The pupil and lens segmentation results undergo some post-processing approaches to compute lens instability and lens unfolding delay. More precisely, we draw the smallest convex polygon surrounding pupil's and lens' masks using binary morphological operations. For lens instability, we use the normalized distance between the lens and pupil centers. For lens unfolding, we track the lens' area over time, considering its relative position.

Phase Recognition. As shown in Figure 6.1-B, we use a pre-trained backbone followed by global average pooling to obtain a feature vector per each input frame. These features undergo a sequence of Dense, Dropout, and ReLU layers to extract higher-order semantic features. A recurrent layer with five units is then employed to improve the feature representation by taking advantage of temporal dependencies. These features are then fed into a sequence of layers to finally output the predicted class for each input frame.

Lens & Pupil Segmentation. In cataract surgery, a folded artificial lens is implanted inside the eye. The lens is transparent and inherits the pupil's color after implantation. Moreover, it is usually being unfolded very fast (sometimes with the help of an instrument). The transparency and unpredictable formation of this object, as well as occlusion, defocus blur, and motion blur [83], make lens segmentation and tracking more challenging. Hence, we require a semantic segmentation network that can be adapted to the changes in the artificial lens's shape and scale. We adopt a U-Net-based encoder-decoder architecture for the proposed semantic segmentation network termed as AdaptNet. AdaptNet consists of three main components: encoder, *cascade pooling fusion (CPF)* module, and *shape/scale-adaptive feature fusion (SSF)* module. We use the VGG16 network as the encoder network. The encoder's output feature map is fed into the CPF module to enhance the feature representation using pyramid features. This feature map is then fed into a sequence of SSF modules, which decode low-resolution semantic features.

As shown in Figure 6.2, the CPF module applies a sequence of three average pooling layers (with a stride of two pixels) followed by a global average pooling layer to the input features. The obtained feature maps are upsampled to the original size of the input and concatenated together with the input feature map in a depth-wise manner. Each group of five channels in the generated feature map undergoes a distinct convolution for intra-channel feature refinement (which is performed using a convolutional layer with C groups). Besides, the upsampled features are mapped into a smaller channel space while extracting higher-order semantic features using convolutional layers with shared weights. The obtained features are concatenated with the intra-channel refined features and undergo a convolutional layer for inter-channel feature refinement.

The *SSF* module starts with concatenating the upsampled semantic feature map with the fine-grained feature map coming from the encoder. The concatenated feature map is fed into a sequence of convolutional, layer normalization, and ReLU layers for

feature enhancement and dimensionality reduction. The resulting features are fed into the *scale-adaptive block*, which aims to fuse the features coming from cascade convolutional blocks. This succession of convolutional layers with small filter sizes can factorize the large and computationally expensive receptive fields [108]. Moreover, the fusion of these successive feature maps can play the role of scale-awareness for the network. The *shape-adaptive (SHA) block* is responsible for fusing the resulting feature maps of deformable and structured convolutions. At first, a convolutional layer followed by a hard tangent hyperbolic function is employed to produce the offsets for the deformable convolutional layer [54]. The input features are also fed into a regular convolutional layer that shares the weights with the deformable layer for structured-feature extraction. These features are then fused to induce the awareness of shape and deformation to the network.

The *feature fusion decision (FFD)* block inspired by CPFNet [69] accounts for determining the importance of each input feature map in improving semantic features. Figure 6.2 shows the *FFD Block* in the case of two input branches. At first, shared convolutional layers are applied to the input feature maps to extract the shared semantic features. The resulting feature maps undergo shared pixel-wise convolutions to produce the pixel-wise attention maps. The concatenated attention maps are fed into a softmax activation layer for normalization. The obtained features are used as pixel-wise weights of the shared-semantic feature maps. The shape/scale adaptive features are computed as the sum of pixel-wise multiplications ($\odot$) between the normalized attention maps and their corresponding semantic feature maps.

6.3 Experimental Setup

We use three datasets for this study: (i) a large dataset containing the annotations for the lens implantation phase versus the rest of phases from 100 videos of cataract surgery, (ii) a dataset containing the lens segmentation of 401 frames from 27 videos

(292 images from 21 videos for training, and 109 images from six videos for testing), and (iii) a dataset containing the pupil segmentation of 189 frames from 16 videos (141 frames from 13 videos for training, and 48 frames from three videos for testing). Regarding the phase recognition dataset, since lens implantation is a very short phase (around four seconds) compared to the whole surgery (seven minutes on average), creating a balanced dataset that can cover the entire content of videos from the "Rest" class is quite challenging. Hence, we propose a video clip generator that can provide diverse training sequences for the recurrent neural network by employing stochastic functions. At first, 12 three-second video clips with overlapping frames are extracted from the implantation phase of each cataract surgery video. Besides, the video segments before and after the implantation phase are divided into eight and four video clips, respectively (these clips have different lengths depending on the length of the input video). Accordingly, we have a balanced dataset containing 2040 video clips from 85 videos for training and 360 video clips from the other 15 videos for testing. For each training example, the video generator uses a stochastic variable to randomly select a three-second clip from the input clip. We divide this clip into N sub-clips, and N stochastic variables are used to randomly select one frame per sub-clip (in our experiments, N is set to five to reduce computational complexity and avoid network overfitting).

For phase recognition, all networks are trained for 20 epochs. The initial learning rate for these networks is set to 0.0002 and 0.0004 for the networks with VGG19 and Resnet50 backbones, respectively, and halved after ten epochs. Since the segmentation networks used for evaluations have different depths, backbones, and the number of trainable parameters, all networks are trained with three different initial learning rates ($lr_0 \in \{0.0005, 0.001, 0.002\}$). For each network, the results with the highest Dice coefficient are listed. All segmentation networks are trained for 30 epochs, and the learning rate is decreased by a factor of 0.8 in every other epoch. To prevent overfitting and improve generalization performance, we have used

motion blur, Gaussian blur, random contrast, random brightness, shift, scale, and rotation for data augmentation. The backbones of all networks evaluated for phase recognition and lens/pupil semantic segmentation are initialized with ImageNet [55] weights. The size of input images to all networks is set to $512 \times 512 \times 3$. The loss function for the phase recognition network is set to *Binary Cross Entropy*. For the semantic segmentation task, we adopt a loss function consisting of categorical cross entropy and logarithm of soft Dice coefficient as follows (in Eq. (6.1), *CE* stands for *Cross Entropy*, and $\mathcal{X}_{Pred}$ and $\mathcal{X}_{True}$ denote the predicted and ground-truth segmentation images, respectively. Besides, we use a Dice smoothing factor equal to 1, and set $\lambda = 0.8$ in our experiments):

$$\mathcal{L} = \lambda \times CE(\mathcal{X}_{Pred}, \mathcal{X}_{True}) - (1 - \lambda) \times \log_2 Dice(\mathcal{X}_{Pred}, \mathcal{X}_{True}) \tag{6.1}$$

To evaluate the performance of phase recognition networks, we use *Precision*, *Recall*, *F1-Score*, and *Accuracy*, which are the common classification metrics. The semantic segmentation performance is evaluated using *Dice coefficient* and *Intersection over Union (IoU)*. We compare the segmentation accuracy of the proposed approach (AdaptNet) with ten state-of-the-art approaches including UNet++ (and UNet++/DS) [272], MultiResUNet [108], CPFNet [69], dU-Net [264], CE-Net [90], FEDNet [43], PSPNet [268], SegNet [17], and U-Net [203]. It should be mentioned that the rival approaches employ different backbone networks, loss functions (cross entropy or cross entropy log Dice), and upsampling methods (bilinear, transposed convolution, pixel-shuffling, or max unpooling).

6.4 Experimental Results and Discussion

Table 6.1 compares the classification reports of the proposed architecture for phase recognition considering two different backbone networks and four different recurrent layers. Thanks to the large training set and taking advantage of recurrent layers,

Table 6.1: Phase recognition results of the end-to-end recurrent convolutional networks.

RNN	Backbone: VGG19				Backbone: ResNet50			
	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score	Accuracy
GRU	0.97	0.96	0.96	0.96	0.9	0.94	0.94	0.94
LSTM	0.98	0.98	0.98	0.98	0.96	0.96	0.96	0.96
BiGRU	0.97	0.96	0.96	0.96	0.95	0.95	0.95	0.95
BiLSTM	1.00	1.00	1.00	1.00	0.98	0.98	0.98	0.98

all networks have shown superior performance in classifying the implantation phase versus other phases. However, the LSTM and bidirectional LSTM (BiLSTM) layers have shown better performance compared to GRU and BiGRU layers, respectively. Surprisingly, the network with a VGG19 backbone and BiLSTM layer has achieved 100% accuracy in classifying the test clips extracted from the videos which are not used during training. Figure 6.3 compares the segmentation results (mean and standard deviation of IoU and Dice coefficient) of AdaptNet and ten rival state-of-the-art approaches. Overall, it can be perceived that AdaptNet, UNet++, UNet++/DS, and FEDNet have achieved the top four segmentation results. However, AdaptNet has achieved the highest mean IoU and Dice coefficient compared to the rival approaches. In particular, the proposed approach achieves 3.48% improvement in mean IoU and 2.22% improvement in mean Dice for lens segmentation compared to the best rival approach (UNet++). Moreover, the smaller standard deviation of IoU (10.56% vs. 12.34%) and Dice (8.56% vs. 9.65%) for AdaptNet compared to UNet++ confirms the reliability and effectiveness of the proposed architecture. For pupil segmentation, AdaptNet shows subtle improvement over the best rival approach (UNet++) regarding mean IoU and Dice while showing significant improvement regarding the standard deviation of IoU (1.91 vs. 4.05). Table 6.2 provides an ablation study of AdaptNet. We have listed the Dice and IoU percentage with two different learning rates by gradually adding the proposed modules and blocks (for lens segmentation). It can be perceived from the results that regardless of the learning

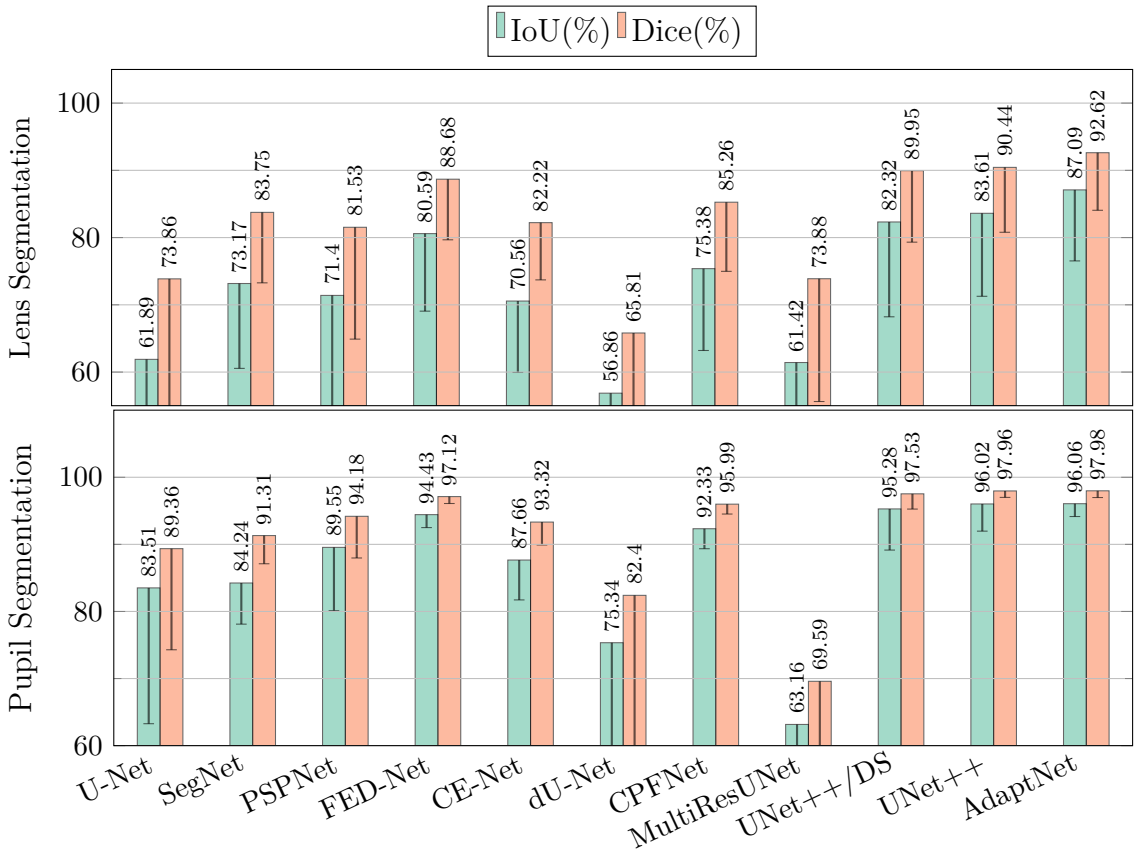


Figure 6.3: Quantitative comparison of segmentation results for the proposed approach (AdaptNet) and rival approaches.

Table 6.2: Impact of different modules on the segmentation results of AdaptNet.

Components				lr = 0.001		lr = 0.002	
Baseline	SSF	SHA	CPF	IoU(%)	Dice(%)	IoU(%)	Dice(%)
✓	✗	✗	✗	82.79	89.94	84.33	90.90
✓	✓	✗	✗	83.54	90.33	84.99	91.22
✓	✓	✓	✗	84.76	91.12	86.34	92.17
✓	✓	✓	✓	85.03	91.28	87.09	92.62

rate, each distinctive module and block has a positive impact on segmentation performance. We cannot test the FFD block separately since it is bound with the SSF module.

Figure 6.4 shows the post-processed lens segments (pink) and pupil segments (cyan) from a representative video in different time slots (a), the relative lens area over time (b), and relative lens movements over time (c). Due to lens instability, a part of the lens is sometimes placed behind the iris, as shown in the segmentation results in the

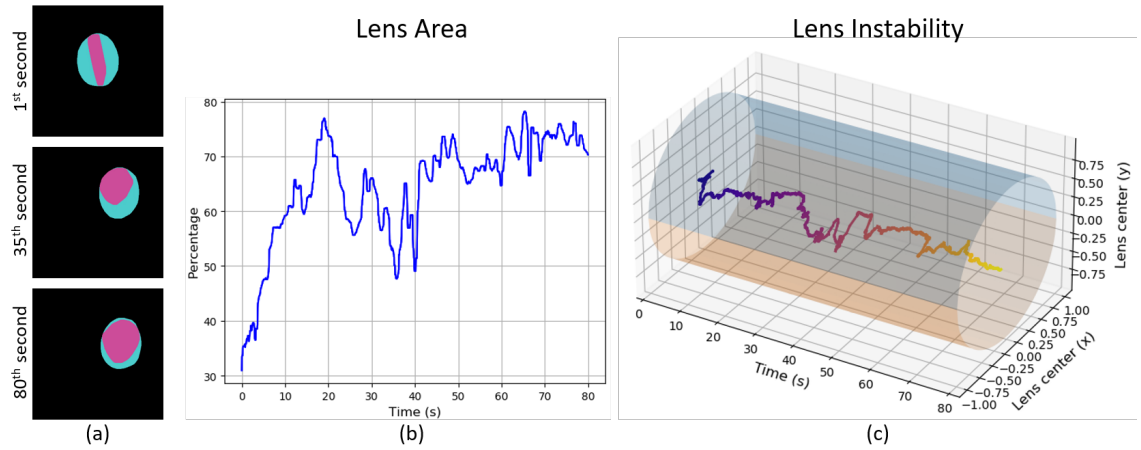


Figure 6.4: The lens statistics for one representative cataract surgery video.

35th second. Accordingly, the visible area of the lens can change independently of the unfolding state. Hence, the relative position of the lens should also be taken into account for lens unfolding delay computations. As can be perceived, the visible area of the lens is near maximum at 20 seconds after the implantation phase, and the lens is located nearly at the center of the pupil at this time. Therefore, the lens unfolding delay is 20 seconds in this case. However, the lens is quite unstable until 70 seconds after implantation.

6.5 Conclusion

Lens irregularity detection is a highly relevant problem in ophthalmology, which can play a prominent role in predicting and preventing lens relocation after surgery. This study focuses on two significant steps towards lens irregularity detection: (i) "lens implantation phase" detection and (ii) lens/pupil segmentation. We propose an end-to-end recurrent convolutional network to detect the lens implantation phase. Moreover, we propose a novel semantic segmentation network termed as AdaptNet. The proposed approach can deal with severe deformations and scale variations in the intraocular lens by adaptively fusing sequential and parallel feature maps. Experimental results reveal the effectiveness of the proposed phase recognition and

semantic segmentation networks.

Appendix 1

Table 6.3 lists the specifications of the rival state-of-the-art approaches used in our evaluations. In "Upsampling" column, "Trans Conv" stands for *Transposed Convolution*.

Table 6.3: Specifications of the proposed and rival segmentation approaches.

Model	Backbone	Params	Upsampling	Reference	Year
UNet++ (/DS)	VGG16	24.24 M	Bilinear	[272]	2020
MultiResUNet	✗	9.34 M	Trans Conv	[108]	2020
CPFNet	ResNet34	34.66 M	Bilinear	[69]	2020
dU-Net	✗	31.98 M	Trans Conv	[264]	2020
CE-Net	ResNet34	29.90 M	Trans Con	[90]	2019
FED-Net	ResNet50	59.52 M	Trans Conv & PixelShuffle	[43]	2019
PSPNet	ResNet50	22.26 M	Bilinear	[268]	2017
SegNet	VGG16	14.71 M	Max Unpooling	[17]	2017
U-Net	✗	17.26 M	Bilinear	[203]	2015
AdaptNet	VGG16	23.61 M	Bilinear	Proposed	

Figure 6.5 presents qualitative comparisons among the top five approaches for lens segmentation in three representative frames. It can be perceived from the figure that AdaptNet can provide the most visually close segmentation results to the ground truth. Moreover, AdaptNet is more robust against lens deformations as it provides the most delineated predictions compared to the rival approaches.

Figure 6.6 demonstrates the effect of post-processing on segmentation results. We use three morphological operations to improve the semantic segmentation results: (i) opening (with the kernel size of 10×10) to attach the separated regions due to instrument covering, (ii) closing (with the kernel size of 15×15) to remove the distant wrong detections, and (iii) convex polygon. Since instruments usually cover a part of the pupil and intraocular lens during surgery, the segmentation results may contain some holes in the location of instruments. However, the occluded parts should be included in the lens and pupil area. Since the pupil is inherently a convex

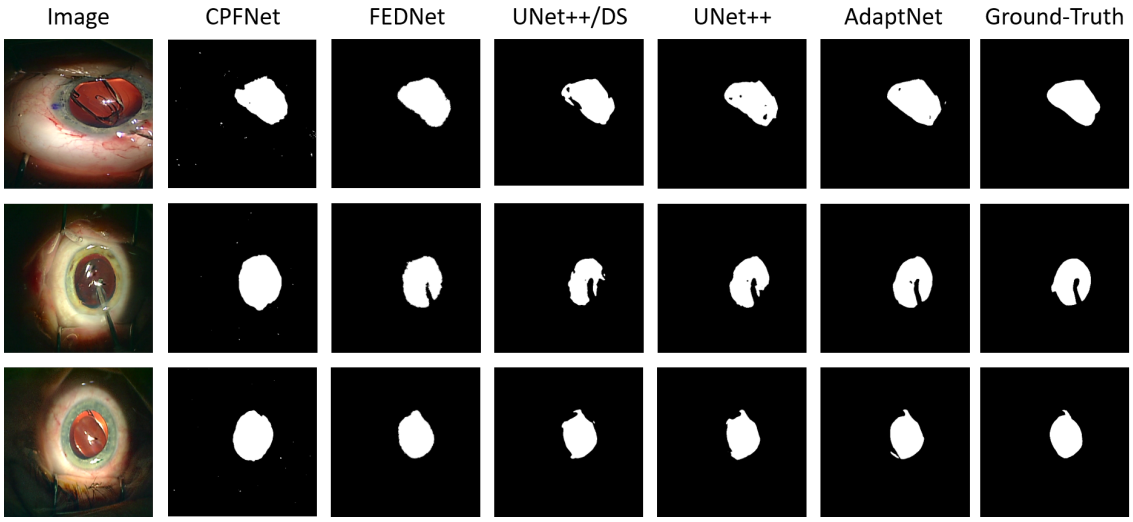


Figure 6.5: Qualitative comparisons among the top five segmentation approaches.

object, and the intraocular lens is usually convex during unfolding, we draw the smallest convex polygon around these objects to retrieve the occluded segments. For convex polygons, we used the "Scipy ConvexHull" function.

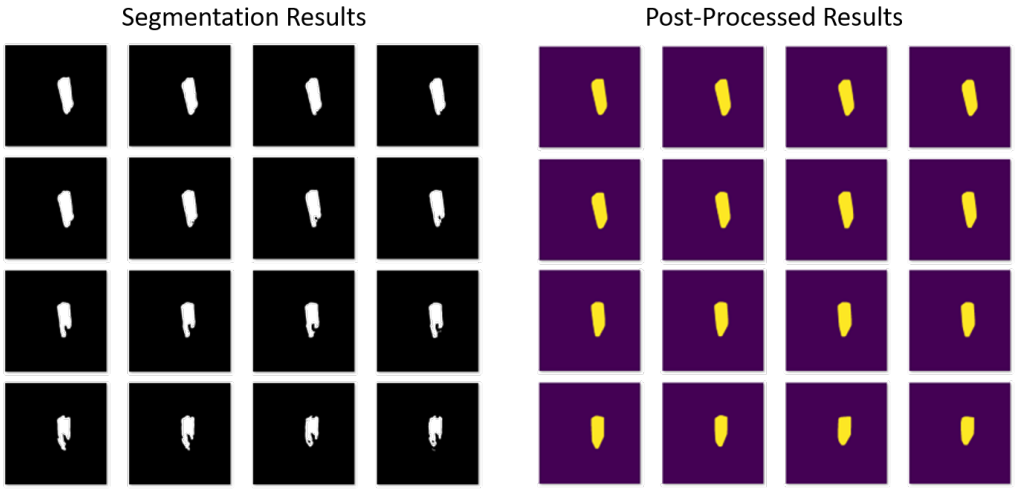


Figure 6.6: Qualitative comparisons among the top five segmentation approaches.

Appendix 2

Herein, we highlight the application of the proposed LensID framework in comparing the statistical distributions of different intraocular lenses. In particular, we compare three types of statistics naming lens unfolding time (per seconds), lens instability

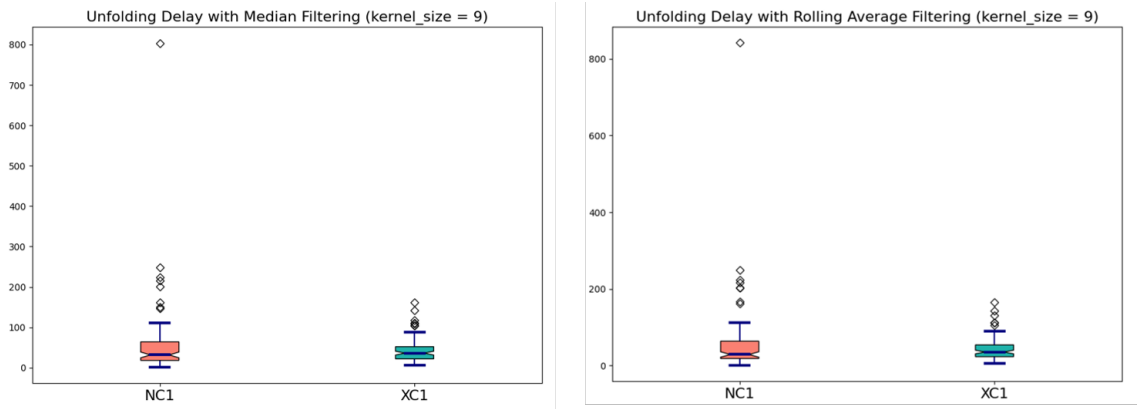


Figure 6.7: Statistical comparison between the unfolding delay of NC1 and XC1 lenses.

(based on absolute lens' relative movements), and lens rotation (based on absolute lens' degree changes) for two brands of IoL labeled as "NC1" and "XC1".

6.5.1 Lens Unfolding Delay

Figure 6.7 compares the distribution of lens unfolding time for 100 surgeries with NC1 lenses versus 100 surgeries with XC1 lenses using boxplot. This figure reveals that the interquartile range (IQR) of XC1 lenses is more compact than that of NC1 lenses. Moreover, there are many outliers being far away from the median value in the distribution of NC1 lenses. These results suggest that XC1 lenses have a more predictable behavior compared to NC1 lenses.

6.5.2 Lens Instability

Lens instability is computed based on the sum of the lens' absolute relative movements inside the pupil. As shown in Figure 6.8, both NC1 and XC1 lenses offer a relatively close interquartile range (IQR). However, NC1 lenses have a larger value in the lower and upper whisker of lens instability distribution. Besides, the NC1 group has a very distant outlier in the instability distribution.

Moving on to Figure 6.9, we can infer that there are some correlations between the lens unfolding time and lens instability. Indeed, NC1 and XC1 lenses' statistical

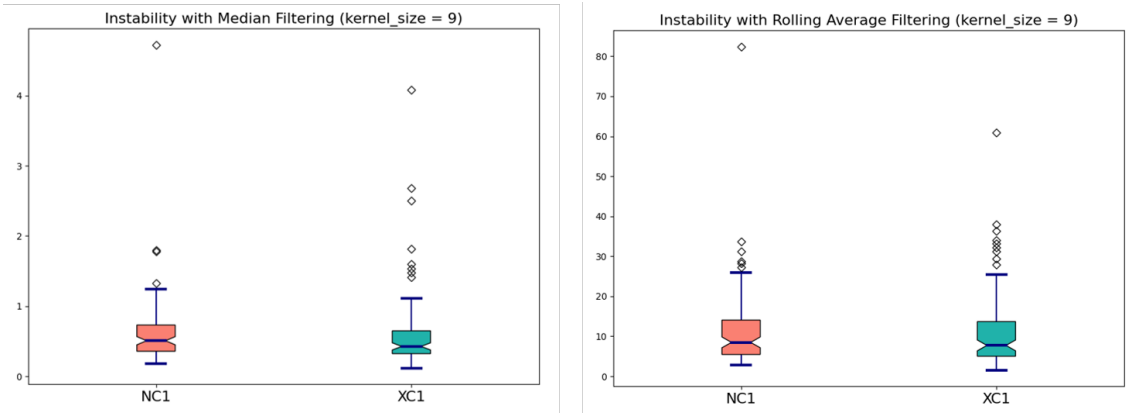


Figure 6.8: Statistical comparison between the instability of NC1 and XC1 lenses.

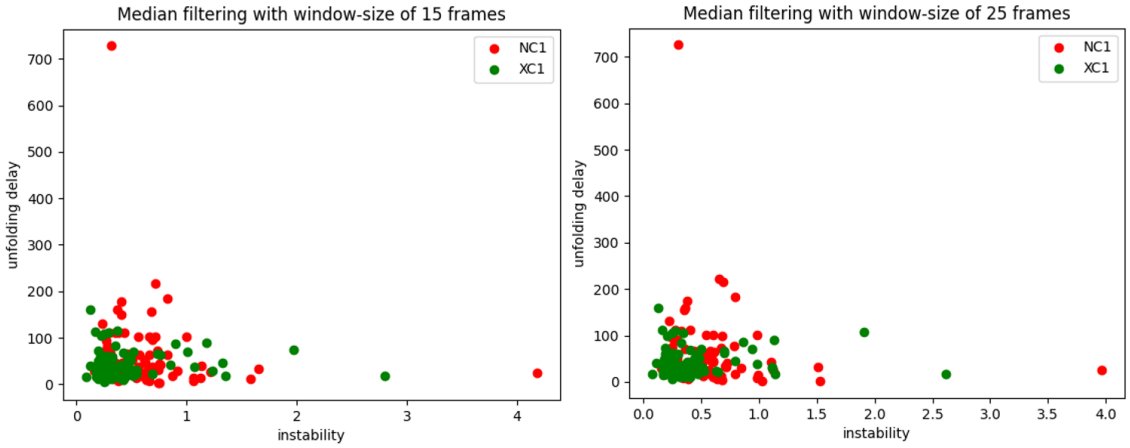


Figure 6.9: Joint distribution of lens unfolding delay and lens instability for NC1 and XC1 lenses.

distribution suggests that lenses with more unfolding delay are usually more unstable. Besides, the distribution of XC1 lenses is compact and focused near the origin, whereas the distribution of XC1 lenses is more scattered and farther from the origin. This suggests NC1 lenses have some irregular behavior.

6.5.3 Lens Rotation

For lens rotation computation, we need a pose estimation approach in addition to the LensID framework. We use Faster R-CNN to compute the degree of IoL when its hooks are visible. We start calculating the lens rotation statistics after lens unfolding. Then lens rotation is calculated based on the sum of absolute IoL degree changes over time. Figure 6.10-left compares the boxplots of rotation irregularity for

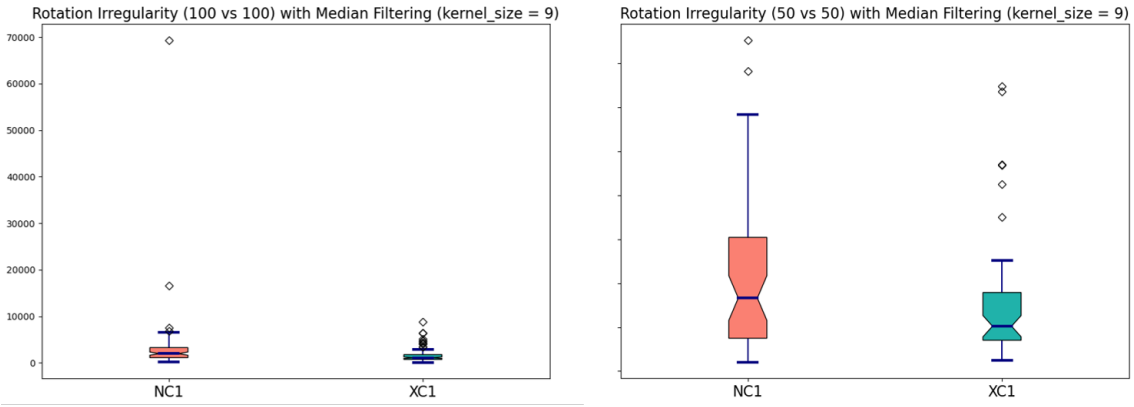


Figure 6.10: Statistical comparison between the rotation of NC1 and XC1 lenses.

NC1 vs. XC1 lenses (100 vs. 100). It is evident from the figure that the NC1 lens' distribution contains very distant outliers. For better visualization, we have also compared the distributions of 50 NC1 lenses with 50 XC1 lenses in Figure 6.10-right. It is evident from the boxplots that the NC1 lens' rotation distribution has a very high upper whisker being more than two times the upper whisker of the XC1 lens' rotation distribution. Besides, the NC1 lens's rotation distribution has a much larger median value and much wider interquartile range compared to the NC1 lens' rotation distribution. This suggests NC1 lenses have less predictable behavior compared to XC1 lenses.

CHAPTER

7 Semantic Segmentation using ReCal-Net

Chapter overview — In this chapter, we propose a novel convolutional module termed as *ReCal* module, which can calibrate the feature maps by employing region intra-and-inter-dependencies and channel-region cross-dependencies. This calibration strategy can effectively enhance semantic representation by correlating different representations of the same semantic label, considering a multi-angle local view centering around each pixel. Thus the proposed module can deal with distant visual characteristics of unique objects as well as cross-similarities in the visual characteristics of different objects. Moreover, we propose a novel network architecture based on the proposed module termed as *ReCal-Net*. Experimental results confirm the superiority of ReCal-Net compared to rival state-of-the-art approaches for all relevant objects in cataract surgery. Moreover, ablation studies reveal the effectiveness of the ReCal module in boosting semantic segmentation accuracy.

This chapter is an adapted version of:

"Ghamsarian, N., Taschwer, M., Putzgruber-Adamitsch, D., Sarny, S., El-Shabrawi, Y., and Schoeffmann, K. ReCal-Net: Joint Region-Channel-Wise Calibrated Network for Semantic Segmentation in Cataract Surgery Videos. Under Review."

7.1 Introduction

In recent years, a large body of research has been focused on computerized surgical workflow analysis in cataract surgery [232, 82, 79, 164, 121, 79], with a majority of approaches relying on semantic segmentation. Hence, improving semantic segmenta-

tion accuracy in cataract surgery videos can play a leading role in the development of a reliable computerized clinical diagnosis or surgical analysis approach [176, 175]. Semantic segmentation of the relevant objects in cataract surgery videos is quite challenging due to (i) transparency of the intraocular lens, (ii) color and contextual variation of the pupil and iris, (iii) blunt edges of the iris, and (iv) severe motion blur and reflection distortion of the instruments. In this chapter, we propose a novel module for joint Region-channel-wise Calibration, termed as *ReCal* module. The proposed module can simultaneously deal with the various segmentation challenges in cataract surgery videos. In particular, the ReCal module is able to (1) employ multi-angle pyramid features centered around each pixel position to deal with transparency, blunt edges, and motion blur, (2) employ cross region-channel dependencies to handle texture and color variation through interconnecting the distant feature vectors corresponding to the same object. The proposed module can be added on top of every convolutional layer without changing the output feature dimensions. Moreover, the ReCal module does not impose a significant number of trainable parameters on the network and thus can be used after several layers to calibrate their output feature maps. Besides, we propose a novel semantic segmentation network based on the ReCal module termed as *ReCal-Net*. The experimental results show significant improvement in semantic segmentation of the relevant objects with ReCal-Net compared to the best-performing rival approach (85.38% compared to 83.32% overall IoU (intersection over union) for ReCal-Net vs. UNet++).

The rest of this work is organized as follows. In Section 7.2, we first discuss two convolutional blocks from which the proposed approach is inspired, and then delineate the proposed ReCal-Net and ReCal module. We detail the experimental settings in Section 7.3 and present the experimental results in Section 7.4. We finally conclude the paper in Section 7.5.

7.2 Methodology

Notations. Everywhere in this chapter, we show convolutional layer with the kernel-size of $(m \times n)$, P output channels, and g groups as $*^{P,g}_{(m\times n)}$ (we consider the default dilation rate of 1 for this layer). Besides, we show average-pooling layer with a kernel-size of $(m \times n)$ and a stride of s pixels as $\sum^{s}_{(m\times n)}$, and global average pooling as $\sum^{G}$.

Feature Map Recalibration. The Squeeze-and-Excitation (SE) block [103] was proposed to model inter-channel dependencies through squeezing the spatial features into a channel descriptor, applying fully-connected layers, and rescaling the input feature map via multiplication. This low-complexity operation unit has proved to be effective, especially for semantic segmentation. However, the SE block does not consider pixel-wise features in recalibration. Accordingly, scSE block [205] was proposed to exploit pixel-wise and channel-wise information concurrently. This block can be split into two parallel operations: (1) spatial squeeze and channel excitation, exactly the same as the SE block, and (2) channel squeeze and spatial excitation. The latter operation is conducted by applying a pixel-wise convolution with one output channel to the input feature map, followed by multiplication. The final feature maps of these two parallel computational units are then merged by selecting the maximum feature in each location.

ReCal-Net. Figure 7.1 depicts the architecture of ReCal-Net. Overall, the network consists of three types of blocks: (i) encoder blocks that transform low-level features to semantic features while compressing the spatial representation, (ii) decoder blocks that are responsible for improving the semantic features in higher resolutions by employing the symmetric low-level feature maps from the encoder blocks, (iii) and *ReCal* modules that account for calibrating the semantic feature maps. We use the VGG16 network as the encoder network. The ith encoder block (E$i, i \in \{1, 2, 3, 4\}$)

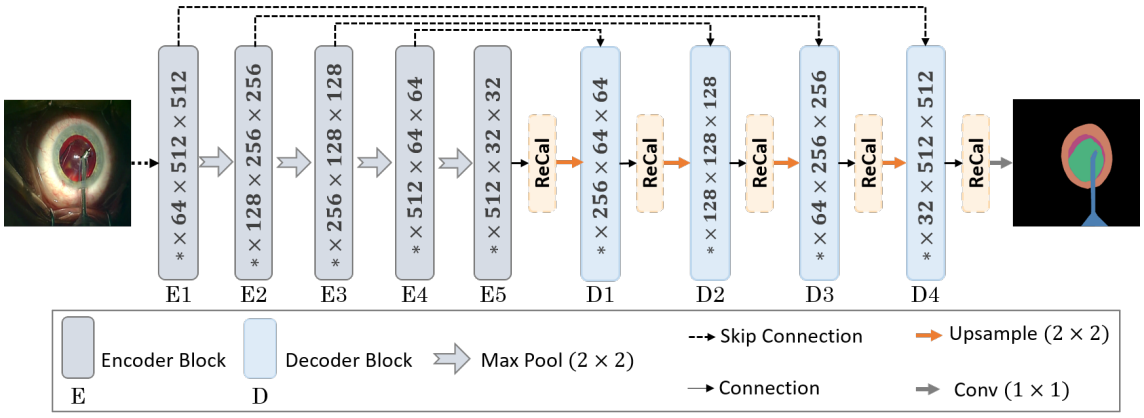


Figure 7.1: The overall architecture of ReCal-Net containing five ReCal blocks.

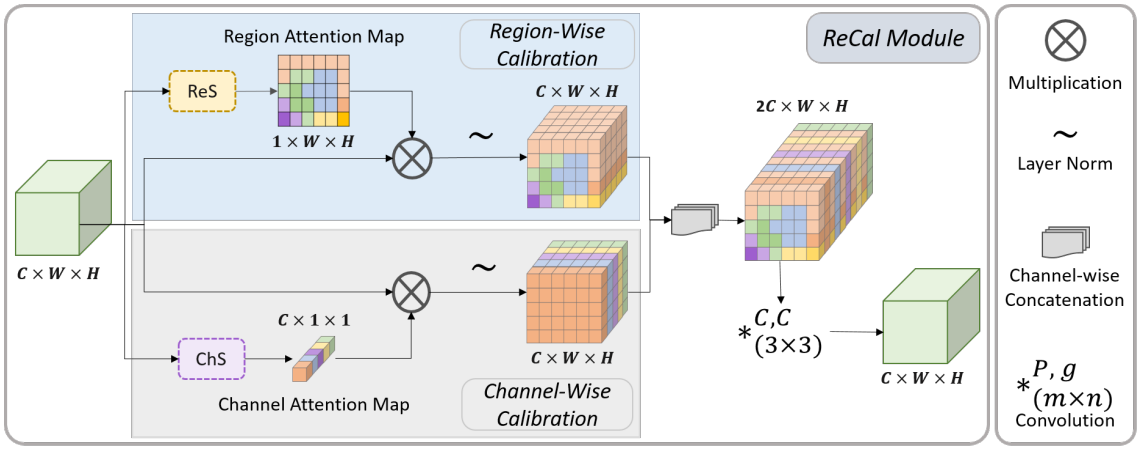


Figure 7.2: The detailed architecture of ReCal block containing regional squeeze block (ReS) and channel squeeze block (ChS).

in Figure 7.1 correspond to all layers between the *i-1*th and *i*th max-pooling layers in the VGG16 network (max-pooling layers are indicated with gray arrows). The last encoder block (E5) corresponds to the layers between the last max-pooling layer and the average pooling layer. Each decoder block follows the same architecture of decoder blocks in U-Net [203], including two convolutional layers, each of which being followed by batch normalization and ReLU.

ReCal Module. Despite the effectiveness of SE and scSE blocks in boosting feature representation, both fail to exploit region-wise dependencies. However, employing region-wise inter-dependencies and intra-dependencies can significantly enhance semantic segmentation performance. We propose a joint region-channel-wise cali-

bration (ReCal) module to calibrate the feature maps based on joint region-wise and channel-wise dependencies. Figure 7.2 demonstrates the architecture of the proposed ReCal module inspired by [103, 205]. This module aims to reinforce a semantic representation considering inter-channel dependencies, inter-region and intra-region dependencies, and channel-region cross-dependencies. The input feature map of ReCal module $\mathcal{F}_{In} \in \mathbb{R}^{C\times W\times H}$ is first fed into two parallel blocks: (1) the Region-wise Squeeze block (ReS), and (2) the Channel-wise Squeeze block (ChS). Afterward, the region-wise and channel-wise calibrated features ($\mathcal{F}_{Re} \in \mathbb{R}^{C\times W\times H}$ and $\mathcal{F}_{Ch} \in \mathbb{R}^{C\times W\times H}$) are obtained by multiplying ($\otimes$) the input feature map to the region-attention map and channel-attention map, respectively, followed by the layer normalization function. In this stage, each particular channel $\mathcal{F}_{In}(C_j) \in \mathbb{R}^{W\times H}$ in the input feature map of a ReCal module has corresponding region-wise and channel-wise calibrated channels ($\mathcal{F}_{Re}(C_j) \in \mathbb{R}^{W\times H}$ and $\mathcal{F}_{Ch}(C_j) \in \mathbb{R}^{W\times H}$). To enable the utilization of cross-dependencies between the region-wise and channel-wise calibrated features, we concatenate these two feature maps in a depth-wise manner. Indeed, the concatenated feature map ($\mathcal{F}_{Concat}$) for each $p \in [1, C]$, $x \in [1, W]$, and $y \in [1, H]$ can be formulated as (7.1).

$$\begin{cases} \mathcal{F}_{Concat}(2p, x, y) = \mathcal{F}_{Re}(p, x, y) \\ \mathcal{F}_{Concat}(2p-1, x, y) = \mathcal{F}_{Ch}(p, x, y) \end{cases} \tag{7.1}$$

The cross-dependency between region-wise and channel-wise calibrated features is computed using a convolutional layer with C groups. More concretely, every two consecutive channels in the concatenated feature map undergo a distinct convolution with a kernel-size of (3×3). This convolutional layer considers the local contextual features around each pixel (a 3×3 window around each pixel) to determine the contribution of each of region-wise and channel-wise calibrated features in the output features. Using a kernel size greater than one unit allows jointly considering inter-region dependencies.

Region-Wise Squeeze Block. Figure 7.3 details the architecture of the ReS block, which is responsible for providing the region attention map. The region attention map is obtained by taking advantage of multi-angle local content based on narrow to wider views around each distinct pixel in the input feature map. We model multi-angle local features using average pooling layers with different kernel sizes and the stride of one pixel. The average pooling layers do not any number of impose trainable parameters on the network and thus ease using the ReS block and ReCal module in multiple locations. Besides, the stride of one pixel in the average pooling layer can stimulate a local view centered around each distinctive pixel. We use three average pooling layers with kernel-sizes of (3×3), (5×5), and (7×7), followed by pixel-wise convolutions with one output channel $(*^{1,1}_{(1\times1)})$ to obtain the region-wise descriptors. In parallel, the input feature map undergoes another convolutional layer to obtain the pixel-wise descriptor. The local features can indicate if some particular features (could be similar or dissimilar to the centering pixel) exist in its neighborhood, and how large is the neighborhood of each pixel containing particular features. The four attention maps are then concatenated and fed into a convolutional layer $(*^{1,1}_{(1\times1)})$ that is responsible for determining the contribution of each spatial descriptor in the final region-wise attention map.

Channel-Wise Squeeze Block. For ChS Block, we follow a similar scheme as in [103]. At first, we apply global average pooling $(\sum^{G})$ on the input convolutional feature map. Afterward, we form a bottleneck via a pixel-wise convolution with C/r output channels $(*^{C/r,1}_{(1\times1)})$ followed by ReLU non-linearity. The scaling parameter r can curbs the computational complexity. Besides, it can act as a smoothing factor that can yield a better-generalized model by preventing the network from learning outliers . In experiments, we set $r = 2$ as it is proved to have the best performance [205]. Finally, another pixel-wise convolution with C output channels $(*^{C,1}_{(1\times1)})$ followed by ReLU non-linearity is used to provide the channel attention map.

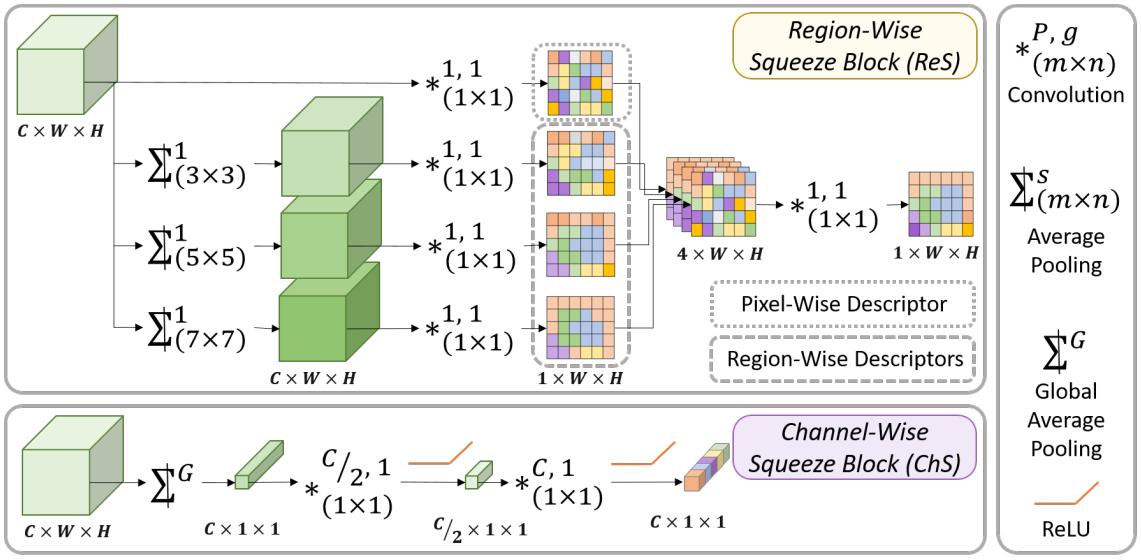


Figure 7.3: Demonstration of regional squeeze block (ReS) and channel squeeze block (CS).

Module Complexity. Suppose we have an intermediate layer in the network with convolutional response map $\mathcal{X} \in \mathbb{R}^{C\times H\times W}$. Adding a ReCal module on top of this layer with its scaling parameter being equal to 2, amounts to "C^2+22C+4" additional trainable weights. More specifically, each convolutional layer $*^{P,g}_{(m\times n)}$ applied to C input channels amounts to ${}^{((m\times n)\times C\times P)}/_{g}$ trainable weights. Accordingly, we need "$4C+4$" weights for the ReS block, "C^2" weights for the ChS block, and "$18C$" weights for the last convolution operation of the ReCal module. In our proposed architecture, adding five ReCal modules on convolutional feature maps with 512, 256, 128, 64, and 32 channels sums up to $371K$ additional weights, and only $21K$ more trainable parameters compared to the SE block [103] and scSE block [205].

7.3 Experimental Settings

Datasets. We use four datasets in this study. The iris dataset is created by annotating the cornea and pupil from 14 cataract surgery videos using "supervisely" platform. The iris annotations are then obtained by subtracting the convex-hull

Table 7.1: Specifications of the proposed and rival segmentation approaches.

Model	Backbone	Params	Upsampling	Reference	Year
UNet++ (/DS)	VGG16	24.24 M	Bilinear	[272]	2020
MultiResUNet	✗	9.34 M	Trans Conv	[108]	2020
BARNet	ResNet34	24.90 M	Bilinear	[176]	2020
PAANet	ResNet34	22.43 M	Trans Conv & Bilinear	[175]	2020
CPFNet	ResNet34	34.66 M	Bilinear	[69]	2020
dU-Net	✗	31.98 M	Trans Conv	[264]	2020
CE-Net	ResNet34	29.90 M	Trans Conv	[90]	2019
scSE-Net	VGG16	22.90 M	Bilinear	[205]	2019
U-Net	✗	17.26 M	Bilinear	[203]	2015
ReCal-Net	VGG16	22.92 M	Bilinear	Proposed	

of the pupil segment from the cornea segment. This dataset contains 124 frames from 12 videos for training and 23 frames from two videos for testing. For lens and pupil segmentation, we employ the two public datasets of the LensID framework [80], containing the annotation of the intraocular lens and pupil. The lens dataset consists of lens annotation in 401 frames sampled from 27 videos. From these annotations, 292 frames from 21 videos are used for training, and 109 frames from the remaining six videos are used for testing. The pupil segmentation dataset contains 189 frames from 16 videos. The training set consists of 141 frames from 13 videos, and the testing set contains 48 frames from three remaining videos. For instrument segmentation, we use the instrument annotations of the CaDIS dataset [89]. We use 3190 frames from 18 videos for training and 459 frames from three other videos for testing.

Rival Approaches. Table 7.1 lists the specifications of the rival state-of-the-art approaches used in our evaluations. In "Upsampling" column, "Trans Conv" stands for *Transposed Convolution*. To enable direct comparison between the ReCal module and scSE block, we have formed scSE-Net by replacing the ReCal modules in ReCal-Net with scSE modules. Indeed, the baseline of both approaches are the same, and the only difference is the use of scSE blocks in scSE-Net at the position of ReCal modules in ReCal-Net.

Data Augmentation Methods. We use the Albumentations [27] library for image and mask augmentation during training. Considering the inherent features of the relevant objects and problems of the recorded videos [83], we apply motion blur, median blur, brightness and contrast change, shifting, scaling, and rotation for augmentation. We use the same augmentation pipeline for the proposed and rival approaches.

Neural Network Settings. We initialize the parameters of backbones for the proposed and rival approaches (in case of having a backbone) with ImageNet [55] training weights. We set the input size of all networks to $(3 \times 512 \times 512)$.

Training Settings. During training with all networks, a threshold of 0.1 is applied for gradient clipping. This strategy can prevent the gradient from exploding and result in a more appropriate behavior during learning in the case of irregularities in the loss landscape. Considering the different depths and connections of the proposed and rival approaches, all networks are trained with two different initial learning rates ($lr \in \{0.005, 0.002\}$) for 30 epochs with SGD optimizer. The learning scheduler decreases the learning rate every other epoch with a factor of 0.8. We list the results with the highest IoU for each network.

Loss Function. To provide a fair comparison, we use the same loss function for all networks. The loss function is set to a weighted sum of binary cross-entropy (BCE) and the logarithm of soft Dice coefficient as follows.

$$\begin{aligned}\mathcal{L} =&(\lambda) \times BCE(\mathcal{X}_{true}(i,j), \mathcal{X}_{pred}(i,j)) \\ &- (1-\lambda) \times (\log \frac{2 \sum \mathcal{X}_{true} \odot \mathcal{X}_{pred} + \sigma}{\sum \mathcal{X}_{true} + \sum \mathcal{X}_{pred} + \sigma})\end{aligned} \tag{7.2}$$

Soft Dice refers to the dice coefficient computed directly based on predicted probabilities rather than the predicted binary masks after thresholding. In (7.2), $\mathcal{X}_{true}$ refers to the ground truth mask, $\mathcal{X}_{pred}$ refers to the predicted mask, $\odot$ refers to

Table 7.2: Quantitative comparisons among the semantic segmentation results of Recal-Net and rival approaches based on IoU(%).

Network	Lens	Pupil	Iris	Instruments	Overall
U-Net	61.89 ±20.93	83.51 ±20.24	65.89 ±16.93	60.78 ±26.04	68.01 ±21.03
CE-Net	78.51 ±11.56	92.07 ± 4.24	71.74 ± 6.19	69.44 ±17.94	77.94 ± 9.98
dU-Net	60.39 ±29.36	68.03 ±35.95	70.21 ±12.97	61.24 ±27.64	64.96 ±26.48
scSE-Net	86.04 ±11.36	96.13 ± 2.10	78.58 ± 9.61	71.03 ±23.25	82.94 ±11.58
CPFNet	80.65 ±12.16	93.76 ± 2.87	77.93 ± 5.42	69.46 ±17.88	80.45 ± 9.58
BARNet	80.23 ±14.57	93.64 ± 4.11	75.80 ± 8.68	69.76 ±21.29	79.86 ±12.16
PAANet	80.30 ±11.73	94.35 ± 3.88	75.73 ±11.67	68.01 ±22.29	79.59 ±12.39
MultiResUNet	61.42 ±19.91	76.46 ±29.43	49.99 ±28.73	61.01 ±26.94	62.22 ±26.25
UNet++/DS	84.53 ±13.42	96.18 ± 2.62	74.01 ±13.13	65.99 ±25.66	79.42 ±14.75
UNet++	85.74 ±11.16	96.50 ± 1.51	81.98 ± 6.96	69.07 ±23.89	83.32 ±10.88
ReCal-Net	**87.94** ±10.72	**96.58** ± 1.30	**85.13** ± 3.98	**71.89** ±19.93	**85.38** ± 8.98

Table 7.3: Quantitative comparisons among the semantic segmentation results of Recal-Net and rival approaches based on Dice(%).

Network	Lens	Pupil	Iris	Instruments	Overall
U-Net	73.86 ±20.39	89.36 ±15.07	78.12 ±13.01	71.50 ±25.77	78.21 ±18.56
CE-Net	87.32 ± 9.98	95.81 ± 2.39	83.39 ± 4.25	80.30 ±15.97	86.70 ± 8.15
dU-Net	69.99 ±29.40	73.72 ±34.24	81.76 ± 9.73	71.30 ±27.62	74.19 ±25.24
scSE-Net	91.95 ± 9.14	98.01 ± 1.10	87.66 ± 6.35	80.18 ±21.49	89.45 ± 9.52
CPFNet	88.61 ±10.20	96.76 ± 1.53	87.48 ± 3.60	80.33 ±15.85	88.29 ± 7.79
BARNet	88.16 ±10.87	96.66 ± 2.30	85.95 ± 5.73	79.72 ±19.95	87.62 ± 9.71
PAANet	88.46 ± 9.59	97.05 ± 2.16	85.62 ± 8.50	78.15 ±21.51	87.32 ±10.44
MultiResUNet	73.88 ±18.26	82.45 ±25.49	61.78 ±25.96	71.35 ±26.88	72.36 ±24.14
UNet++/DS	90.80 ±11.41	98.03 ± 1.41	84.38 ± 9.06	75.64 ±25.38	87.21 ±11.81
UNet++	91.80 ±11.16	98.26 ± 0.79	89.93 ± 4.51	78.54 ±22.76	89.63 ± 9.80
ReCal-Net	**93.09** ± 8.56	**98.26** ± 0.68	**91.91** ± 2.47	**81.62** ±17.75	**91.22** ± 7.36

Hadamard product (element-wise multiplication), and σ refers to the smoothing factor. In experiments, we set $\lambda = 0.8$ and $\sigma = 1$.

Soft Dice refers to the dice coefficient computed directly based on predicted probabilities rather than the predicted binary masks after thresholding. In (7.2), $\mathcal{X}_{true}$ refers to the ground truth mask, $\mathcal{X}_{pred}$ refers to the predicted mask, $\odot$ refers to Hadamard product (element-wise multiplication), and σ refers to the smoothing factor. In experiments, we set $\lambda = 0.8$ and $\sigma = 1$.

7.4 Experimental Results

Table 7.2 and Table 7.3 compare the segmentation performance of ReCal-Net and ten rival state-of-the-art approaches based on the average and standard deviation of

IoU and Dice coefficient, respectively [1]. Overall, ReCal-Net, UNet++, scSE-Net, and CPFNet have shown the top four segmentation results. Moreover, the experimental results reveal that ReCal-Net has achieved the highest average IoU and Dice coefficient for all relevant objects compared to state-of-the-art approaches. Considering the IoU report, ReCal-Net has gained considerable enhancement in segmentation performance compared to the second-best approach in lens segmentation (87.94% vs. 86.04% for scSE-Net) and iris segmentation (85.13% vs. 81.98% for UNet++). Having only 21k more trainable parameters than scSE-Net (0.08% additive trainable parameters), ReCal-Net has achieved 8.3% relative improvement in iris segmentation, 2.9% relative improvement in instrument segmentation, and 2.2% relative improvement in lens segmentation in comparison with scSE-Net. Regarding the Dice coefficient, ReCal-Net and UNet++ show very similar performance in pupil segmentation. However, with 1.32M fewer parameters than UNet++ as the second-best approach, ReCal-Net shows 1.7% relative improvement in overall Dice coefficient (91.22% vs. 89.63%). Surprisingly, replacing the scSE blocks with the ReCal modules results in 4.25% higher Dice coefficient for iris segmentation and 1.44% higher Dice coefficient for instrument segmentation.

Figure 7.4 provides qualitative comparisons among the top four segmentation approaches for lens, iris, and instrument segmentation. Comparing the visual segmentation results of ReCal-Net and scSE-Net further highlights the effectiveness of region-wise and cross channel-region calibration in boosting semantic segmentation performance.

Table 7.4 reports the ablation study by comparing the segmentation performance of the baseline approach with ReCal-Net considering two different learning rates. The baseline approach refers to the network obtained after removing all ReCal modules of ReCal-Net in Figure 7.1. These results approve of the ReCal module's effectiveness regardless of the learning rate.

[1]The "Overall" column in Table 7.2 and Table 7.3 is the mean of the other four values.

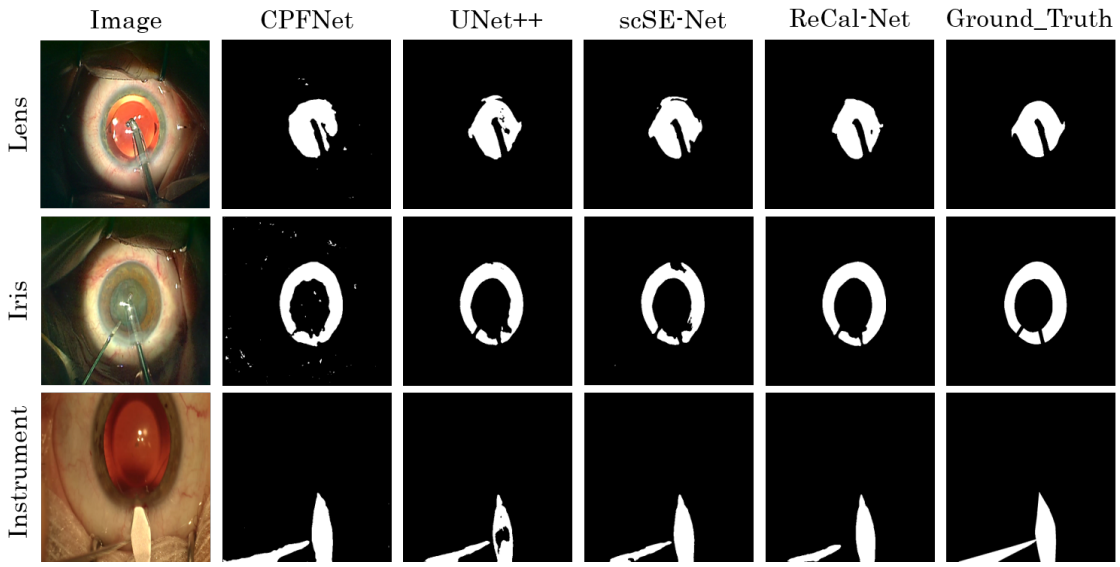


Figure 7.4: Qualitative comparisons among the top four segmentation approaches.

Table 7.4: Impact of adding ReCal modules on the segmentation accuracy based on IoU(%).

Learning Rate	Network	Lens	Iris	Instrument
0.002	Baseline	84.83 ±11.62	81.49 ± 6.82	70.04 ±23.94
	ReCal-Net	85.77 ±12.33	83.29 ± 5.82	71.89 ±19.93
0.005	Baseline	86.13 ±11.63	81.00 ± 8.06	67.16 ±24.67
	ReCal-Net	87.94 ±10.72	85.13 ± 3.98	70.43 ±21.17

To further investigate the impact of the ReCal modules on segmentation performance, we have visualized two intermediate filter response maps for iris segmentation in Figure 7.5. The E5 output corresponds to the filter response map of the last encoder block, and the D1 output corresponds to the filter response map of the first decoder block (see Figure 7.1). A comparison between the filter response maps of the baseline and ReCal-Net in the same locations indicated the positive impact of the ReCal modules on the network's semantic discrimination capability. Indeed, employing the correlations between the pixel-wise, region-wise, and channel-wise descriptors can reinforce the network's semantic interpretation.

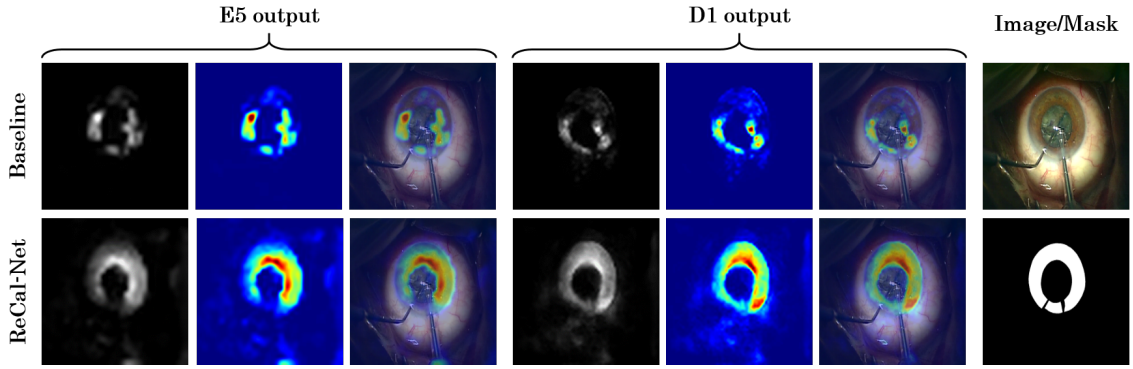


Figure 7.5: Visualizations of the intermediate outputs in the baseline approach and ReCal-Net based on class activation maps [237]. For each output, the figures from left to right represent the gray-scale activation maps, heatmaps, and heatmaps on images.

7.5 Conclusion

This chapter presents a novel convolutional module, termed as ReCal module, that can adaptively calibrate feature maps considering pixel-wise, region-wise, and channel-wise descriptors. The ReCal module can effectively correlate intra-region information and cross-channel-region information to deal with severe contextual variations in the same semantic labels and contextual similarities between different semantic labels. The proposed region-channel recalibration module is a very light-weight computational unit that can be applied to any feature map $\mathcal{X} \in \mathbb{R}^{C \times H \times W}$ and output a recalibrated feature map $\mathcal{Y} \in \mathbb{R}^{C \times H \times W}$. Moreover, we have proposed a novel network architecture based on the ReCal module for semantic segmentation in cataract surgery videos, termed as ReCal-Net. The experimental evaluations confirm the effectiveness of the proposed ReCal module and ReCal-Net in dealing with various segmentation challenges in cataract surgery. The proposed ReCal module and ReCal-Net can be adopted for various medical image segmentation and general semantic segmentation problems.

CHAPTER

8 Semantic Segmentation using DeepPyram

Chapter overview — Semantic segmentation in cataract surgery has a wide range of applications contributing to surgical outcome enhancement and clinical risk reduction. However, the varying issues in segmenting the different relevant instances make the designation of a unique network quite challenging. This chapter proposes a semantic segmentation network termed as DeepPyram that can achieve superior performance in segmenting relevant objects in cataract surgery videos with varying issues. This superiority mainly originates from three modules: (i) Pyramid View Fusion, which provides a varying-angle global view of the surrounding region centering at each pixel position in the input convolutional feature map; (ii) Deformable Pyramid Reception, which enables a wide deformable receptive field that can adapt to geometric transformations in the object of interest; and (iii) Pyramid Loss that adaptively supervises multi-scale semantic feature maps. These modules can effectively boost semantic segmentation performance, especially in the case of transparency, deformability, scalability, and blunt edges in objects. The proposed approach is evaluated using four datasets of cataract surgery for objects with different contextual features and compared with thirteen state-of-the-art segmentation networks. The experimental results confirm that DeepPyram outperforms the rival approaches without imposing additional trainable parameters. Our comprehensive ablation study further proves the effectiveness of the proposed modules.

This chapter is an adapted version of:

"Ghamsarian, N., Taschwer, and Schoeffmann, K. DeepPyram: Enabling Pyramid View and Deformable Pyramid Reception for Semantic Segmentation in Cataract Surgery Videos. Under Review."

8.1 Introduction

Semantic segmentation plays a prominent role in computerized surgical workflow analysis. Especially in cataract surgery, where workflow analysis can highly contribute to the reduction of intra-operative and post-operative complications[78], semantic segmentation is of great importance. Cataract refers to the eye's natural lens having become cloudy and causing vision deterioration. Cataract surgery is the procedure of restoring clear eye vision via cataract removal followed by artificial lens implantation. This surgery is the most common ophthalmic surgery and one of the most frequent surgical procedures worldwide [231]. Semantic segmentation in cataract surgery videos has several applications ranging from phase and action recognition [82, 280], irregularity detection (pupillary reaction, lens rotation, lens instability, and lens unfolding delay detection), objective skill assessment, relevance-based compression[79], and so forth [262, 195, 170, 93, 11]. Accordingly, there exist four different relevant objects in videos from cataract surgery, namely Intraocular Lens, Pupil, Cornea, and Instruments. The diversity of features of different relevant objects in cataract surgery imposes a challenge on optimal neural network architecture designation. More concretely, a semantic segmentation network is required that can simultaneously deal with (I) deformability and transparency in case of the artificial lens, (II) color and texture variation in case of the pupil, (III) blunt edges in case of the cornea, and (IV) harsh motion blur degradation, reflection distortion, and scale variation in case of the instruments (Figure 8.1).

Thischapter presents a U-Net-based CNN for semantic segmentation that can adaptively capture the semantic information in cataract surgery videos. The proposed network, termed as DeepPyram, mainly consists of three modules: (i) Pyramid View Fusion (PVF) module enabling a varying-angle surrounding view of feature map for each pixel position, (ii) Deformable Pyramid Reception (DPR) module being responsible for performing shape-wise feature extraction on the input convolutional

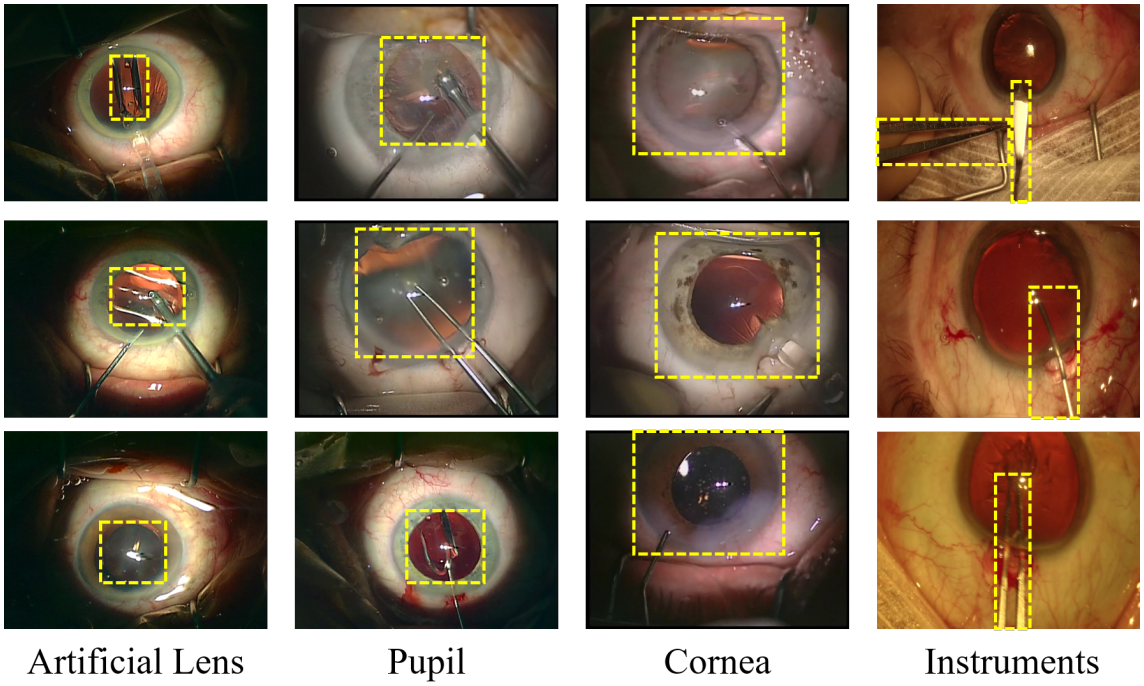


Figure 8.1: Semantic Segmentation difficulties for different relevant objects in cataract surgery videos.

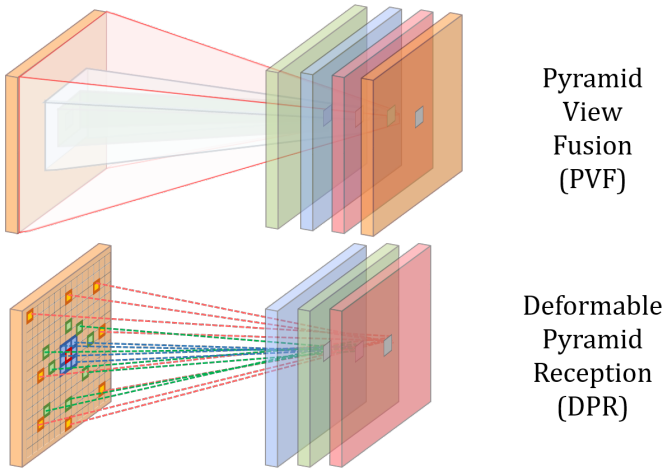


Figure 8.2: Two major operations in DeepPyram.

feature map (Figure 8.2), and (iii) Pyramid Loss ($P\mathcal{L}$) module that directly supervises the multi-scale semantic feature maps. We have provided a comprehensive study to compare the performance of DeepPyram with thirteen rival state-of-the-art approaches for relevant-instance segmentation in cataract surgery. The experimental results affirm the superiority of DeepPyram, especially in the case of scalable and

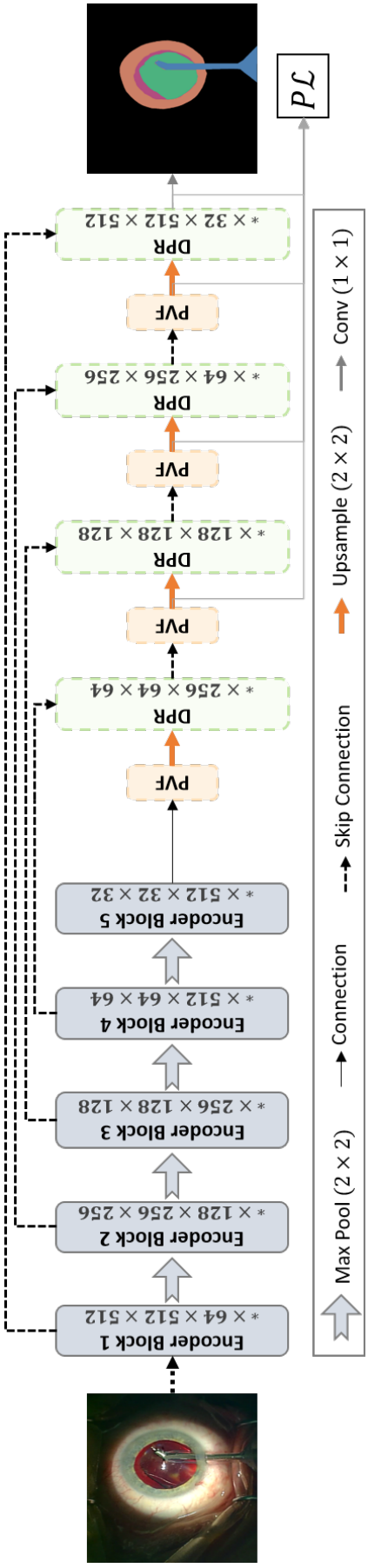


Figure 8.3: The overall architecture of the proposed DeepPyram network. It contains Pyramid View Fusion (PVF), Deformable Pyramid Reception (DPR), and Pyramid Loss ($\mathcal{PL}$) modules.

deformable transparent objects.

The rest of the paper is organized as follows. In Section 8.2, we position our approach in the literature by reviewing the state-of-the-art semantic segmentation approaches. We delineate the proposed network (DeepPyram) in Section 8.3. We describe the experimental settings in Section 8.4 and analyze the experimental results in Section 8.5. We also provide an ablation study on DeepPyram in Section 8.5 and conclude the paper in Section 8.9.

[67] for deep supervision

8.2 Related Work

This section briefly reviews U-Net-based approaches, state-of-the-art semantic segmentation approaches related to attention and fusion modules, and multi-branch supervision.

U-Net-based Networks. U-Net [203] was initially proposed for medical image segmentation and achieved succeeding performance being attributed to its skip connections. In the encoder side of this encoder-decoder network, low-level features are combined and transformed into semantic information with low resolution. The decoder network improves these low-resolution semantic features and converts them to semantic segmentation results with the same resolution as the input image. The role of skip connections is to transmit the fine-grained low-level feature maps from the encoder to the decoder layers. The coarse-grained semantic feature maps from the decoder and fine-grained low-level feature maps from the encoder are accumulated and undergo convolutional operations. This accumulation technique helps the decoder retrieve the object of interest's high-resolution features to provide delineated semantic segmentation results. Many U-Net-based architectures have been proposed over the past five years to improve the segmentation accuracy and address different flaws and restrictions in the previous architectures [17, 43, 177, 90, 108, 264, 175, 176, 69, 272].

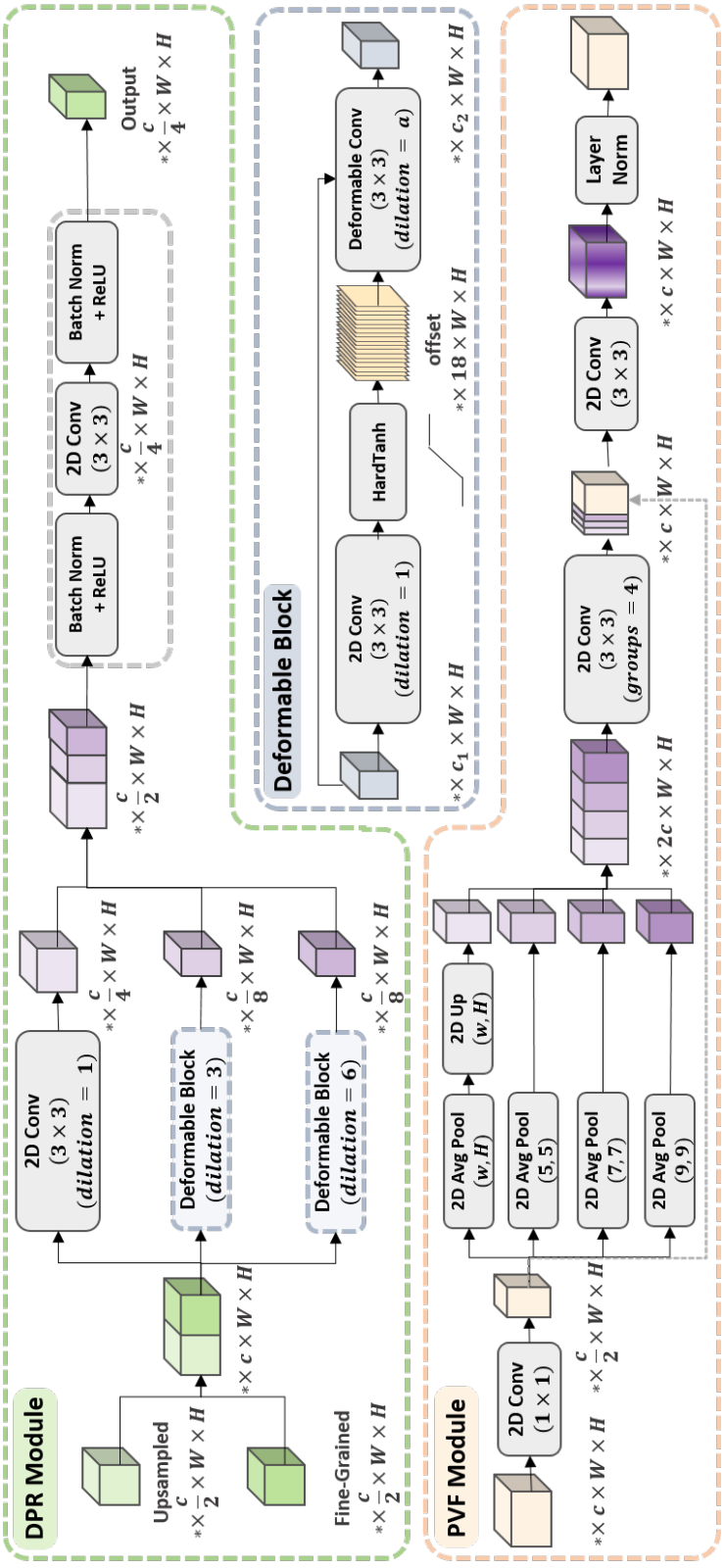


Figure 8.4: The detailed architecture of the Deformable Pyramid Reception (DPR) and Pyramid View Fusion (PVF) modules.

In SegNet [17], non-linear upsampling in a decoder's layer is performed using the stored max-pooling indices in its symmetrical encoder's layer. The sparse upsampled feature maps are then converted to dense feature maps using convolutional layers. In another work, dU-Net [264] replaces the classic convolutional operations in U-Net with deformable convolutions [54] to deal with shape variations. Full-Resolution Residual Networks (FRRN) adopt a two-stream architecture, namely the residual and the pooling stream [189]. The residual stream preserves the full resolution of the input image. The pooling stream employs consecutive full-resolution residual units (FRRU) to add semantic information to the residual stream gradually. Inspired by FRNN, Jue *et al.* [115] developed two types of multiple resolution residually connected networks (MRRN) for lung tumor segmentation. The two versions of MRNN, namely incremental-MRNN and dense-MRNN, aim to fuse varying-level semantic feature maps with different resolutions to deal with size variance in tumors. It is argued that the optimal depth of a U-Net architecture for different datasets is different. UNet++ [272] as an ensemble of varying-depth UNets is proposed to address this depth optimization problem. Hejie *et al.* [51] employe UNet++ as their baseline and add dense connections between the blocks with matching-resolution feature maps for pulmonary vessel segmentation.

Attention Modules. Attention mechanisms can be broadly described as the techniques to guide the network's computational resources (*i.e.*, the convolutional operations) towards the most determinative features in the input feature map. Such mechanisms have been especially proven to be gainful in the case of semantic segmentation. Inspired by Squeeze-and-Excitation block [103], the SegSE block [187] and scSE block [205] aim to utilize inter-channel dependencies and recalibrate the channels spatially by applying fully connected operations on the globally pooled feature maps. RAUNet [177] includes an attention module to merge the multi-level feature maps from the encoder and decoder using global average pooling. BARNet [176]

adopts a bilinear-attention module to extract the cross semantic dependencies between the different channels of a convolutional feature map. This module is specially designed to enhance the segmentation accuracy in the case of illumination and scale variation for surgical instruments. PAANET [175] uses a double attention module (DAM) to model semantic dependencies between channels and spatial positions in the convolutional feature map.

Fusion Modules. Fusion modules can be characterized as modules designed to improve semantic representation via combining several feature maps. The input feature maps could range from varying-level semantic features to the features coming from parallel operations. PSPNet [268] adopts a pyramid pooling module (PPM) containing parallel sub-region average pooling layers followed by upsampling to fuse the multi-scale sub-region representations. This module has shown significant improvement and is frequently used in semantic segmentation architectures [276]. Atrous spatial pyramid pooling (ASPP) [38, 40] was proposed to deal with objects' scale variance as an efficient alternative to Share-net [39]. Indeed, the ASPP module aggregates multi-scale features extracted using parallel varying-rate dilated convolutions and obviates the need to propagate and aggregate the features of multi-scale inputs. This module is employed in many segmentation approaches due to its effectiveness in capturing multi-scale contextual features [254, 271, 191].
Autofocus Convolutional Layer [193] uses a novel approach to fuse resulting feature maps of dilated convolutions adaptively. CPFNet [69] uses another fusion approach for scale-aware feature extraction. MultiResUNet [108] and Dilated MultiResUNet [250] factorize the large and computationally expensive receptive fields into a fusion of successive small receptive fields.

Multi-Branch Supervision. The idea of deep supervision was initially proposed by Chen-Yu *et al.* [143]. The authors proved that introducing a classifier (SVM or Softmax) on top of hidden layers can improve the learning process and minimize

classification failure. This idea is simultaneously adopted by GoogleNet [223] to facilitate gradient flow in deep neural network architectures. The auxiliary loss in PSPNet [268] follows the same strategy and guides one feature map in the encoder network to reinforce learning discriminative features in shallower layers. The architecture of DensNets [105] implicitly enables such deep supervision. Multi-branch supervision approaches are frequently used for improving semantic segmentation. Qi *et al.* [63] suggest directly supervising multi-resolution feature maps of the encoder network by adding deconvolutional layers followed by a Softmax activation layer on top of them. Zhu *et al.* [274] adopt five deep supervision modules for the multi-scale feature maps in the encoder network and three deep supervision modules in the decoder network. In each module, the input feature map is upsampled to its original version and undergoes a deconvolutional layer which extracts semantic segmentation results. The nested architecture of UNet++ [272] inherently provides such multi-depth semantic feature maps with original resolution.

8.3 Methodology

8.3.1 Overview

Figure 8.3 depicts the architecture of the proposed network. Overall, the network consists of a contracting path and an expanding path. The contracting path is responsible for converting low-level to semantic features. The expanding path accounts for performing super-resolution on the coarse-grained semantic feature maps and improving the segmentation accuracy by taking advantage of the symmetric[1] fine-grained feature maps. The encoder network in the baseline approach is VGG16, pretrained on ImageNet. The decoder network consists of three modules: (i) Pyramid View Fusion (PVF), which induces a large-scale view with progressive angles, (ii) Deformable Pyramid Reception (DPR), which enables a large, sparse, and learnable

[1]Symmetric feature maps are the feature maps with the same spatial resolution.

receptive field, which can sample from up to seven pixels far from each pixel position in the input convolutional feature map, and (iii) Pyramid Loss ($P\mathcal{L}$), which is responsible for the direct supervision of multi-scale semantic feature maps. In the following subsections, we detail each of the proposed modules.

8.3.2 Pyramid View Fusion (PVF)

By this module, we aim to stimulate a neural network deduction process analogous to the human visual system. Due to non-uniformly distributed photoreceptors on the retina, the perceived region with high resolution by the human visual system is up to $2^\circ - 5^\circ$ of visual angle centering around the gaze [112]. Correspondingly, we infer that the human eye recognizes the semantic information considering not only the internal object's content but also the relative information between the object and the surrounding area. The pyramid view fusion (PVF) module's role is to reinforce the feeling of such relative information at every distinct pixel position. One way to exploit such relative features is to apply convolutional operations with large receptive fields. However, increasing the receptive field's size is not recommended due to imposing huge additional trainable parameters, and consequently (i) escalating the risk of overfitting and (ii) increasing the requirement to more annotations. Alternatively, we use average pooling for fusing the multi-angle local information in our novel attention mechanism. At first, as shown at the bottom of Figure 8.4, a bottleneck is formed by employing a convolutional layer with a kernel size of one to curb the computational complexity. After this dimensionality reduction stage, the convolutional feature map is fed into four parallel branches. The first branch is a global average pooling layer followed by upsampling. The other three branches include average pooling layers with progressive filter sizes and the stride of one. Using a one-pixel stride is specifically essential for obtaining pixel-wise centralized pyramid view in contrast with region-wise pyramid attention in PSPNet [268]. The output feature maps are then concatenated and fed into a convolutional layer with four groups. This

layer is responsible for extracting inter-channel dependencies during dimensionality reduction. A regular convolutional layer is then applied to extract joint intra-channel and inter-channel dependencies before being fed into a layer-normalization function.

Discussion. There are three significant differences between the PVF module in DeepPyram and pyramid pooling module in PSPNet [268]: (i) All average pooling layers in the PVF module use a stride of one pixel, whereas the average pooling layers in PSPNet adopt different strides of 3, 4, and 6 pixels. The pyramid pooling module separates the input feature map into varying-size sub-regions and plays the role of object detection as region proposal networks (RPNs [201]). However, the stride of one pixel in the PVF module is used to capture the subtle changes in pyramid information that is especially important for segmenting the narrow regions of objects such as instruments and the artificial lens hooks[2]. These subtle differences can be diluted after fusing the pyramid information in the pyramid pooling module in PSPNet. (ii) In contrast with PSPNet, the PVF module applies three functions to the concatenated features to capture high-level contextual features: (1) a group convolution function to fuse the features obtained from the average pooling filters independently (as in PSPNet), (2) a convolutional operation being responsible for combining the multi-angle local features to reinforce the semantic features corresponding to each target label, and (3) a layer normalization function that accelerates training and avoids overfitting to the color and contrast information in case of few annotations.

8.3.3 Deformable Pyramid Reception (DPR)

Figure 8.4 (top) demonstrates the architecture of the deformable pyramid reception (DPR) module. At first, the fine-grained feature map from the encoder and coarse-grained semantic feature map from the previous layer are concatenated. Afterward,

[2]Precise segmentation of the artificial lens hooks is crucial for lens rotation and irregularity detection.

these features are fed into three parallel branches: a regular 3×3 convolution and two deformable blocks with different dilation rates. These layers together cover a learnable sparse receptive field of size 15×15[3]. The output feature maps are then concatenated before undergoing a sequence of regular layers for higher-order feature extraction and dimensionality reduction.

Deformable Block. Dilated convolutions can implicitly enlarge the receptive field without imposing additional trainable parameters. Dilated convolutional layers can recognize each pixel position's semantic label based on its cross-dependencies with varying-distance surrounding pixels. These layers are exploited in many architectures for semantic segmentation [254, 38, 271]. Due to the inflexible rectangle shape of the receptive field in regular convolutional layers, however, the feature extraction procedure cannot be adapted to the target semantic label. Dilated deformable convolutional layers can effectively support the object's geometric transformations in terms of scale and shape. As shown in the *Deformable Block* in Figure 8.4, a regular convolutional layer is applied to the input feature map to compute the offset field for deformable convolution. The offset field provides two values per element in the convolutional filter (horizontal and vertical offsets). Accordingly, the number of offset field's output channels for a kernel of size 3×3 is equal to 18. Inspired by dU-Net [264], the convolutional layer for the offset field is followed by an activation function. We use the hard tangent hyperbolic function (HardTanh), which is computationally cheap, to clip the offset values in the range of $[-1, 1]$. The deformable convolutional layer uses the learned offset values along with the convolutional feature map with a predetermined dilation rate to extract object-adaptive features.

The output feature map (y) for each pixel position (p_0) and the receptive field ($\mathcal{RF}$) for a regular 2D convolution with a 3×3 filter and dilation rate of one can be defined

[3]The structured 3×3 filter covers up to one pixel far from the central pixel). The deformable filter with $dilation = 3$ covers an area from two to four pixels far away from each central pixel. The second deformable convolution with $dilation = 6$ covers an area from five to seven pixels far away from each central pixel. Therefore, these layers together form a sparse filter of size 15×15 pixels. This sparse kernel can be better seen in Figure 8.2.

as:

$$y(p_o) = \sum_{p_i \in \mathcal{RF}_1} x(p_0 + p_i).w(p_i) \tag{8.1}$$

$$\mathcal{RF}_1 = \{(-1,-1),(-1,0),...,(1,0),(1,1)\} \tag{8.2}$$

Where x denotes the input convolutional feature map and w refers to the weights of the convolutional kernel. In a dilated 2D convolution with a dilation rate of α, the receptive field can be defined as $\mathcal{RF}_\alpha = \alpha \times \mathcal{RF}_1$. Although the sampling locations in a dilated receptive field have a greater distance with the central pixel, they follow a firm structure. In a deformable dilated convolution with a dilation rate of α, the sampling locations of the receptive field are dependent to the local contextual features. In the proposed deformable block, the sampling location for the ith element of the receptive field and the input pixel p_0 can be formulated as:

$$\mathcal{RF}_{def,\alpha}[i,p_0] = \mathcal{RF}_\alpha[i] + f(\sum_{p_j \in \mathcal{RF}_1} x(p_0 + p_j).\hat{w}(p_j)) \tag{8.3}$$

In (8.3), f denotes the activation function, which is the tangent hyperbolic function in our case, and $\hat{w}$ refers to the weights of the offset filter. This learnable receptive field can be adapted to every distinct pixel in the convolutional feature map and allows the convolutional layer to extract more informative semantic features compared to the regular convolution.

Discussion. deformable convolutions are usually used for video object detection, tracking, and segmentation since a deformable layer needs offsets. These offsets are usually provided by optical flow computation, subtracting consecutive frames, or substracting the corresponding feature maps of consecutive frames (as in MaskProp [21] and MF-TAPNet [118]). Since we use videos recorded in usual and non-laboratory conditions, we have many problems in the video dataset such as defocus blur and

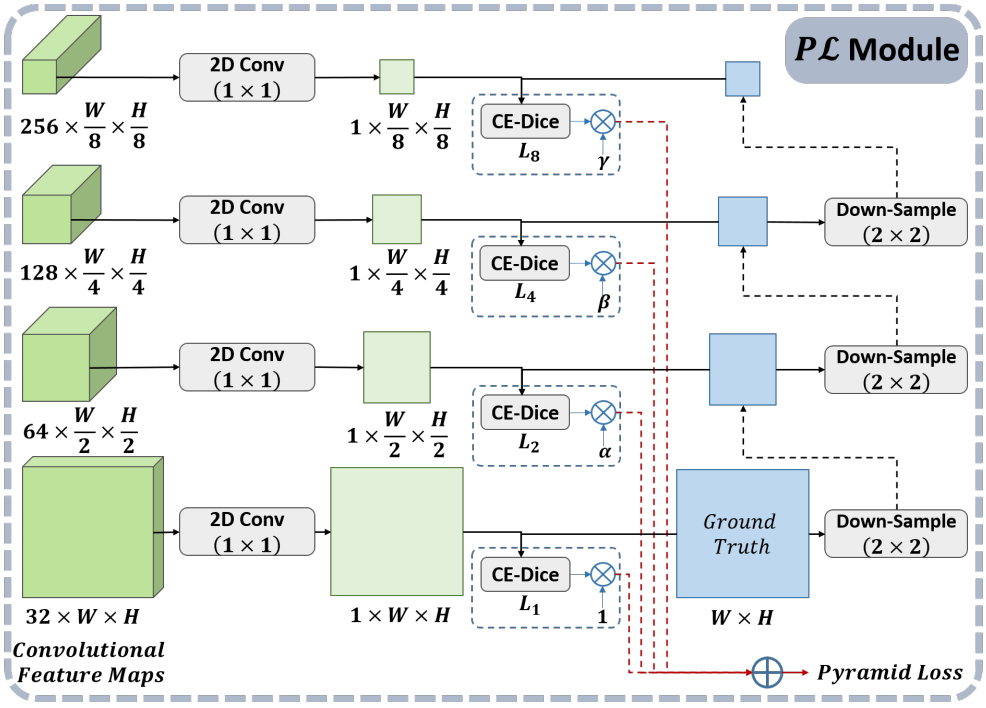


Figure 8.5: Demonstration of the *Pyramid Loss* module.

harsh motion blur (due to fast movements of eye and motions of the instruments [83]). Using temporal information for semantic segmentation (as in MaskProp [21]) may lead to error accumulation and less precise results. Moreover, video object segmentation requires much more annotations which is an additional burden on the expert surgeons' time. For the DPR module, we propose (1) a deformable block to apply deformable convolutions based on learned offsets from static information, (2) a combination of deformable and static filters which can not only capture the edge-sensitive information in the case of sharp edges (as in pupil and lens), but also can cover a large area (up to a size of 15×15 pixels) to deal with color and scale variations, blunt edges in case of cornea, and reflections in case of instruments.

8.3.4 Pyramid Loss (P$\mathcal{L}$)

The role of this module is to directly supervise the multi-scale semantic feature maps on the decoder's side. As shown in Fig 8.5, in order to enable direct supervision, a fully connected layer is formed using a pixel-wise convolution operation. The output feature

map presents the semantic segmentation results with the same resolution as the input feature map. To compute the loss for varying-scale outputs, we downscale the ground-truth masks using nearest-neighbor downsampling for multi-class segmentation and max-pooling for binary segmentation. The overall loss is defined as:

$$P\mathcal{L} = \mathcal{L}_1 + \alpha\mathcal{L}_2 + \beta\mathcal{L}_4 + \gamma\mathcal{L}_8 \tag{8.4}$$

Where α, β, and γ are predetermined weights in the range of $[0, 1]$ (In the experiments, we have set $\alpha = 0.75$, $\beta = 0.5$, and $\gamma = 0.25$). Besides, $\mathcal{L}_i$ denotes the loss of output mask segmentation result with the resolution of $(1/i)$ compared to the input resolution.

Discussion. The auxiliary losses for deep semantic segmentation that are proposed to date have two different purposes: (1) Many auxiliary losses aim to guide one or more feature maps in the encoder subnetwork to prevent gradient vanishing due to using a deep architecture (for instance ResNet50) as the backbone. (2) Some auxiliary losses attempt to improve the segmentation accuracy by directly guiding different layers of the encoder or decoder network. In the second case, the loss is computed by performing super-resolution through interpolating the output feature map (to obtain a feature map with the same resolution as the ground truth) and comparing it to the ground truth. In contrast, the $P\mathcal{L}$ module directly compares each feature map to the downsampled version of the ground truth. This strategy is more time-efficient and computationally less expensive compared to the previously proposed auxiliary losses. Moreover, in contrast with state-of-the-art approaches, each loss branch in the $P\mathcal{L}$ module only consists of $Conv2D + Softmax$ to keep the number of trainable parameters as few as possible.

Table 8.1: Specifications of the proposed and rival approaches. In the “loss" column, “CE" and “CE-Dice" stand for *Cross Entropy* and *Cross Entropy Log Dice*. In “Upsampling" column, “Trans Conv" stands for *Transposed Convolution*.

Model	Backbone	Params	Loss	Upsampling	Target	Year	Reference
UNet++	VGG16	24.24 M	CE-Dice	Bilinear	Medical Images	2020	[272]
UNet++/DS	VGG16	24.24 M	CE-Dice	Bilinear	Medical Images	2020	[272]
CPFNet	ResNet34	34.66 M	CE-Dice	Bilinear	Medical Images	2020	[69]
BARNet	ResNet34	24.90 M	CE-Dice	Bilinear	Surgical Instruments	2020	[176]
PAANet	ResNet34	22.43M	CE-Dice	Trans Conv & Bilinear	Surgical Instruments	2020	[175]
dU-Net	✗	31.98 M	CE-Dice	Trans Conv	Blood Cells	2020	[264]
MultiResUNet	✗	9.34 M	CE	Trans Conv	Medical Images	2020	[108]
CE-Net	ResNet34	29.9 M	CE	Trans Conv	Medical Images	2019	[90]
RAUNet	ResNet34	22.14 M	CE-Dice	Trans Conv	Cataract Surgical Instruments	2019	[177]
FED-Net	ResNet50	59.52 M	CE-Dice	Trans Conv & PixelShuffle	Liver Lesion	2019	[43]
PSPNet	ResNet50	22.26 M	CE	Bilinear	Scene	2017	[268]
SegNet	VGG16	14.71 M	CE	Max Unpooling	Scene	2017	[17]
U-Net	✗	17.26 M	CE	Bilinear	Medical Images	2015	[203]
DeepPyram	VGG16	23.62 M	CE-Dice	Bilinear	Cataract Surgery	Proposed Approach	

8.4 Experimental Setup

Datasets. We have used four datasets with varying instance features to provide extensive evaluations for the proposed and rival approaches. The two public datasets include the "Cornea" [79] and "Instruments" [89] mask annotations. Additionally, we have prepared a customized dataset for "Intraocular Lens" and "Pupil" pixel-wise segmentation[4]. The number of training and testing images for the aforementioned objects are 178:84, 3190:459, 141:48, and 141:48, respectively. All training and testing images are sampled from distinctive videos to meet real-world conditions. Our annotations are performed using the "supervise.ly" platform[5] based on the guidelines from cataract surgeons.

Rival Approaches. Table 8.1 details the specifications of the proposed approach and rival approaches employed in our experiments. To provide a fair comparison, we adopt our improved version of PSPNet, featuring a decoder designed similarly to U-Net (with four sequences of double-convolution blocks). Besides, our version of du-Net has the same number of filter-response-maps as for U-Net.

Data Augmentation Methods. Data augmentation is a vital step during training, which prevents network overfitting and boosts the network performance in the case of unseen data. Accordingly, the training images for all evaluations undergo various augmentation approaches before being fed into the network. We chose the transformations considering the inherent and statistical features of datasets. For instance, we use motion blur transformation to encourage the network to deal with harsh motion blur regularly occurring in cataract surgery videos [83]. Table 8.2 details the adopted augmentation pipeline.

[4]The dataset will be publicly released with the acceptance of the paper.
[5]https://supervise.ly/

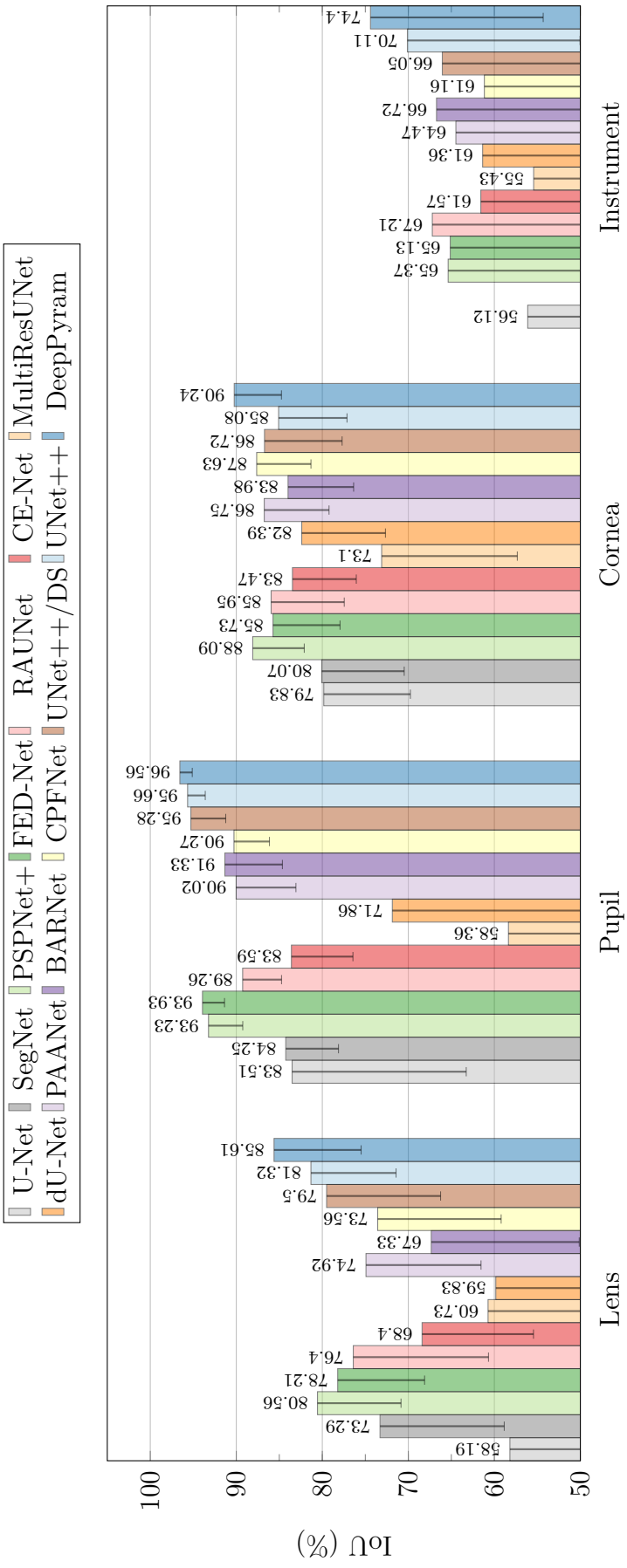


Figure 8.6: Quantitative comparisons among DeepPyram and rival approaches based on average and standard deviation of IoU.

Table 8.2: Augmentation Pipeline.

Augmentation Method	Property	Value
Brightness & Contrast	Factor Range	(-0.2,0.2)
Shift & Scale	Percentage	10%
Rotate	Degree Range	[-10,10]
Motion Blur	Kernel-Size Range	(3,7)

Neural Network Settings. As listed in Table 8.1, U-Net, MultiResUNet, and dU-Net do not adopt a pretrained backbone. For the other approaches, the weights of the backbone are initialized with ImageNet [55] training weights. The input size of all models is set to $3 \times 512 \times 512$. In the evaluation and testing stages of UNet++/DS and DeepPyram, we disregard the additional output branches (branches related to auxiliary loss functions) and only consider the master output branch.

Training Settings. Due to the different depth and connections of the proposed and rival approaches, all networks are trained with three different initial learning rates ($lr \in \{0.0005, 0.0002, 0.001\}$), and the results with the highest IoU for each network are listed. The learning rate is scheduled to decrease every two epochs with the factor of 0.8. In all different evaluations, the networks are trained end-to-end and for 30 epochs. We use a threshold of 0.1 for gradient clipping during training[6].

Loss Function. The *cross entropy log dice* loss, which is used during training, is a weighted sum of binary cross-entropy (BCE) and the logarithm of soft Dice coefficient as follows:

$$\begin{aligned}\mathcal{L} =&(\lambda) \times BCE(\mathcal{X}_{true}(i,j), \mathcal{X}_{pred}(i,j)) \\ &- (1-\lambda) \times (\log \frac{2 \sum \mathcal{X}_{true} \odot \mathcal{X}_{pred} + \sigma}{\sum \mathcal{X}_{true} + \sum \mathcal{X}_{pred} + \sigma})\end{aligned} \tag{8.5}$$

Where $\mathcal{X}_{true}$ denote the ground truth binary mask, and $\mathcal{X}_{pred}$ denote the predicted mask ($0 \leq \mathcal{X}_{pred}(i,j) \leq 1$). The parameter $\lambda \in [0,1]$ is set to 0.8 in our experiments,

[6]Gradient clipping is used to clip the error derivatives during back-propagation to prevent gradient explosion.

Table 8.3: Impact of different modules on the segmentation results (IoU% and Dice%) of DeepPyram.

Modules				Lens		Pupil		Cornea		Instrument		Overall	
PVF	DPR	PC	Params	IoU(%)	Dice(%)	IoU(%)	Dice(%)	IoU(%)	Dice(%)	IoU(%)	Dice(%)	IoU(%)	Dice(%)
✗	✗	✗	22.55 M	82.98	90.44	95.13	97.48	86.02	92.28	69.82	79.05	83.49	89.81
✓	✗	✗	22.99 M	83.73	90.79	96.04	97.95	88.43	93.77	72.58	81.84	85.19	91.09
✗	✓	✗	23.17 M	81.85	89.58	95.32	97.59	86.43	92.55	71.57	80.60	83.79	90.08
✓	✓	✗	23.62 M	83.85	90.89	95.70	97.79	89.36	94.29	72.76	82.00	85.42	91.24
✓	✓	✓	23.62 M	**85.84**	**91.98**	**96.56**	**98.24**	**90.24**	**94.77**	**74.40**	**83.30**	**86.76**	**92.07**

and $\odot$ refers to Hadamard product (element-wise multiplication). Besides, the parameter s is the Laplacian smoothing factor which is added to (i) prevent division by zero, and (ii) avoid overfitting (in experiments, $\sigma = 1$).

Evaluation Metrics. The Jaccard metric (Intersection-over-Union – IoU) and the Dice Coefficient (F1-score) are regarded as the major semantic segmentation indicators. Accordingly, we evaluate the proposed and rival approaches using average IoU and dice. In order to enable a broader analysis of the networks' performance, the standard deviation of IoU, and minimum and maximum of dice coefficient over all of the testing images are additionally compared.

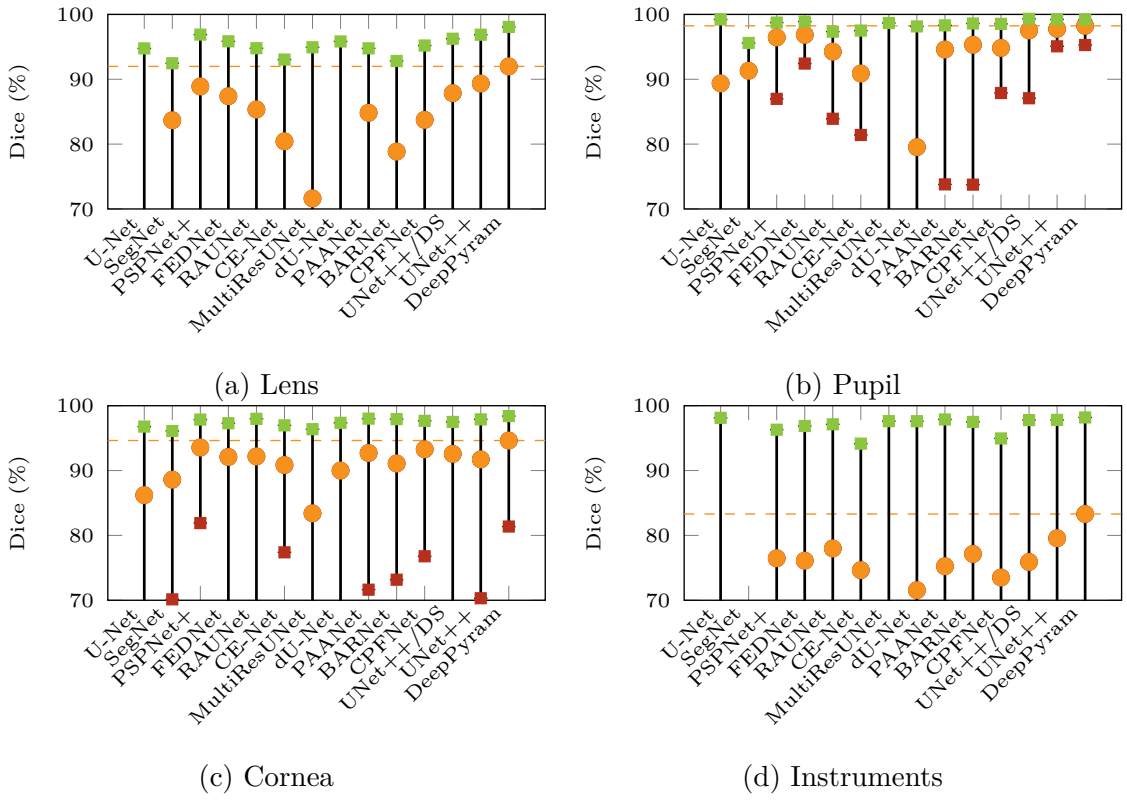
Ablation Study Settings. To evaluate the effectiveness of different modules, we have implemented another segmentation network, excluding all the proposed modules. This network has the same backbone as our baseline (VGG16). However, the PVF module is completely removed from the network. Besides, the DPR module is replaced with a sequence of two convolutional layers, each of which is followed by a batch normalization layer and a ReLU layer. The connections between the encoder and decoder remain the same as DeepPyram. Besides, DeepPyram++ in ablation studies is the nested version of DeepPyram implemented based on UNet++.

8.5 Experimental Results

8.5.1 Relevant Object Segmentation

Figure 8.6 compares the resulting IoU of DeepPyram and thirteen rival approaches[7]. Overall, the rival approaches have shown a different level of performance for each of the four relevant objects. Based on the mean IoU, the best four segmentation approaches for each of the four relevant objects are listed in descending order below:

[7]SegNet did not converge during training for instrument segmentation with different initial learning rates.



(a) Lens (b) Pupil

(c) Cornea (d) Instruments

Figure 8.7: Quantitative comparison of segmentation results for the proposed (DeepPyram) and rival architectures (some minimum and average values are not visible due to y-axis clipping).

- *Lens:* DeepPyram, Unet++, PSPNet+, UNet++/DS
- *Pupil:* DeepPyram, Unet++, UNet++/DS, FEDNet
- *Cornea:* DeepPyram, PSPNet+, CPFNet, UNet++/DS
- *Instrument:* DeepPyram, Unet++, RAUNet, BARNet

Accordingly, DeepPyram, Unet++, and PSPNet+ contribute to the top three segmentation results for the relevant objects in cataract surgery videos. However, DeepPyram shows considerable improvement in segmentation accuracy compared to the second-best approach in each class. Specifically, DeepPyram has achieved more than 4% improvement in lens segmentation (85.61% vs. 81.32%) and more than 4% improvement in instrument segmentation (74.40% vs. 70.11%) compared to UNet++ as the second-best approach. Moreover, DeepPyram appears to be the

most reliable approach considering the smallest standard deviation compared to the rival approaches. This significant improvement is attributed to the PVF, DPR, and $P\mathcal{L}$ modules.

As shown in Fig 8.7, DeepPyram has achieved the highest dice coefficient compared to the rival approaches for the lens, pupil, cornea, and instrument segmentation. Moreover, DeepPyram is the most reliable segmentation approach based on achieving the highest minimum dice percentage.

Figure 8.8 further affirms the effectiveness of DeepPyram in enhancing the segmentation results. Taking advantage of the pyramid view provided by the PVF module, DeepPyram can handle reflection and brightness variation in instruments, blunt edges in the cornea, color and texture variation in the pupil, as well as transparency in the lens. Furthermore, powering by deformable pyramid reception, DeepPyram can tackle scale variations in instruments and blunt edges in the cornea. In particular, we can perceive from Figure 8.8 that DeepPyram shows much less distortion in the region of edges (especially in the case of the cornea. Furthermore, based on these qualitative experiments, DeepPyram shows much better precision and recall in the narrow regions for segmenting the instruments and other relevant objects in the case of occlusion by the instruments.

8.5.2 Ablation Study

Table 8.3 validates the effectiveness of the proposed modules in segmentation enhancement. The PVF module can notably enhance the performance for cornea and instrument segmentation (2.41% and 2.76% improvement in IoU, respectively). This improvement is due to the ability of the PVF module to provide a global view of varying-size sub-regions centering around each distinctive spatial position. Such a global view can reinforce semantic representation in the regions corresponding to blunt edges and reflections. Due to scale variance in instruments, the DPR module can effectively boost the segmentation performance for instruments. The addition of

$P\mathcal{L}$ module results in the improvement of IoU for all relevant segments, especially lens segmentation (around 2% improvement) and Instrument (1.64% improvement). The combination of PVF, DPR, and $P\mathcal{L}$ modules can contribute to 4.58% improvement in instrument segmentation and 4.22% improvement in cornea segmentation (based on IoU%). These modules have improved the IoU for the lens and pupil by 2.85% and 1.43%, respectively.

Overall[8], the addition of different modules of DeepPyram has led to considerable improvement of segmentation performance (3.27% improvement in IoU) on average compared to the baseline approach. The PVF module can provide varying-angle global information centering around each pixel in the convolutional feature map to support relative information access. The DPR module enables large-field content-adaptive reception while minimizing the additive trainable parameters.

It should be noted that even our baseline approach has much better performance compared to some rival approaches. We argue that the fusion modules adopted in some rival approaches may lead to the dilution of discriminative semantic information.

elFig-TMI:IoU

8.6 Comparisons with Alternative Modules

Herein, we compare the effectiveness of the proposed modules with our enhanced version of the ASPP module [38] and PPM [268]. More concretely, we replace the PVF module with PPM (referring to as $Alternative^{PPM}$) and DPR module with our improved version of the ASPP module (referring to as $Alternative^{ASPP+}$). Figure 8.9 demonstrates the architecture of DeepPyram versus the alternative networks. The difference between the modules ASPP+ and ASPP lies in the filter size of the convolutional layers. In the ASPP module, there are four parallel convolutional layers: a pixel-wise convolutional layer and three dilated convolutions with different dilation rates. In ASPP+, the pixel-wise convolution is replaced with a 3×3

[8]The "Overall" column in Table 8.3 is the mean of the other four average values.

Table 8.4: Impact of alternative modules on the segmentation results (IoU% and Dice%) of DeepPyram.

		Lens		Pupil		Cornea		Instrument		Overall	
Network	Params	IoU(%)	Dice(%)	IoU(%)	Dice(%)	IoU(%)	Dice(%)	IoU(%)	Dice(%)	IoU(%)	Dice(%)
$Alternative^{ASPP+}$	22.99 M	85.02	91.51	96.41	98.17	88.83	94.00	**74.81**	**83.66**	86.26	91.83
$Alternative^{PPM}$	23.44 M	83.40	90.66	95.70	97.79	87.44	92.88	72.51	81.03	84.76	90.59
DeepPyram	23.62 M	**85.84**	**91.98**	**96.56**	**98.24**	**90.24**	**94.77**	74.40	83.30	**86.76**	**92.07**

Table 8.5: Impact of different backbones and combinations on the segmentation results (IoU% and Dice%) of DeepPyram.

			Lens		Pupil		Cornea		Instrument		Overall	
		Params	IoU(%)	Dice(%)	IoU(%)	Dice(%)	IoU(%)	Dice(%)	IoU(%)	Dice(%)	IoU(%)	Dice(%)
Backbone	ResNet50	85.52 M	81.71	89.54	95.09	97.46	89.32	94.23	71.79	80.73	84.48	90.49
	ResNet34	25.77 M	82.77	90.22	95.06	97.45	88.68	93.84	72.58	80.99	84.77	90.62
	VGG19	28.93 M	85.33	91.66	96.36	98.14	88.77	93.93	**74.70**	**83.49**	86.29	91.80
	VGG16	23.62 M	**85.84**	**91.98**	**96.56**	**98.24**	**90.24**	**94.77**	74.40	83.30	**86.76**	**92.07**
DeepPyram++		28.48 M	84.83	91.54	96.30	98.11	89.48	94.34	74.64	83.34	86.31	91.83

Table 8.6: Impact of different super-resolution functions on the segmentation results (IoU% and Dice%) of DeepPyram.

		Lens		Pupil		Cornea		Instrument		Overall	
Upsampling	Params	IoU(%)	Dice(%)	IoU(%)	Dice(%)	IoU(%)	Dice(%)	IoU(%)	Dice(%)	IoU(%)	Dice(%)
Trans Conv	25.01 M	84.37	91.22	96.01	97.96	88.80	93.97	**75.16**	**83.71**	86.08	91.71
PixelShuffle	36.15 M	84.31	91.00	96.49	98.20	89.17	94.15	74.55	83.41	86.13	91.69
Bilinear	23.62 M	**85.84**	**91.98**	**96.56**	**98.24**	**90.24**	**94.77**	74.40	83.30	**86.76**	**92.07**

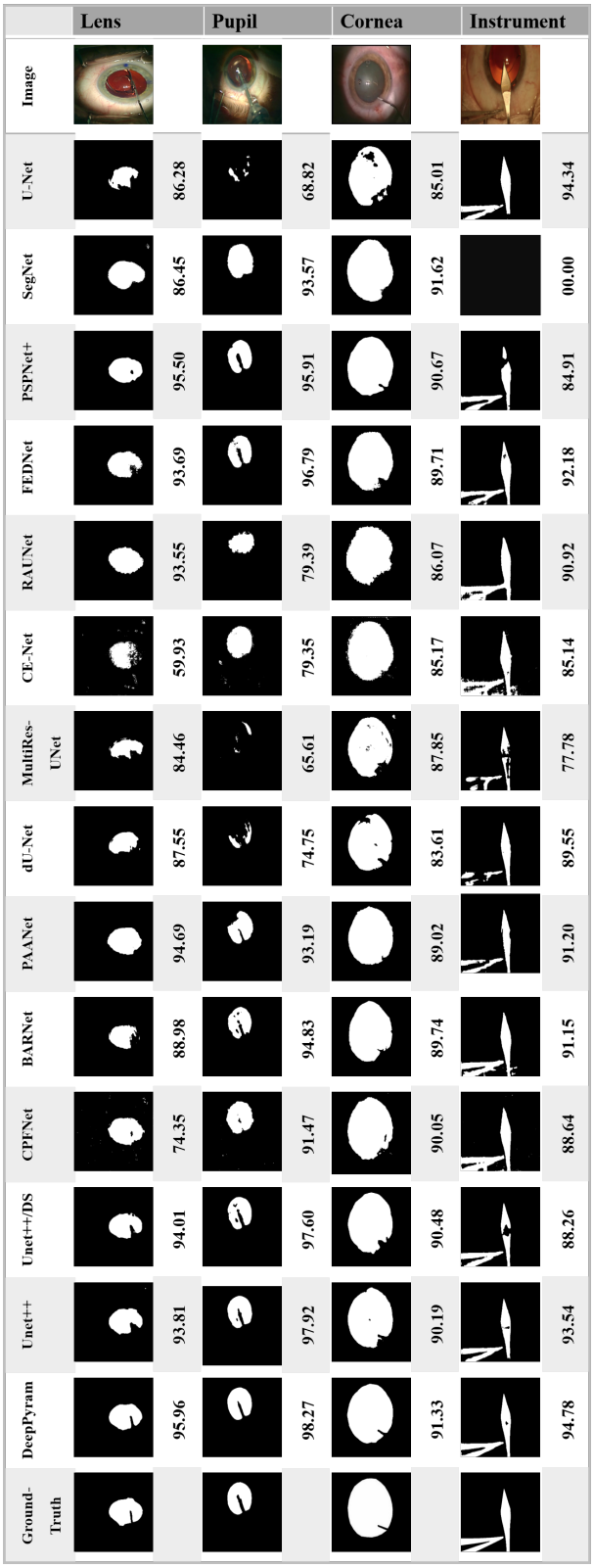


Figure 8.8: Qualitative comparisons among DeepPyram and the rival approaches for the relevant objects in cataract surgery videos (the numbers denote the Dice(%) coefficient for each detection).

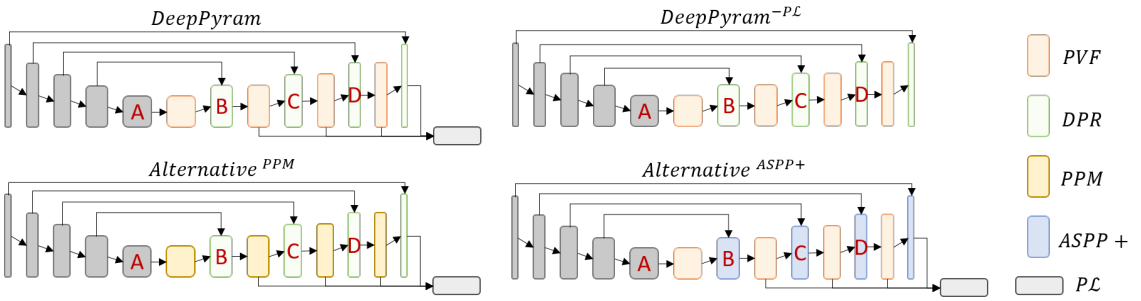


Figure 8.9: The overall architecture of DeepPyram compared to its three alternatives. The locations A, B, C, and D in each architecture correspond to the four modules for which we visualize the feature representations in Figure 8.10.

convolutional layer that effectively enhances the segmentation performance. Instead of three dilated convolutions in ASPP+, we use two dilated convolutions with the same dilation rates as in the DPR module. Moreover, the parallel feature-maps in the ASPP module are fused using a pixel-wise convolution, whereas the ASPP+ module adopts a kernel-size of 3×3 to boost the segmentation performance. Besides, we have removed the $P\mathcal{L}$ module and referred to it as $DeepPyram^{-P\mathcal{L}}$, to qualitatively compare its performance with DeepPyram. As illustrated in Figure 8.9, location "A" corresponds to the output of the last encoder's layer in the bottleneck. locations "B-D" are the outputs of three modules in the same locations of the decoder networks in DeepPyram and the three alternative networks. Figure 8.10 compares the class activation maps corresponding to these four locations in DeepPyram and alternative approaches for cornea segmentation in a representative image [9]. A comparison between the activation maps of DeepPyram and $DeepPyram^{-P\mathcal{L}}$ indicates how negatively removing the $P\mathcal{L}$ module affects the discrimination ability in different semantic levels. It is evident that the activation map of block "C" in DeepPyram is even more concrete compared to the activation map of block "D" (which is in a higher semantic level) in $DeepPyram^{-P\mathcal{L}}$. We can infer that the $P\mathcal{L}$ module can effectively reinforce the semantic representations in different semantic levels of the network. The effect of pixel-wise global view (PVF module) versus region-wise global

[9] These activation maps are obtained using Score-CAM [237] visualization approach.

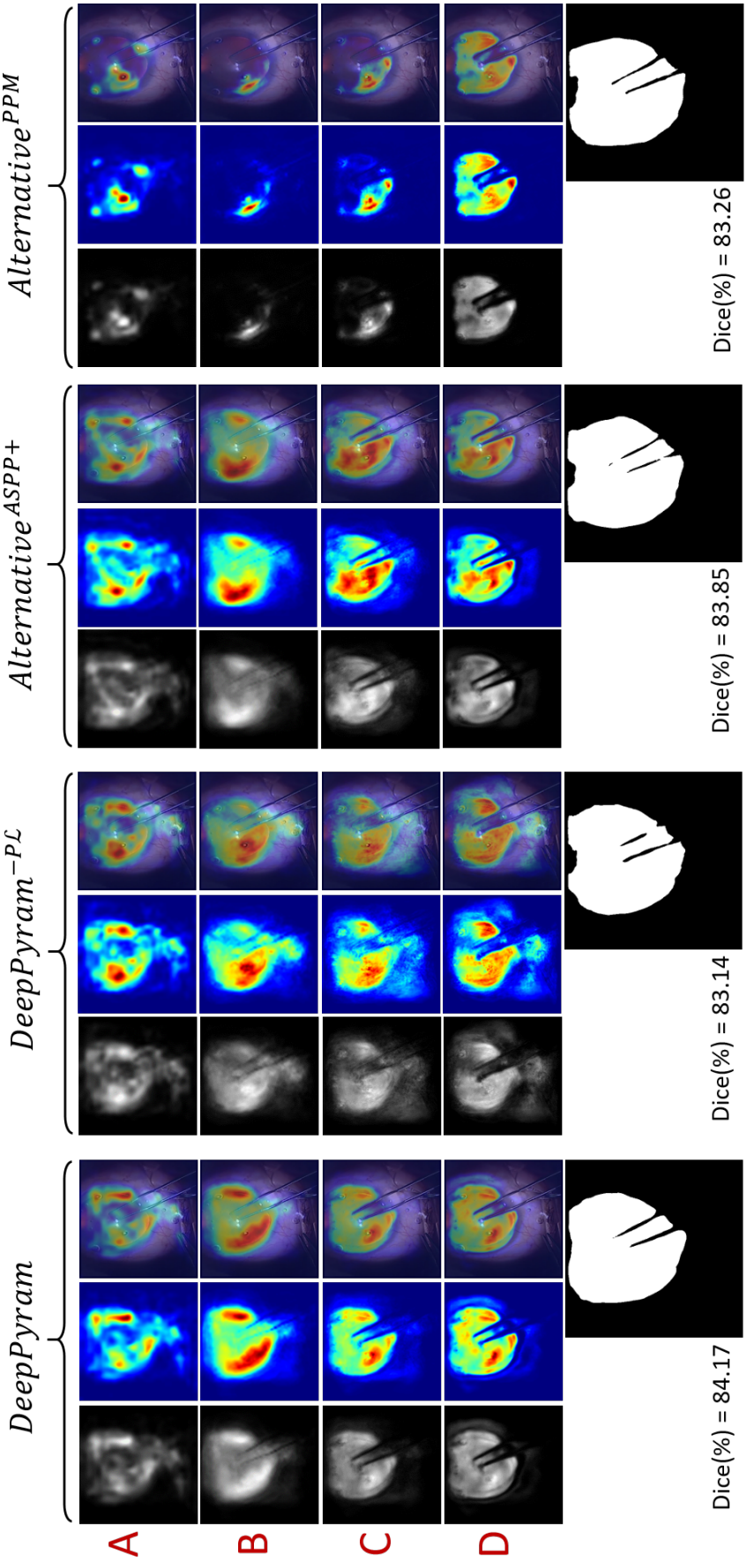


Figure 8.10: Visualization of the effect of the proposed and alternative modules based on class activation maps [237] using the network architectures demonstrated in Figure 8.9. For each approach, the figures from left to right represent the gray-scale activation maps, heatmaps, and heatmaps on images.

view (PPM) can be inferred by comparing the activation maps of DeepPyram and $Alternative^{PPM}$. The activation maps of $Alternative^{PPM}$ are impaired and distorted in different regions, especially in the lower semantic layers. The activation maps of $Alternative^{ASPP+}$ compared to DeepPyram confirm that replacing the deformable convolutions with regular convolutions can negatively affect semantic representation in narrow regions and the object borders.

8.7 Effect of Different Backbones and Nested Architecture

We have evaluated the achievable segmentation accuracy using different backbone networks to decide which backbone performs the best, considering the trade-off between the number of trainable parameters and the Dice percentage. As listed in Table 8.5, the experimental results show that the higher-depth networks such as ResNet50 cannot improve the segmentation accuracy. In contrast, VGG16 with the fewest number of trainable parameters has achieved the best segmentation performance. Moreover, VGG19, having around 5.3M parameters more than VGG16, performs just slightly better than the baseline backbone in instrument segmentation. In Table 8.5, DeepPyram++ is the nested version of DeepPyram [10] (with the same connections as in UNet++). This nested architecture shows around a one percent drop in IoU percentage on average.

[10] UNet++ uses the encoder-decoder architecture of U-Net as baseline, and adds additional convolutional layers between different encoder's and decoder's layers to form a nested multi-depth architecture. In DeepPyram++, we replace the the encoder-decoder baseline of UNet++ (which is U-Net) with our proposed DeepPyram network to see if these additional layers and connections can improve the segmentation performance of DeepPyram.

8.8 Effect of Different Super-resolution Functions

Table 8.6 compares the effect of three different super-resolution functions on the segmentation performance, including transposed convolution, Pixel-Shuffle [212], and bilinear upsampling. Overall, the network with bilinear upsampling function with the fewest parameters has achieved the best performance among the networks with different upsampling functions. Besides, the results reveal that the bilinear upsampling function has the best performance in segmenting all relevant objects except for the instruments.

8.9 Conclusion

In recent years, considerable attention has been devoted to computerized surgical workflow analysis for various applications such as action recognition, irregularity detection, objective skill assessment, and so forth. A reliable relevant-instance-segmentation approach is a prerequisite for a majority of these applications. In this chapter, we have proposed a novel network architecture for semantic segmentation in cataract surgery videos. The proposed architecture takes advantage of three modules, namely "Pyramid View Fusion", "Deformable Pyramid Reception", and "Pyramid Loss", to simultaneously deal with different challenges. These challenges include: (i) geometric transformations such as scale variation and deformability, (ii) blur degradation and blunt edges, and (iii) transparency, and texture and color variation. Experimental results have shown the effectiveness of the proposed network architecture (DeepPyram) in retrieving the object information in all mentioned situations. DeepPyram stands in the first position for cornea, pupil, lens, and instrument segmentation compared to all rival approaches. The proposed architecture can also be adopted for various other medical image segmentation and general semantic segmentation problems.

CHAPTER

9 Self-Supervised Pretraining for Semantic Segmentation

Chapter overview — The performance of a supervised-deep-learning approach is heavily reliant on annotations. This key demand is hard to meet, especially in the case of semantic segmentation. In the medical domain, where domain knowledge is a prerequisite, providing adequate annotations is even more expensive and challenging. This chapter proposes a novel self-supervised learning approach based on contrastive learning for semantic segmentation in surgical videos. In particular, the proposed method selects and augments a pair of temporally close frames with moderate-to-harsh frequency-based and region-based augmenters. The contrastive loss encourages a close representation for the original and augmented versions while forcing distant representations between the temporally close frames. We exploit a set of progressive discrimination impediments to avoid network under-fitting due to a complex learning task in the beginning and learning suspension due to task easiness in later epochs.

This chapter is an adapted version of:

"Ghamsarian, N., Taschwer, M., Putzgruber-Adamitsch, D., Sarny, S., El-Shabrawi, Y., and Schoeffmann, K. Self-Supervised Progressive Representation Learning For Semantic Segmentation in Surgical Videos. "

9.1 Introduction

A major contributing factor in the performance of supervised deep-learning approaches is a large labeled dataset. This requirement cannot usually be met in the case of medical image analysis since it requires expert knowledge and is consequently

costly. This dearth of annotations is more severe in medical image segmentation since pixel-wise annotation is a time-extensive procedure. On the other hand, the common pre-trained backbones cannot provide optimal initialization for medical image analysis due to the large gap between the statistical distributions and semantic characteristics of the natural and medical images. During the past few years, many techniques have been developed to alleviate annotations' requirements and negate the dearth of annotations. These techniques include but are not limited to (i) data augmentation approaches such as affine and random transformations, mixup, and generative adversarial networks, (ii) semi-supervised learning, and (iii) self-supervised learning. Self-supervised learning is regarded as an effective technique to mitigate the negative impact of inadequate annotations.

Self-supervised learning refers to the methods employed to encourage the network to learn semantic features from unlabeled data. This objective is met in two ways: (1) using the inherent labels in data such as spatial and temporal characteristics, and (2) forcing the network to solve a game with the input dataset.

Despite the success of state-of-the-art approaches in alleviating the requirement for annotations, we argue that one critical aspect has not yet been explored. That is, how to stimulate a human-like semantic feature extraction through self-supervised learning. In this chapter, we propose a novel self-supervised learning framework for surgical videos. We aim to encourage learning the object's shape and configurations regardless of independent local characteristics and encapsulated rich statistics to bridge the gap between human and network interpretation. In particular, we propose:

1. A novel progressive-learning strategy by providing easy-to-hard learning tasks for the network,

2. A novel global contrastive strategy to encourage feature extraction and semantic feature deduction analogous to the human visual system (HVS),

3. Two novel region-based contrastive strategies to reinforce the learning of local

representations being advantageous for the segmentation tasks.

We argue that the proposed self-supervised learning approach can prevent capturing unnecessary local information and consequently yield better generalization capability. We compare the achievable segmentation performance with and without pre-training with the proposed approach using different amounts of annotations. Finally, we assess the impact of two modules in self-supervised learning compared to the baseline network.

9.2 Related Work

The self-supervised learning approaches can be categorized into (i) pretext-task-based approaches and (ii) contrastive learning approaches.

9.2.1 Pretext Task

In the pretext approaches, we withhold some visual characteristics of the input data and encourage the network to predict them. This objective can be achieved in two ways: regression or classification. The regression-based pretext tasks include Context restoration through context-based pixel prediction or inpainting [186], colorization [266], cross-channel prediction using split-brain autoencoders [267], and motion segmentation [184]. Orientation degree prediction [136], multi-task learning [61, 36, 226], spatial position prediction [19, 265], Discriminating between surrogate classes [62], and transformation type prediction [113] are some examples of classification-based pretext tasks. Another work [60] proposed to encourage representation learning through predicting the relative position of neighboring patches. Solving a Jigsaw puzzle [180] and Rubik cube [277, 273], and non-parametric instance discrimination [247] are other variants of representation learning through classification-based pretext tasks.

Regarding self-supervised learning from video sequences, many approaches have

been recently proposed based on video frame ordering [145, 168], video clip order prediction [248], predicting arrow of time [244], wrong order prediction [70], object tracking [185], and space-time puzzle solving [6, 126].

9.2.2 Contrastive Learning

Contrastive learning can be broadly described as the methods employed to reinforce a semantically close representation for similar pairs and distant representation for dissimilar pairs [32]. This goal can be met using mutual information maximization [100, 16], transformation-invariant instance representation learning [42, 167], momentum contrast [96], multiview coding [228], and contrastive predictive coding [99].

Contrastive learning from videos is performed via tracked patches versus random patches from videos [241], temporal cycle-consistency [66], and multiple viewpoints correspondance [211].

9.2.3 Shortcomings of State-of-The-Art Approaches

Despite the outstanding performance of state-of-the-art self-supervised learning approaches, scant attention has been devoted to self-supervised learning for semantic segmentation in surgical videos. Some approaches exploit temporal-coherence in surgical videos with contrastive and ranking loss [73]. However, since surgical videos usually contain many repetitive actions, the sequence sorting task can confuse the network and avoid semantic representation learning.

Besides, the general self-supervised learning approaches are not capable of reinforcing region-wise semantic interpretation. For instance, rotation degrees in surgical videos can be learned through simple concepts such as instrument orientation. Hence, rotation prediction seems to be an effortless task for the network to learn. Moreover, rotation prediction is not a suitable method for learning the representation of circular, deformable, and orientable objects. In surgical videos, however, we have

many relevant deformable and circular objects. Examples of cataract surgery are: (i) the implanted artificial lens as a deformable object, (ii) pupil, cornea, and iris as circular objects.

Moreover, learning to classify particular transformations such as inpainting and rotation might encourage the network to look for easy clues instead of learning shape-wise representations. Indeed, an optimal self-supervised learning approach for semantic segmentation should encourage the network to not fire on the local regions independently but on cross-region dependencies.

9.3 Methodology

Inspired by the recent advancements in self-supervised contrastive learning, we propose a self-supervised progressive representation learning for semantic segmentation in surgical videos termed VidSeg-SSL. The proposed framework encourages the network to learn the semantic features based on a contrastive loss derived from the latent representations of close frames and their augmented versions. The VidSeg-SSL framework can be flexibly used for every neural network architecture and every video dataset.

9.3.1 Notations

Everywhere in this chapter:

- $||\cdot||$ denotes the Euclidean norm.
- $|\cdot|$ denotes the absolute value function.
- $\#A$ for each arbitrary matrix or vector A, denotes the cardinality of A.
- We define a uniformly random selection function $\mathcal{R}and(K, [x_1, x_2])$ that outputs an array of K non-repetitive integer numbers between x_1 and x_2. Besides, $A[k]$ for each A and k refers to the kth element of A.

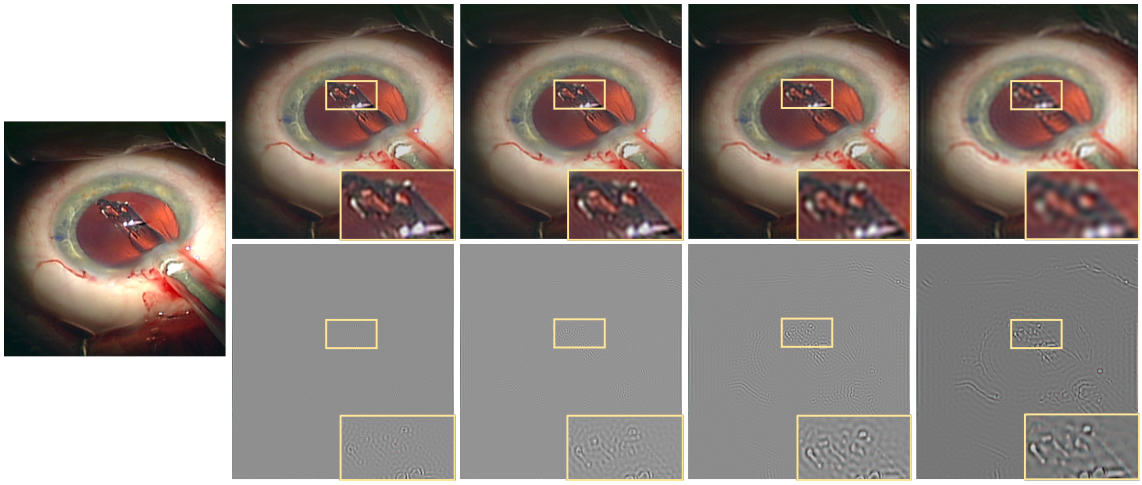


Figure 9.1: Effect of removing high-frequency components on the visual characteristics of images [238].

9.3.2 Self-Supervised Learning Algorithm

As the first step, we need to create the training set for self-supervised learning. For this purpose, we extract non-overlapping ten-second clips with the temporal resolution of 10 frames per second. Accordingly, each video clip V_i in our dataset has the dimension of $100 \times 3 \times W \times H$.

Supposing we are given a video-clip dataset $\mathcal{D} = \{V_1, V_2, ..., V_{\#\mathcal{D}}\}$ with $V_i \in \mathbb{R}^{100\times3\times W\times H}$, a stochastic function $\mathcal{F}_t$ is employed to randomly select a training clip V_t as follows:

$$V_t = \mathcal{D}[\mathcal{F}_t], \mathcal{F}_t = \mathcal{R}and(1, [1, \#\mathcal{D}]) \tag{9.1}$$

We formulate the selected video-clip as a sequence of its frames as follows:

$$V_t = (\mathrm{I}_1, \mathrm{I}_2, ..., \mathrm{I}_{100}) \tag{9.2}$$

A second stochastic function $\mathcal{F}_e$ is used to randomly select an exemplary frame I_e

from 100 possible frames in the selected video as follows:

$$\mathrm{I}_e = V_t[\mathcal{F}_e], \mathcal{F}_e = \mathcal{R}and(1, [1, 100]) \tag{9.3}$$

Depending on the pre-determined frame-distance threshold T_{dist} and the index of exemplary frame $\mathrm{ind}(\mathrm{I}_e)$, the third stochastic function $\mathcal{F}_a$ selects an adversarial frame I_a as follows:

$$\mathrm{I}_a = V_t[\mathcal{F}_a], \mathcal{F}_a = \mathcal{R}and(1, [max(0, \mathrm{ind}(\mathrm{I}_e) - \mathrm{T}_{dist}), min(50, \mathrm{ind}(\mathrm{I}_e) + \mathrm{T}_{dist})] \tag{9.4}$$

In order to be sure that the sampled pair of frames I_e and I_a have enough visual discriminative features, we subtract these frames ($\mathrm{I}_{diff} = |\mathrm{I}_e - \mathrm{I}_e|$) and search for the elements with large absolute values in the difference matrix I_{diff} based on a pre-determined pixel-wise difference threshold T_{diff} as follows:

$$C(\mathrm{I}_{diff}) = \{x \in \mathrm{I}_{diff} : x > \mathrm{T}_{diff}\} \tag{9.5}$$

We check the suitability of the two frames by counting the number of large values in the difference frame as follows:

$$S(\mathrm{I}_{diff}) = \begin{cases} 1, & \text{if } \frac{\#C(\mathrm{I}_{diff})}{\#\mathrm{I}_{diff}} \geq 0.1 \\ 0, & \text{otherwise} \end{cases} \tag{9.6}$$

We pass the frames to the next step if $S(\mathrm{I}_{diff}) = 1$. Otherwise, we through the sampled frames away, choose an alternative clip, and repeat sampling until the sampled pair satisfy the mentioned condition. Afterwards, exemplary frame and adversarial frame undergo gradual transformations to produce contrastive pairs with

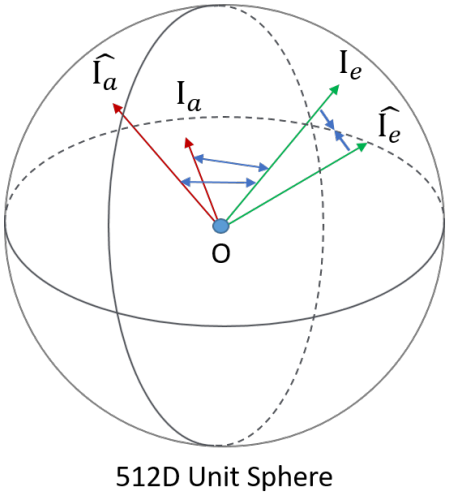


Figure 9.2: The proposed contrastive learning strategy.

different strategies:

$$\begin{aligned} \hat{\mathrm{I}}_e &= \mathcal{T}(\mathrm{I}_e) \\ \hat{\mathrm{I}}_a &= \mathcal{T}(\mathrm{I}_a) \end{aligned} \tag{9.7}$$

We explain our proposed novel strategies in the following three subsections.

In contrastive learning, the objective is to increase (i) the similarity between the representation of the exemplar image I_e and its transformed version $\hat{\mathrm{I}}_e$, and (ii) cross dissimilarities between all versions of exemplar and adversarial frames. To this end, we use a contrastive loss based on cosine similarity between the representations of exemplary and adversarial frames and their transformations as follows [41].

$$\mathcal{L}(E(\mathrm{I}_e), E(\mathrm{I}_a)) = -\log \frac{exp(sim(E(\mathrm{I}_e), E(\hat{\mathrm{I}}_e))/\tau)}{\sum_{\mathrm{I}_1,\mathrm{I}_2} exp(sim(E(\mathrm{I}_1), E(\mathrm{I}_2))/\tau)} \tag{9.8}$$

In (9.8), $\mathrm{I}_1 \in \{\mathrm{I}_e, \hat{\mathrm{I}}_e\}$, $\mathrm{I}_2 \in \{\mathrm{I}_a, \hat{\mathrm{I}}_a\}$, and τ is the pre-determined scaling parameter. The function $E(\cdot)$ refers to the representation of the encoder network $e(\cdot)$ appended by a projection head $p(\cdot)$ as $E(x) = p(e(x))$. The projection head is used to provide

Algorithm 1 VidSeg-SSL's learning algorithm

Input: batch size N, Network $E(\cdot)$, epoch-adaptive transformation $\mathcal{T}$, Video clip dataset $\mathcal{D}$.
Parameter: frame-distance threshold T_{dist}, pixel-wise difference threshold T_{diff},
Output: pre-trained encoder $e(\cdot)$.

1: Let $t = 0$.
2: **for** sampled video-clip batch **do**
3: **for** training clip V_t in the video-clip batch b_V ($b_V = \{V_i | i \in [1, \#\mathcal{D}\}, \#b_V = N$) **do**
4: Select the exemplary frame I_e using eq. (9.3).
5: Select the adversarial frame I_a using eq. (9.4).
6: Compute the frame differences $C(\mathrm{I}_{diff})$ using eq. (9.5).
7: **if** $S(\mathrm{I}_{diff}) = 0$ (eq. (9.6)) **then**
8: Remove V_t from the video-clip batch b_V and add an alternative video clip to the batch.
9: Repeat the frame selection step (go to line 3).
10: **end if**
11: Transform I_e and I_a to obtain $\hat{\mathrm{I}_e}$ and $\hat{\mathrm{I}_a}$ using eq. (9.7).
12: Compute the latent representations of the exemplary and adversarial frames and their transformations using feed-forward:
$E(\mathrm{I}_e)$, $E(\mathrm{I}_a)$, $E(\hat{\mathrm{I}_e})$, $E(\hat{\mathrm{I}_a})$.
13: Compute the contrastive loss for the current samples using eq. (9.8).
14: Set $l[t] = \mathcal{L}(E(\mathrm{I}_e), E(\mathrm{I}_a))$.
15: **end for**
16: Compute the overall loss for the batch:
$L = \frac{1}{N} \sum_{t=1}^{N} l[t]$.
17: Update $E(\cdot)$ including the encoder $e(\cdot)$ and projection the head $p(\cdot)$ using the overall batch loss L.
18: **end for**
19: Remove the projection head from $E(\cdot)$ to obtain the encoder network $e(\cdot)$.
20: **return** the pretrained network $e(\cdot)$.

a lower-dimensionality representation as well as allowing the encoder network $e(\cdot)$ to maintain transformation-related features [41]. Beside, $sim(x, y)$ refers to the cosine similarity between the vectors x and y and is formulated as $sim(x, y) = (x^T y)/max(\| x \| \| y \|, \epsilon)$ where ϵ is a very small value to avoid division by zero.

The threshold T_a is reset to its highest value at the beginning of each training strategy and is gradually decreases to increase the task difficulty.

9.3.3 Strategy 1: contrastive learning based on high-frequency component removal

In this stage, we aim to encourage the network to output a very close representation for each examplar frame I_e and its transformed version $\hat{\mathrm{I}_e}$ obtained via removing the high-frequency components. In contrast, we reinforce distant representations for the examplar and adversarial frames. The removal of high-frequency components does not affect the semantic information perceived by the human visual system. Accordingly, we reinforce a neural network deduction of semantic information analogous to the human visual system.

The high-frequency-removal transformation function $\mathcal{T}^r_{HF}$ uses a pre-determined parameter r to remove the high-frequency components outside of the circle with radius r at the center of the fast Fourier transform (FFT) version of the exemplary and adversarial frames. Then we obtain:

$$\hat{\mathrm{I}_e} = \mathcal{T}^r_{HF}(\mathrm{I}_e)$$
$$\hat{\mathrm{I}_a} = \mathcal{T}^r_{HF}(\mathrm{I}_a)$$

The radial is scheduled to gradually decrease in order to provide a more complicated task for the network. However, we consider a pre-determined minimum radial to avoid the removal of semantically relevant components.

Discussion: With this strategy, we aim to address the misalignment between the semantic interpretation of CNNs and the human visual system (HVS). While HVS determines the label purely based on semantic information, the convolutional neural networks unintentionally learn joint high-frequency and semantic correlations between the visual signals and their corresponding labels.

9.3.4 Strategy 2: contrastive learning based on block-wise augmentation

In this stage, we further change the frames transformed by Strategy 1 to increase the difficulty of the task and as a result, encourage the network to extract more discriminative information from the video frames. We use a set of non-geometric transformations (such as Gaussian and motion blur, brightness, and contrast transformations) for block-wise augmentation. More concretely, the examplar and adversarial frames are split into $k \times k$ blocks. Supposing $k = 4$, the frames after high-frequency removal can be shown as the sequence of their blocks as follows:

$$\mathcal{T}^r_{HF}(\mathrm{I}_e) = (b_{e1}, b_{e2}, ..., b_{e15}, b_{e16})$$
$$\mathcal{T}^r_{HF}(\mathrm{I}_a) = (b_{a1}, b_{a2}, ..., b_{a15}, b_{a16})$$

Each block is distinctively augmented using two stochastic functions $\mathcal{F}_b$ and $\mathcal{F}_{Aug}$. The function $\mathcal{F}_b$ randomly selects n blocks in the current frame to be augmented. For instance, supposing $n = 2$, $\mathcal{F}_b(\mathrm{I}_e) = \{b_{e1}, b_{e15}\}$, and $\mathcal{F}_b(\mathrm{I}_a) = \{b_{e2}, b_{e16}\}$, the frames $\hat{\mathrm{I}_e}$ and $\hat{\mathrm{I}_a}$ can be shown as:

$$\hat{\mathrm{I}_e} = \mathcal{F}_{Aug}(\mathcal{T}^r_{HF}(\mathrm{I}_e), \mathcal{F}_b(\mathrm{I}_e)) = (\hat{b_{e1}}, b_{e2}, ..., \hat{b_{e15}}, b_{e16})$$
$$\hat{\mathrm{I}_a} = \mathcal{F}_{Aug}(\mathcal{T}^r_{HF}(\mathrm{I}_a), \mathcal{F}_b(\mathrm{I}_a)) = (b_{a1}, \hat{b_{a2}}, ..., b_{a15}, \hat{b_{a16}})$$

The function $\mathcal{F}_{Aug}(a, b)$ transforms the input frame a by augmenting the blocks in the array b using a random selection of augmentation for each block independently of the transformations applied to other blocks. The decomposition size (k) progressively decreases, and the number of augmented blocks n progressively increases during training to provide a more complex task for the network.

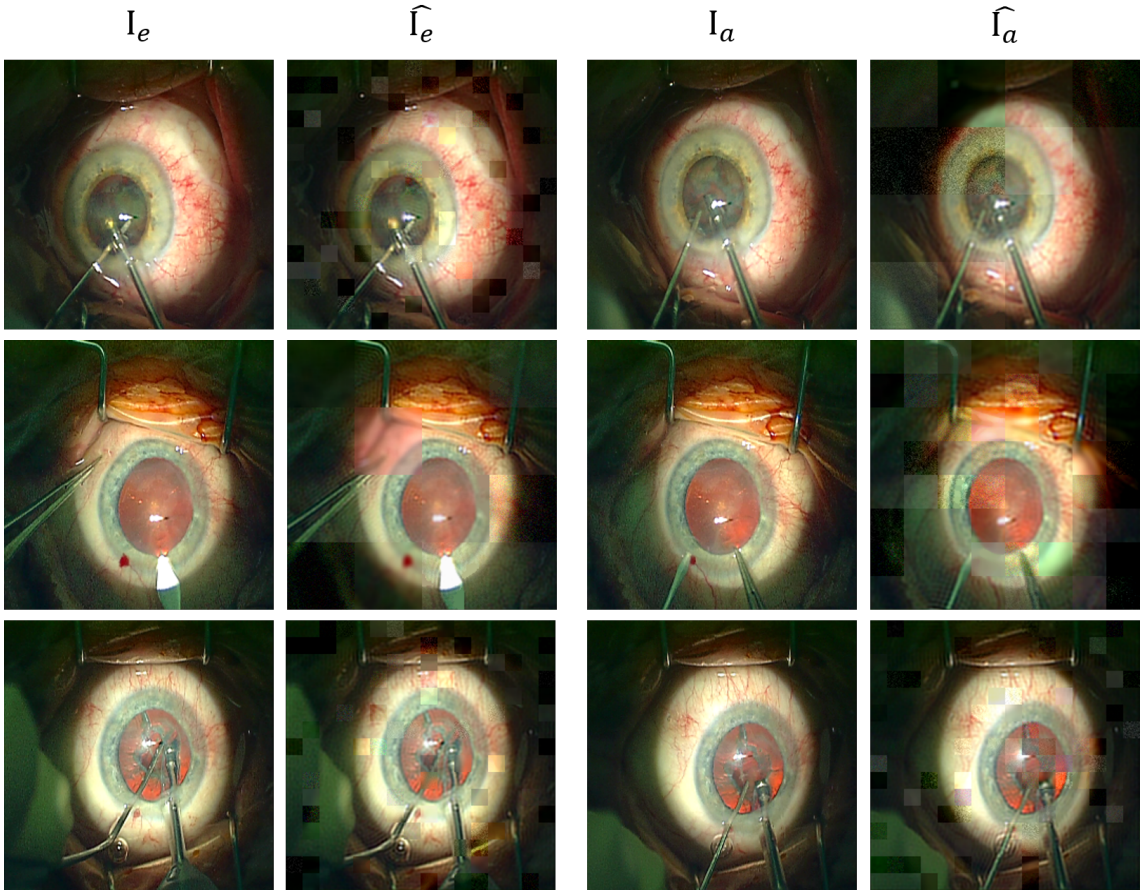


Figure 9.3: The training quadruple images generated with Strategy 2.

Discussion: As shown in Figure 9.3, block-wise augmentation produces many irrelevant edges and thus encourages the network to learn the relevant edges to maximize the representation agreements between the positive pairs.

9.3.5 Strategy 3 : contrastive learning based on deformable-block-wise augmentation

Block-wise augmentation leads to straight edges in the images. As a result, a network trained on such augmented images may learn to disregard the straight edges after some epochs, leading to stopping learning the underlying semantic representation. Hence, we propose to impede the contrastive learning task via augmenting deformable blocks. To produce such deformations, we take advantage of piece-wise affine transformations (Figure 9.4). After splitting the image into $k \times k$ blocks and selecting the blocks to be

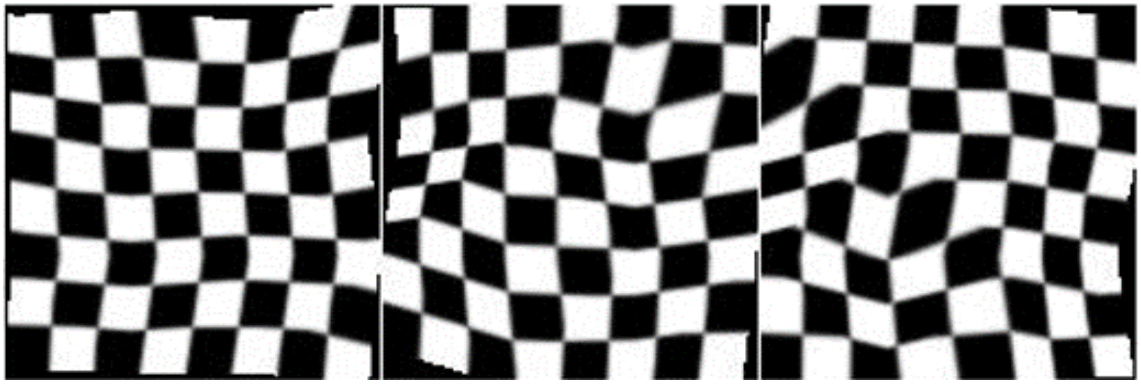


Figure 9.4: Deformation of blocks' borders via piece-wise affine transformation (picture from https://imgaug.readthedocs.io/).

augmented, we form a mask image per block with the white pixels corresponding to the selected block's pixels. We then apply the affine transformation on the mask to obtain a deformed block. By multiplying the obtained mask to the original image, we obtain the pixels corresponding to the deformable block that should be augmented. Similar to Strategy 2, we apply a random selection of pre-determined (non-geometric) transformations on the selected images. The augmented image is finally obtained by replacing the pixels in the original image with non-zero pixels in the augmented masks for all selected blocks. As shown in Figure 9.5, deformable-block-wise augmentation leads to some irrelevant curved edges that are more difficult than the straight lines in strategy 2 to distinguish from the relevant edges.

9.4 Experimental Settings

In this chapter, we evaluate and compare the performance of representation learning frameworks using four evaluation setups as follows:

- **Setup 1.** To evaluate the learned representations for classification, many approaches freeze the self-supervised pre-trained encoder, train a linear classifier added on top of the encoder network, and evaluate the whole network [135, 16, 42]. We use the same approach in this setup by adding the decoder network on top of the frozen pre-trained encoder, training the whole network for semantic

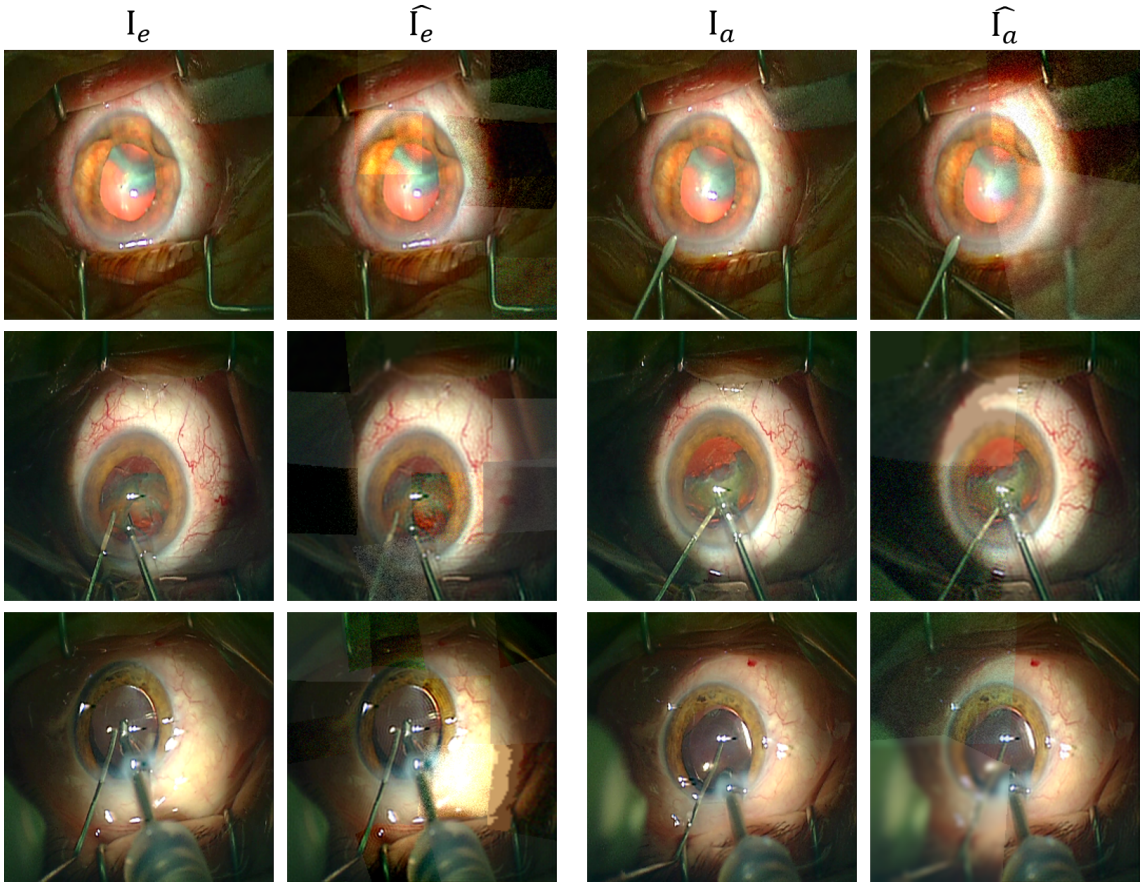


Figure 9.5: The training quadruple images generated with Strategy 3.

segmentation. We also use this approach as the baseline for other setups.

- **Setup 2.** In this setup, we do not freeze the encoder network and fine-tune the whole network. The results reveal how the pre-trained weights affect the speed and performance of learning.

- **Setup 3.** We adopt a semi-supervised learning approach for semantic segmentation to see which of the self-supervised learning approaches leads to less outlier and more centralized semantic segmentation interpretation. In particular, we want to figure out which SSL approach can prevent the network from biasing to unexpected features within the dataset.

- **Setup 4.** As the main reason behind self-supervised learning, we assess the performance of a pre-trained network with supervised learning on a dataset such

as ImageNet with the pre-trained networks using self-supervised learning. More concretely, in this setup, we evaluate the percentage of annotations required after self-supervised learning to achieve the same semantic segmentation performance.

- **Setup 5.** As an ablation study, we assess the effect of three modules based on average pooling on SSL performance.

9.5 Conclusion

Considering that supervised learning based on manual annotation is time-consuming, expensive, prone to human error, and subject to distribution shift, self-supervised representation learning has witnessed considerable attention in recent years. The proposed framework does not impose any constraints on the choice of the neural network architecture and video dataset. The proposed representation learning approach can initialize the networks for semantic segmentation tasks and other various downstream tasks.

CHAPTER

10 Concluding Remarks

Chapter overview — This chapter concludes the contributions of this thesis and discusses the open-ended research questions regarding deep-learning-assisted analysis of cataract surgery videos.

The recent technological advancements in robotic and minimally invasive surgeries have enabled recording and collecting surgical videos. Analyzing such videos can provide valuable insights to enhance post-surgical care, speed up surgical training, discover unexplored symptoms of surgical complications, and provide real-time guidance for OR planning. Indeed, computerized surgical video analysis can offer solutions to numerous demands of the modern operating rooms.

This thesis investigates novel frameworks and neural network architectures to address the significant research questions about the computerized analysis of cataract surgery videos. In particular, this thesis focuses on four critical subjects in computerized analysis of cataract surgery videos: (1) phase recognition, (2) semantic segmentation, (3) adaptive compression, and (4) irregularity detection.

10.1 Contributions of This Dissertation

Providing a powerful basis for high-level tasks. Phase recognition and semantic segmentation are two integral parts of many surgical video analysis approaches. Hence, accurate semantic segmentation and phase recognition can play a critical role in

subsequent downstream tasks such as objective skill assessment, relevance detection, irregularity detection, adaptive compression, etc. To fulfill these requirements, this thesis proposes (i) novel frameworks, recurrent convolutional networks, and training strategies to improve phase recognition accuracy upon state-of-the-art approaches, and (ii) several novel convolutional modules to deal with diverse semantic segmentation challenges in cataract surgery videos.

Relevance-Based Compression. High visual quality of relevant content is a key factor in the usefulness of cataract surgery videos. This pre-condition notwithstanding, compression is a necessity for real-time streaming and efficient storage of cataract surgery videos. To accommodate both demands, we propose relevance-based compression of cataract surgery videos considering different scenarios for the relevant content.

Irregularity Detection. Deep neural networks are powerful means for performing large-scale evaluations to detect the symptoms and reason intra-operative irregularities and post-operative complications in surgical videos. We employ the capacity of deep neural networks and propose the first framework for automatic lens irregularity detection in cataract surgery videos.

Self-supervised pre-training. To alleviate the requirement for manual annotations, we propose novel self-supervised learning strategies. These strategies especially focus on reinforcing high-level semantic interpretation of raw video frames being determinative for semantic segmentation.

10.2 Future Work

Despite the recent advancements in computerized surgical video analysis, two critical aspects are mainly unexplored: (1) facilitating the use of 3D CNNs via self-supervised

pre-training and (2) enhancing the generalization performance in semantic segmentation via domain adaptation.

10.2.1 Self-Supervised Pre-Training of 3D CNNs

Although recurrent CNNs have shown superior performance in phase recognition and relevance detection, they may not present a satisfactory performance in action recognition and skill assessment. This is due to the fact that actions involve intertwined spatio-temporal features, the property that recurrent CNNs cannot efficiently capture. On the other hand, action recognition and skill assessment via relevant object segmentation and motion trajectories appear to be suboptimal since (i) providing the supervisory signal for semantic segmentation is more time-consuming and expensive compared to action annotation, and (ii) action recognition and skill assessment using a sequence of steps may decrease the time and computation efficiency.

Three-dimensional CNNs can effectively capture joint spatio-temporal features associated with a particular action or skill level. However, employing 3D CNNs involves its specific challenges. Unlike 2D CNNs, for which many large-scale datasets for pre-training exist (examples are ImageNet [55], COCO [153], and Cityscapes [50]), no large-scale dataset for pre-training the 3D CNNs exists. As a result, training 3D CNNs involves starting from scratch with random parameter initialization. Comparing to a pre-trained network, starting from random weights requires substantially more supervisory signals, imposing higher costs and more burden for further annotations. As mentioned in chapter 9, self-supervised learning from raw video frames can provide optimal initial states for the neural networks and negate the lack of large-scale supervisory signal. Accordingly, using domain-specific (cataract surgery features) and problem-specific (action recognition or skill assessment) strategies for self-supervised learning can make ground on exploiting 3D CNNs.

10.2.2 Domain Adaptation for Semantic Segmentation

The rapid technological advancements have resulted in continuous changes in image and video capturing tools, lighting conditions, and compression standards that in turn lead to large distribution shifts from previous to new datasets. This distribution shift exists not only between different dataset generations but also between the contemporary datasets captured with different cameras and in different conditions. Such domain mismatch between the datasets cripples the performance of models trained on one dataset when being tested on the other datasets.

Domain mismatch necessitates unjustifiable human effort and cost for new annotations per dataset, a condition which is very hard to meet, especially in the case of pixel-level annotation for semantic segmentation. Domain adaptation suggests techniques to cut down or bypass the requirement for annotations in new domains. In the last couple of years, many attentions have been focused on domain adaptation [150, 33, 127, 206, 87, 95]. However, domain adaptation for semantic segmentation in cataract surgery videos taking advantage of domain-specific knowledge (unique characteristics of cataract surgery videos) is an important subject that is negated.

Bibliography

[1] Sub-challenge workflow detection from laparoscopic videos. https://endovissub-workflow.grand-challenge.org/.

[2] Abdulla, W. Mask r-cnn for object detection and instance segmentation on keras and tensorflow. https://github.com/matterport/Mask_RCNN, 2017.

[3] Achiron, A., Haddad, F., Gerra, M., Bartov, E., and Burgansky-Eliash, Z. Predicting cataract surgery time based on preoperative risk assessment. *European Journal of Ophthalmology 26*, 3 (2016), 226–229. PMID: 26541113.

[4] Ahmidi, N., Tao, L., Sefati, S., Gao, Y., Lea, C., Haro, B. B., Zappella, L., Khudanpur, S., Vidal, R., and Hager, G. D. A dataset and benchmarks for segmentation and recognition of gestures in robotic surgery. *IEEE Transactions on Biomedical Engineering 64*, 9 (2017), 2025–2041.

[5] Ahmidi, N., Tao, L., Sefati, S., Gao, Y., Lea, C., Haro, B. B., Zappella, L., Khudanpur, S., Vidal, R., and Hager, G. D. A dataset and benchmarks for segmentation and recognition of gestures in robotic surgery. *IEEE Transactions on Biomedical Engineering 64*, 9 (2017), 2025–2041.

[6] Ahsan, U., Madhok, R., and Essa, I. Video jigsaw: Unsupervised learning of spatiotemporal context for video action recognition. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)* (2019), pp. 179–189.

[7] Aizenberg, I., Paliy, D., Moraga, C., and Astola, J. Blur Identification Using Neural Network for Image Restoration. In *Computational Intelligence, Theory and Applications* (Berlin, Heidelberg, 2006), B. Reusch, Ed., Springer Berlin Heidelberg, pp. 441–455.

[8] Aksamentov, I., Twinanda, A. P., Mutter, D., Marescaux, J., and Padoy, N. Deep neural networks predict remaining surgery duration from cholecystectomy videos. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2017* (Cham, 2017), Springer International Publishing, pp. 586–593.

[9] Al Hajj, H., et al. Automatic tool annotation for surgical workflow analysis. https://cataracts.grand-challenge.org/, 2017.

[10] Al Hajj, H., et al. Jigsaws: The jhu-isi gesture and skill assessment working set. https://cirl.lcsr.jhu.edu/research/hmm/datasets/jigsaws_release/, 2017.

[11] Al Hajj, H., Lamard, M., Conze, P.-H., Cochener, B., and Quellec, G. Monitoring tool usage in surgery videos using boosted convolutional and recurrent neural networks. *Medical Image Analysis 47* (2018), 203–218.

[12] 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., and Quellec, G. Cataracts: Challenge on automatic tool annotation for cataract surgery. *Medical Image Analysis 52* (2019), 24 – 41.

[13] Allan, M., Kondo, S., Bodenstedt, S., Leger, S., Kadkhodamohammadi, R., Luengo, I., Fuentes-Hurtado, F., Flouty, E., Mohammed, A. K., Pedersen, M., Kori, A., Varghese, A., Krishnamurthi, G., Rauber, D., Mendel, R., Palm, C., Bano, S., Saibro, G., Shih, C., Chiang, H., Zhuang, J., Yang, J., Iglovikov, V., Dobrenkii, A., Reddiboina, M., Reddy, A., Liu, X., Gao, C., Unberath, M., Azizian, M., Stoyanov, D., Maier-Hein, L., and Speidel, S. 2018 robotic scene segmentation challenge. *CoRR abs/2001.11190* (2020).

[14] Allan, M., Shvets, A., Kurmann, T., Zhang, Z., Duggal, R., Su, Y., Rieke, N., Laina, I., Kalavakonda, N., Bodenstedt, S., García-Peraza, L. C., Li, W., Iglovikov, V., Luo, H., Yang, J., Stoyanov, D., Maier-Hein, L., Speidel, S., and Azizian, M. 2017 robotic instrument segmentation challenge. *CoRR abs/1902.06426* (2019).

[15] Arnab, A., and Torr, P. H. S. Pixelwise instance segmentation with a dynamically instantiated network. *CoRR abs/1704.02386* (2017).

[16] Bachman, P., Hjelm, R. D., and Buchwalter, W. Learning representations by maximizing mutual information across views. *CoRR abs/1906.00910* (2019).

[17] Badrinarayanan, V., Kendall, A., and Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence 39*, 12 (2017), 2481–2495.

[18] Bai, M., and Urtasun, R. Deep watershed transform for instance segmentation. *CoRR abs/1611.08303* (2016).

[19] Bai, W., Chen, C., Tarroni, G., Duan, J., Guitton, F., Petersen, S. E., Guo, Y., Matthews, P. M., and Rueckert, D. Self-supervised learning for cardiac mr image segmentation by anatomical position prediction. In *Medical Image Computing and Computer Assisted Intervention – MICCAI*

2019 (Cham, 2019), D. Shen, T. Liu, T. M. Peters, L. H. Staib, C. Essert, S. Zhou, P.-T. Yap, and A. Khan, Eds., Springer International Publishing, pp. 541–549.

[20] Bardram, J. E., Doryab, A., Jensen, R. M., Lange, P. M., Nielsen, K. L. G., and Petersen, S. T. Phase recognition during surgical procedures using embedded and body-worn sensors. In *2011 IEEE International Conference on Pervasive Computing and Communications (PerCom)* (2011), pp. 45–53.

[21] Bertasius, G., and Torresani, L. Classifying, segmenting, and tracking object instances in video with mask propagation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), pp. 9736–9745.

[22] Bodenstedt, S., Allan, M., Agustinos, A., Du, X., García-Peraza-Herrera, L. C., Kenngott, H., Kurmann, T., Müller-Stich, B. P., Ourselin, S., Pakhomov, D., Sznitman, R., Teichmann, M., Thoma, M., Vercauteren, T., Voros, S., Wagner, M., Wochner, P., Maier-Hein, L., Stoyanov, D., and Speidel, S. Comparative evaluation of instrument segmentation and tracking methods in minimally invasive surgery. *CoRR abs/1805.02475* (2018).

[23] Bodenstedt, S., et al. Multi-instrument endovis challenge dataset. `https://endovissub-instrument.grand-challenge.org/`, 2015.

[24] Bodenstedt, S., Rivoir, D., Jenke, A., Wagner, M., Breucha, M., Müller-Stich, B., Mees, S. T., Weitz, J., and Speidel, S. Active learning using deep bayesian networks for surgical workflow analysis. *International Journal of Computer Assisted Radiology and Surgery 14*, 6 (Jun 2019), 1079–1087.

[25] Bodenstedt, S., Wagner, M., Katic, D., Mietkowski, P., Mayer, B. F. B., Kenngott, H., Müller-Stich, B. P., Dillmann, R., and Speidel, S. Unsupervised temporal context learning using convolutional neural networks for laparoscopic workflow analysis. *CoRR abs/1702.03684* (2017).

[26] Bossen, F., et al. Common test conditions and software reference configurations. *JCTVC-L1100 12* (2013), 7.

[27] Buslaev, A., Iglovikov, V. I., Khvedchenya, E., Parinov, A., Druzhinin, M., and Kalinin, A. A. Albumentations: Fast and flexible image augmentations. *Information 11*, 2 (Feb 2020), 125.

[28] Cai, C., Chen, L., Zhang, X., and Gao, Z. End-to-end optimized roi image compression. *IEEE Transactions on Image Processing 29* (2020), 3442–3457.

[29] Carreira, J., and 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 (CVPR)* (July 2017).

[30] Carreira, J., and Zisserman, A. Quo vadis, action recognition? A new model and the kinetics dataset. *CoRR abs/1705.07750* (2017).

[31] Castells, X., Comas, M., Castilla, M., Cots, F., and Alarcón, S. Clinical outcomes and costs of cataract surgery performed by planned ECCE and phacoemulsification. *International ophthalmology 22* (02 1998), 363–7.

[32] Chaitanya, K., Erdil, E., Karani, N., and Konukoglu, E. Contrastive learning of global and local features for medical image segmentation with limited annotations. In *Advances in Neural Information Processing Systems* (2020), H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds., vol. 33, Curran Associates, Inc., pp. 12546–12558.

[33] Chang, W.-L., Wang, H.-P., Peng, W.-H., and Chiu, W.-C. All about structure: Adapting structural information across domains for boosting semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2019).

[34] Charrière, K., Quellec, G., Lamard, M., Martiano, D., Cazuguel, G., Coatrieux, G., and Cochener, B. Real-time analysis of cataract surgery videos using statistical models. *Multimedia Tools and Applications 76*, 21 (Nov 2017), 22473–22491.

[35] Charriere, K., Quelled, G., Lamard, M., Martiano, D., Cazuguel, G., Coatrieux, G., and Cochener, B. Real-time multilevel sequencing of cataract surgery videos. In *2016 14th International Workshop on Content-Based Multimedia Indexing (CBMI)* (June 2016), pp. 1–6.

[36] Chen, L., Bentley, P., Mori, K., Misawa, K., Fujiwara, M., and Rueckert, D. Self-supervised learning for medical image analysis using image context restoration. *Medical Image Analysis 58* (2019), 101539.

[37] Chen, L., Hermans, A., Papandreou, G., Schroff, F., Wang, P., and Adam, H. Masklab: Instance segmentation by refining object detection with semantic and direction features. *CoRR abs/1712.04837* (2017).

[38] Chen, L., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *CoRR abs/1606.00915* (2016).

[39] Chen, L.-C., Yang, Y., Wang, J., Xu, W., and Yuille, A. L. Attention to scale: Scale-aware semantic image segmentation. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), pp. 3640–3649.

[40] Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., and Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)* (September 2018).

[41] Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning* (13–18 Jul 2020), H. D. III and A. Singh, Eds., vol. 119 of *Proceedings of Machine Learning Research*, PMLR, pp. 1597–1607.

[42] Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. E. A simple framework for contrastive learning of visual representations. *CoRR abs/2002.05709* (2020).

[43] Chen, X., Zhang, R., and Yan, P. Feature fusion encoder decoder network for automatic liver lesion segmentation. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)* (2019), pp. 430–433.

[44] Chmarra, M. K., Grimbergen, C. A., and Dankelman, J. Systems for tracking minimally invasive surgical instruments. *Minimally Invasive Therapy & Allied Technologies 16*, 6 (2007), 328–340.

[45] Cho, C. ., and Don, H. . Blur identification and image restoration using a multilayer neural network. In *[Proceedings] 1991 IEEE International Joint Conference on Neural Networks* (Nov 1991), pp. 2558–2563 vol.3.

[46] Cho, K., van Merrienboer, B., Gülçehre, Ç., Bougares, F., Schwenk, H., and Bengio, Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *CoRR abs/1406.1078* (2014).

[47] Cho, S. M., Kim, Y., Jeong, J., Lee, H., and Kim, N. Automatic tip detection of surgical instruments in biportal endoscopic spine surgery. *CoRR abs/1911.02755* (2019).

[48] Choi, B., Jo, K., Choi, S., and Choi, J. Surgical-tools detection based on convolutional neural network in laparoscopic robot-assisted surgery. In *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (2017), pp. 1756–1759.

[49] Colleoni, E., Edwards, P., and Stoyanov, D. Synthetic and real inputs for tool segmentation in robotic surgery. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020* (Cham, 2020), A. L. Martel, P. Abolmaesumi, D. Stoyanov, D. Mateus, M. A. Zuluaga, S. K. Zhou, D. Racoceanu, and L. Joskowicz, Eds., Springer International Publishing, pp. 700–710.

[50] Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., and Schiele, B. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016).

[51] Cui, H., Liu, X., and Huang, N. Pulmonary vessel segmentation based on orthogonal fused u-net++ of chest ct images. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019* (Cham, 2019), D. Shen, T. Liu, T. M. Peters, L. H. Staib, C. Essert, S. Zhou, P.-T. Yap, and A. Khan, Eds., Springer International Publishing, pp. 293–300.

[52] Czempiel, T., Paschali, M., Keicher, M., Simson, W., Feussner, H., Kim, S. T., and Navab, N. Tecno: Surgical phase recognition with multi-stage temporal convolutional networks. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020* (Cham, 2020), A. L. Martel, P. Abolmaesumi, D. Stoyanov, D. Mateus, M. A. Zuluaga, S. K. Zhou, D. Racoceanu, and L. Joskowicz, Eds., Springer International Publishing, pp. 343–352.

[53] Dai, J., Li, Y., He, K., and Sun, J. R-FCN: object detection via region-based fully convolutional networks. *CoRR abs/1605.06409* (2016).

[54] Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., and Wei, Y. Deformable convolutional networks. In *2017 IEEE International Conference on Computer Vision (ICCV)* (2017), pp. 764–773.

[55] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09* (2009).

[56] Dexter, F., Epstein, R. H., Lee, J. D., and Ledolter, J. Automatic updating of times remaining in surgical cases using bayesian analysis of historical case duration data and "instant messaging" updates from anesthesia providers. *Anesthesia & Analgesia 108*, 3 (2009).

[57] Ding, Y., Fan, J., Pang, K., Li, H., Fu, T., Song, H., Chen, L., and Yang, J. Surgical workflow recognition using two-stream mixed convolution network. In *2020 3rd International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE)* (2020), pp. 264–269.

[58] DiPietro, R., and Hager, G. D. Unsupervised learning for surgical motion by learning to predict the future. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018* (Cham, 2018), A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López, and G. Fichtinger, Eds., Springer International Publishing, pp. 281–288.

[59] DiPietro, R., and Hager, G. D. Automated surgical activity recognition with one labeled sequence. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019* (Cham, 2019), D. Shen, T. Liu, T. M. Peters, L. H. Staib, C. Essert, S. Zhou, P.-T. Yap, and A. Khan, Eds., Springer International Publishing, pp. 458–466.

[60] Doersch, C., Gupta, A., and Efros, A. A. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (December 2015).

[61] Doersch, C., and Zisserman, A. Multi-task self-supervised visual learning. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (Oct 2017).

[62] Dosovitskiy, A., Springenberg, J. T., Riedmiller, M., and Brox, T. Discriminative unsupervised feature learning with convolutional neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1* (Cambridge, MA, USA, 2014), NIPS'14, MIT Press, p. 766–774.

[63] Dou, Q., Chen, H., Jin, Y., Yu, L., Qin, J., and Heng, P. 3d deeply supervised network for automatic liver segmentation from CT volumes. *CoRR abs/1607.00582* (2016).

[64] Du, X., Allan, M., Dore, A., Ourselin, S., Hawkes, D. J., Kelly, J. D., and Stoyanov, D. Combined 2d and 3d tracking of surgical instruments for minimally invasive and robotic-assisted surgery. *International Journal of Computer Assisted Radiology and Surgery 11* (2016), 1109 – 1119.

[65] Du, X., Kurmann, T., Chang, P.-L., Allan, M., Ourselin, S., Sznitman, R., Kelly, J. D., and Stoyanov, D. Articulated multi-instrument 2-d pose estimation using fully convolutional networks. *IEEE Transactions on Medical Imaging 37*, 5 (2018), 1276–1287.

[66] Dwibedi, D., Aytar, Y., Tompson, J., Sermanet, P., and Zisserman, A. Temporal cycle-consistency learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2019).

[67] Fan, D.-P., Zhou, T., Ji, G.-P., Zhou, Y., Chen, G., Fu, H., Shen, J., and Shao, L. Inf-net: Automatic covid-19 lung infection segmentation from ct images. *IEEE Transactions on Medical Imaging 39*, 8 (2020), 2626–2637.

[68] Fathi, A., Wojna, Z., Rathod, V., Wang, P., Song, H. O., Guadarrama, S., and Murphy, K. P. Semantic instance segmentation via deep metric learning. *CoRR abs/1703.10277* (2017).

[69] Feng, S., Zhao, H., Shi, F., Cheng, X., Wang, M., Ma, Y., Xiang, D., Zhu, W., and Chen, X. Cpfnet: Context pyramid fusion network for medical image segmentation. *IEEE Transactions on Medical Imaging 39*, 10 (2020), 3008–3018.

[70] Fernando, B., Bilen, H., Gavves, E., and Gould, S. Self-supervised video representation learning with odd-one-out networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (July 2017).

[71] Follmann, P., and König, R. Oriented boxes for accurate instance segmentation, 2019.

[72] Funke, I., Jenke, A., Mees, S. T., Weitz, J., Speidel, S., and Bodenstedt, S. Temporal coherence-based self-supervised learning for laparoscopic workflow analysis. *CoRR abs/1806.06811* (2018).

[73] Funke, I., Jenke, A., Mees, S. T., Weitz, J., Speidel, S., and Bodenstedt, S. Temporal coherence-based self-supervised learning for laparoscopic workflow analysis. In *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis* (Cham, 2018), D. Stoyanov, Z. Taylor, D. Sarikaya, J. McLeod, M. A. González Ballester, N. C. Codella, A. Martel, L. Maier-Hein, A. Malpani, M. A. Zenati, S. De Ribaupierre, L. Xiongbiao, T. Collins, T. Reichl, K. Drechsler, M. Erdt, M. G. Linguraru, C. Oyarzun Laura, R. Shekhar, S. Wesarg, M. E. Celebi, K. Dana, and A. Halpern, Eds., Springer International Publishing, pp. 85–93.

[74] Funke, I., Mees, S. T., Weitz, J., and Speidel, S. Video-based surgical skill assessment using 3d convolutional neural networks. *International Journal of Computer Assisted Radiology and Surgery 14*, 7 (Jul 2019), 1217–1225.

[75] Gan, C., Naiyan Wang, Yang, Y., Dit-Yan Yeung, and Hauptmann, A. G. Devnet: A deep event network for multimedia event detection and evidence recounting. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015), pp. 2568–2577.

[76] Gao, X., Jin, Y., Dou, Q., and Heng, P. A. Automatic gesture recognition in robot-assisted surgery with reinforcement learning and tree search. In *2020 IEEE International Conference on Robotics and Automation (ICRA)* (2020), pp. 8440–8446.

[77] Gao, Y., Vedula, S., Reiley, C., Ahmidi, N., Varadarajan, B., Lin, H. C., Tao, L., Zappella, L., Béjar, B., Yuh, D., Chen, C. C. G., Vidal, R., Khudanpur, S., and Hager, G. Jhu-isi gesture and skill assessment working set (jigsaws) : A surgical activity dataset for human motion modeling. In *MICCAI Workshop: M2CAI, vol. 3, 2014, p. 3.* (2014).

[78] Ghamsarian, N. Enabling relevance-based exploration of cataract videos. In *Proceedings of the 2020 International Conference on Multimedia Retrieval* (New York, NY, USA, 2020), ICMR '20, Association for Computing Machinery, p. 378–382.

[79] Ghamsarian, N., Amirpourazarian, H., Timmerer, C., Taschwer, M., and Schöffmann, K. Relevance-based compression of cataract surgery videos using convolutional neural networks. In *Proceedings of the 28th ACM International Conference on Multimedia* (New York, NY, USA, 2020), MM '20, Association for Computing Machinery, p. 3577–3585.

[80] Ghamsarian, N., Taschwer, M., Putzgruber-Adamitsch, D., Sarny, S., El-Shabrawi, Y., and Schoeffmann, K. Lensid: A cnn-rnn-based framework towards lens irregularity detection in cataract surgery videos. In

Medical Image Computing and Computer Assisted Intervention – MICCAI 2021 (Cham, 2021), M. de Bruijne, P. C. Cattin, S. Cotin, N. Padoy, S. Speidel, Y. Zheng, and C. Essert, Eds., Springer International Publishing, pp. 76–86.

[81] Ghamsarian, N., Taschwer, M., Putzgruber-Adamitsch, D., Sarny, S., El-Shabrawi, Y., and Schoeffmann, K. Recal-net: Joint region-channel-wise calibrated network for semantic segmentation in cataract surgery videos. In *28th International Conference on Neural Information Processing (ICONIP)* (2021), p. To Appear.

[82] Ghamsarian, N., Taschwer, M., Putzgruber-Adamitsch, D., Sarny, S., and Schoeffmann, K. Relevance detection in cataract surgery videos by spatio-temporal action localization. In *2020 25th International Conference on Pattern Recognition (ICPR)* (2021), pp. 10720–10727.

[83] Ghamsarian, N., Taschwer, M., and Schoeffmann, K. Deblurring cataract surgery videos using a multi-scale deconvolutional neural network. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)* (2020), pp. 872–876.

[84] Ghamsarian, N., Taschwer, M., and Schoeffmann, K. Deeppyram: Enabling pyramid view and deformable pyramid reception for semantic segmentation in cataract surgery videos. p. Under Review.

[85] Glasner, D., Bagon, S., and Irani, M. Super-resolution from a single image. In *2009 IEEE 12th International Conference on Computer Vision* (Sep. 2009), pp. 349–356.

[86] Gokturk, S. B., Tomasi, C., Girod, B., and Beaulieu, C. Medical image compression based on region of interest, with application to colon ct images. In *2001 Conference Proceedings of the 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (2001), vol. 3, pp. 2453–2456 vol.3.

[87] Gong, R., Chen, Y., Paudel, D. P., Li, Y., Chhatkuli, A., Li, W., Dai, D., and Van Gool, L. Cluster, split, fuse, and update: Meta-learning for open compound domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2021), pp. 8344–8354.

[88] González, C., Bravo-Sánchez, L., and Arbelaez, P. Isinet: An instance-based approach for surgical instrument segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020* (Cham, 2020), A. L. Martel, P. Abolmaesumi, D. Stoyanov, D. Mateus, M. A. Zuluaga, S. K. Zhou, D. Racoceanu, and L. Joskowicz, Eds., Springer International Publishing, pp. 595–605.

[89] Grammatikopoulou, M., Flouty, E., Kadkhodamohammadi, A., Quellec, G., Chow, A., Nehme, J., Luengo, I., and Stoyanov, D. Cadis: Cataract dataset for image segmentation, 2020.

[90] Gu, Z., Cheng, J., Fu, H., Zhou, K., Hao, H., Zhao, Y., Zhang, T., Gao, S., and Liu, J. Ce-net: Context encoder network for 2d medical image segmentation. *IEEE Transactions on Medical Imaging 38*, 10 (2019), 2281–2292.

[91] Hadizadeh, H., and Bajić, I. V. Saliency-aware video compression. *IEEE Transactions on Image Processing 23*, 1 (2014), 19–33.

[92] Halevy, A., Norvig, P., and Pereira, F. The unreasonable effectiveness of data. *IEEE Intelligent Systems 24*, 2 (March 2009), 8–12.

[93] Han, D., and Wang, L. Notice of retraction: Dien network: Detailed information extracting network for detecting continuous circular capsulorhexis boundaries of cataracts. *IEEE Access 8* (2020), 161571–161579.

[94] Harley, A. W., Derpanis, K. G., and Kokkinos, I. Segmentation-aware convolutional networks using local attention masks. *CoRR abs/1708.04607* (2017).

[95] He, J., Jia, X., Chen, S., and Liu, J. Multi-source domain adaptation with collaborative learning for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2021), pp. 11008–11017.

[96] He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2020).

[97] He, K., Gkioxari, G., Dollar, P., and Girshick, R. Mask r-cnn. In *The IEEE International Conference on Computer Vision (ICCV)* (Oct 2017).

[98] He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), pp. 770–778.

[99] Henaff, O. Data-efficient image recognition with contrastive predictive coding. In *Proceedings of the 37th International Conference on Machine Learning* (13–18 Jul 2020), H. D. III and A. Singh, Eds., vol. 119 of *Proceedings of Machine Learning Research*, PMLR, pp. 4182–4192.

[100] Hjelm, R. D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., and Bengio, Y. Learning deep representations by mutual information estimation and maximization, 2019.

[101] Hochreiter, S., and Schmidhuber, J. Long Short-Term Memory. *Neural Computation 9*, 8 (1997), 1735–1780.

[102] Holden, M. S., Ungi, T., Sargent, D., McGraw, R. C., Chen, E. C. S., Ganapathy, S., Peters, T. M., and Fichtinger, G. Feasibility

of real-time workflow segmentation for tracked needle interventions. *IEEE Transactions on Biomedical Engineering 61*, 6 (2014), 1720–1728.

[103] Hu, J., Shen, L., and Sun, G. Squeeze-and-excitation networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018), pp. 7132–7141.

[104] Hu, Q., Zhou, J., Zhang, X., Gao, Z., and Sun, M.-T. In-loop perceptual model-based rate-distortion optimization for hevc real-time encoder. *Journal of Real-Time Image Processing* (04 2018), 1–19.

[105] Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), pp. 2261–2269.

[106] Huang, Z., Huang, L., Gong, Y., Huang, C., and Wang, X. Mask scoring R-CNN. *CoRR abs/1903.00241* (2019).

[107] Huaulmé, A., Jannin, P., Reche, F., Faucheron, J.-L., Moreau-Gaudry, A., and Voros, S. Offline identification of surgical deviations in laparoscopic rectopexy. *Artificial Intelligence in Medicine 104* (2020), 101837.

[108] Ibtehaz, N., and Rahman, M. S. Multiresunet : Rethinking the u-net architecture for multimodal biomedical image segmentation. *Neural Networks 121* (2020), 74–87.

[109] Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., and Brox, T. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (July 2017).

[110] Islam, M., Li, Y., and Ren, H. Learning where to look while tracking instruments in robot-assisted surgery. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019* (Cham, 2019), D. Shen, T. Liu, T. M. Peters, L. H. Staib, C. Essert, S. Zhou, P.-T. Yap, and A. Khan, Eds., Springer International Publishing, pp. 412–420.

[111] Ismail Fawaz, H., Forestier, G., Weber, J., Idoumghar, L., and Muller, P.-A. Evaluating surgical skills from kinematic data using convolutional neural networks. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018* (Cham, 2018), A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López, and G. Fichtinger, Eds., Springer International Publishing, pp. 214–221.

[112] Itti, L. Automatic foveation for video compression using a neurobiological model of visual attention. *IEEE Transactions on Image Processing 13*, 10 (2004), 1304–1318.

[113] Jenni, S., Jin, H., and Favaro, P. Steering self-supervised feature learning beyond local pixel statistics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2020).

[114] Jeong, W. J., Park, J. W., Lee, D., Choi, W., and Moon, Y. S. Weighted linear motion deblurring with blur kernel estimation using consecutive frames. In *The 18th IEEE International Symposium on Consumer Electronics (ISCE 2014)* (June 2014), pp. 1–2.

[115] Jiang, J., Hu, Y.-C., Liu, C.-J., Halpenny, D., Hellmann, M. D., Deasy, J. O., Mageras, G., and Veeraraghavan, H. Multiple resolution residually connected feature streams for automatic lung tumor segmentation from ct images. *IEEE Transactions on Medical Imaging 38*, 1 (2019), 134–144.

[116] Jin, A., et al. m2cai16-tool-locations dataset. `http://ai.stanford.edu/~syyeung/tooldetection.html/`, 2018.

[117] Jin, A., Yeung, S., Jopling, J., Krause, J., Azagury, D., Milstein, A., and Fei-Fei, L. Tool detection and operative skill assessment in surgical videos using region-based convolutional neural networks. *CoRR abs/1802.08774* (2018).

[118] Jin, Y., Cheng, K., Dou, Q., and Heng, P.-A. Incorporating temporal prior from motion flow for instrument segmentation in minimally invasive surgery video. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019* (Cham, 2019), D. Shen, T. Liu, T. M. Peters, L. H. Staib, C. Essert, S. Zhou, P.-T. Yap, and A. Khan, Eds., Springer International Publishing, pp. 440–448.

[119] Jin, Y., Dou, Q., Chen, H., Yu, L., Qin, J., Fu, C., and Heng, P. SV-RCNet: Workflow Recognition From Surgical Videos Using Recurrent Convolutional Network. *IEEE Transactions on Medical Imaging 37*, 5 (May 2018), 1114–1126.

[120] Jin, Y., Dou, Q., Yu, L., and Heng, P.-A. Endorcn : Recurrent convolutional networks for recognition of surgical workflow in cholecystectomy procedure video.

[121] Jin, Y., Li, H., Dou, Q., Chen, H., Qin, J., Fu, C.-W., and Heng, P.-A. Multi-task recurrent convolutional network with correlation loss for surgical video analysis. *Medical Image Analysis 59* (2020), 101572.

[122] Joskowicz, L. Computer-aided surgery meets predictive, preventive, and personalized medicine. *EPMA Journal 8*, 1 (Mar 2017), 1–4.

[123] Kannan, S., Yengera, G., Mutter, D., Marescaux, J., and Padoy, N. Future-state predicting lstm for early surgery type recognition. *IEEE Transactions on Medical Imaging 39*, 3 (2020), 556–566.

[124] Kayış, E., Khaniyev, T. T., Suermondt, J., and Sylvester, K. A robust estimation model for surgery durations with temporal, operational, and surgery team effects. *Health Care Management Science 18*, 3 (Sep 2015), 222–233.

[125] Kayis, E., Wang, H., Patel, M., Gonzalez, T., Jain, S., Ramamurthi, R. J., Santos, C., Singhal, S., Suermondt, J., and Sylvester, K. Improving prediction of surgery duration using operational and temporal factors. *AMIA ... Annual Symposium proceedings. AMIA Symposium 2012* (2012), 456–462. 23304316[pmid].

[126] Kim, D., Cho, D., and Kweon, I. S. Self-supervised video representation learning with space-time cubic puzzles. *Proceedings of the AAAI Conference on Artificial Intelligence 33*, 01 (Jul. 2019), 8545–8552.

[127] Kim, M., and Byun, H. Learning texture invariant representation for domain adaptation of semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2020).

[128] Kim, T. H., Nah, S., and Lee, K. M. Dynamic Video Deblurring Using a Locally Adaptive Blur Model. *IEEE Transactions on Pattern Analysis and Machine Intelligence 40*, 10 (Oct 2018), 2374–2387.

[129] Kim, T. S., O'Brien, M., Zafar, S., Hager, G. D., Sikder, S., and Vedula, S. S. Objective assessment of intraoperative technical skill in capsulorhexis using videos of cataract surgery. *International Journal of Computer Assisted Radiology and Surgery 14*, 6 (Jun 2019), 1097–1105.

[130] Kingma, D., and Ba, J. Adam: A Method for Stochastic Optimization. *International Conference on Learning Representations* (12 2014).

[131] Kirillov, A., Levinkov, E., Andres, B., Savchynskyy, B., and Rother, C. Instancecut: from edges to instances with multicut. *CoRR abs/1611.08272* (2016).

[132] Kitaguchi, D., Takeshita, N., Matsuzaki, H., Oda, T., Watanabe, M., Mori, K., Kobayashi, E., and Ito, M. Automated laparoscopic colorectal surgery workflow recognition using artificial intelligence: Experimental research. *International Journal of Surgery 79* (2020), 88 – 94.

[133] Kitaguchi, D., Takeshita, N., Matsuzaki, H., Takano, H., Owada, Y., Enomoto, T., Oda, T., Miura, H., Yamanashi, T., Watanabe, M., Sato, D., Sugomori, Y., Hara, S., and Ito, M. Real-time automatic surgical phase recognition in laparoscopic sigmoidectomy using the convolutional neural network-based deep learning approach. *Surgical Endoscopy 34* (2020), 4924 – 4931.

[134] Kletz, S., Schoeffmann, K., Leibetseder, A., Benois-Pineau, J., and Husslein, H. Instrument recognition in laparoscopy for technical skill assessment. In *MultiMedia Modeling* (Cham, 2020), Y. M. Ro, W.-H. Cheng, J. Kim, W.-T. Chu, P. Cui, J.-W. Choi, M.-C. Hu, and W. De Neve, Eds., Springer International Publishing, pp. 589–600.

[135] Kolesnikov, A., Zhai, X., and Beyer, L. Revisiting self-supervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2019).

[136] Komodakis, N., and Gidaris, S. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations (ICLR)* (Vancouver, Canada, Apr. 2018).

[137] Kugelman, J., Alonso-Caneiro, D., Read, S. A., Vincent, S. J., Chen, F. K., and Collins, M. J. Effect of altered oct image quality on deep learning boundary segmentation. *IEEE Access 8* (2020), 43537–43553.

[138] Kupyn, O., Budzan, V., Mykhailych, M., Mishkin, D., and Matas, J. DeblurGAN: Blind Motion Deblurring Using Conditional Adversarial Networks. *CoRR abs/1711.07064* (2017).

[139] Kurmann, T., Marquez Neila, P., Du, X., Fua, P., Stoyanov, D., Wolf, S., and Sznitman, R. Simultaneous recognition and pose estimation of instruments in minimally invasive surgery. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2017* (Cham, 2017), Springer International Publishing, pp. 505–513.

[140] Ladický, Ľ., Sturgess, P., Alahari, K., Russell, C., and Torr, P. H. S. What, where and how many? combining object detectors and crfs. In *Computer Vision – ECCV 2010* (Berlin, Heidelberg, 2010), K. Daniilidis, P. Maragos, and N. Paragios, Eds., Springer Berlin Heidelberg, pp. 424–437.

[141] Lalys, F., Riffaud, L., Bouget, D., and Jannin, P. A framework for the recognition of high-level surgical tasks from video images for cataract surgeries. *IEEE Transactions on Biomedical Engineering 59*, 4 (April 2012), 966–976.

[142] Lanza, M., Koprowski, R., Boccia, R., Krysik, K., Sbordone, S., Tartaglione, A., Ruggiero, A., and Simonelli, F. Application of artificial intelligence in the analysis of features affecting cataract surgery complications in a teaching hospital. *Frontiers in medicine 7* (Dec 2020), 607870–607870. 33363188[pmid].

[143] Lee, C.-Y., Xie, S., Gallagher, P., Zhang, Z., and Tu, Z. Deeply-Supervised Nets. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics* (San Diego, California, USA, 09–12 May 2015), G. Lebanon and S. V. N. Vishwanathan, Eds., vol. 38 of *Proceedings of Machine Learning Research*, PMLR, pp. 562–570.

[144] Lee, D., Yu, H. W., Kwon, H., Kong, H.-J., Lee, K. E., and Kim, H. C. Evaluation of surgical skills during robotic surgery by deep learning-based multiple surgical instrument tracking in training and actual operations. *Journal of clinical medicine 9*, 6 (Jun 2020), 1964. 32585953[pmid].

[145] Lee, H.-Y., Huang, J.-B., Singh, M., and Yang, M.-H. Unsupervised representation learning by sorting sequences. In *2017 IEEE International Conference on Computer Vision (ICCV)* (2017), pp. 667–676.

[146] Leibetseder, A., and Schoeffmann, K. Surgxplore: Interactive video exploration for endoscopy. In *Proceedings of the 2020 International Conference on Multimedia Retrieval* (New York, NY, USA, 2020), ICMR '20, Association for Computing Machinery, p. 397–401.

[147] Lhuillier, L., Jeancolas, A. L., Renaudin, L., Goetz, C., Ameloot, F., Premy, S., Ouamara, N., and Perone, J. M. Impact of ophthalmic surgeon experience on early postoperative central corneal thickness after cataract surgery. *Cornea 36*, 5 (May 2017), 541–545.

[148] Li, G., and Yu, Y. Visual saliency detection based on multiscale deep CNN features. *CoRR abs/1609.02077* (2016).

[149] Li, S., Xu, M., Deng, X., and Wang, Z. Weight-based r-λ rate control for perceptual hevc coding on conversational videos. *Signal Processing: Image Communication 38* (2015), 127 – 140. Recent Advances in Saliency Models, Applications and Evaluations.

[150] Li, Y., Yuan, L., and Vasconcelos, N. Bidirectional learning for domain adaptation of semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2019).

[151] Li, Z., Qin, S., and Itti, L. Visual attention guided bit allocation in video compression. *Image and Vision Computing 29*, 1 (2011), 1 – 14.

[152] Lin, S., Qin, F., Bly, R. A., Moe, K. S., and Hannaford, B. Automatic sinus surgery skill assessment based on instrument segmentation and tracking in endoscopic video. In *Multiscale Multimodal Medical Imaging* (Cham, 2020), Q. Li, R. Leahy, B. Dong, and X. Li, Eds., Springer International Publishing, pp. 93–100.

[153] Lin, T., Maire, M., Belongie, S. J., Bourdev, L. D., Girshick, R. B., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft COCO: common objects in context. *CoRR abs/1405.0312* (2014).

[154] Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence 42*, 2 (2020), 318–327.

[155] Liu, T., Yuan, Z., Sun, J., Wang, J., Zheng, N., Tang, X., and Shum, H. Learning to detect a salient object. *IEEE Transactions on Pattern Analysis and Machine Intelligence 33*, 2 (2011), 353–367.

[156] Liu, Y., Li, Z. G., and Soh, Y. C. Region-of-interest based resource allocation for conversational video communication of h.264/avc. *IEEE Transactions on Circuits and Systems for Video Technology 18*, 1 (2008), 134–139.

[157] Lokoč, J., Veselý, P., Mejzlík, F., Kovalčík, G., Souček, T., Rossetto, L., Schoeffmann, K., Bailer, W., Gurrin, C., Sauter, L., Song, J., Vrochidis, S., Wu, J., and Jónsson, B. t. Is the reign of interactive search eternal? findings from the video browser showdown 2020. *ACM Trans. Multimedia Comput. Commun. Appl. 17*, 3 (July 2021).

[158] Long, J., Shelhamer, E., and Darrell, T. Fully convolutional networks for semantic segmentation. *CoRR abs/1411.4038* (2014).

[159] Loukas, C. Video content analysis of surgical procedures. *Surgical Endoscopy 32*, 2 (Feb 2018), 553–568.

[160] Loukas, C. G. Surgical Phase Recognition of Short Video Shots Based on Temporal Modeling of Deep Features. *CoRR abs/1807.07853* (2018).

[161] Low, S. A., Braga-Mele, R., Yan, D. B., and El-Defrawy, S. Intraoperative complication rates in cataract surgery performed by ophthalmology resident trainees compared to staff surgeons in a canadian academic center. *Journal of Cataract & Refractive Surgery 44*, 11 (2018), 1344 – 1349.

[162] Maier-Hein, L., Vedula, S., Speidel, S., Navab, N., Kikinis, R., Park, A., Eisenmann, M., Feussner, H., Forestier, G., Giannarou, S., Hashizume, M., Katic, D., Kenngott, H., Kranzfelder, M., Malpani, A., März, K., Neumuth, T., Padoy, N., Pugh, C., Schoch, N., Stoyanov, D., Taylor, R., Wagner, M., Hager, G., and Jannin, P. Surgical data science for next-generation interventions. *Nature Biomedical Engineering 1*, 9 (9 2017), 691–696.

[163] Maktabi, M., and Neumuth, T. Online time and resource management based on surgical workflow time series analysis. *International Journal of Computer Assisted Radiology and Surgery 12*, 2 (Feb 2017), 325–338.

[164] Marafioti, A., Hayoz, M., Gallardo, M., Neila, P. M., Wolf, S., Zinkernagel, M., and Sznitman, R. Catanet: Predicting remaining cataract surgery duration, 2021.

[165] Mayer-Xanthaki, C. F., Pregartner, G., Hirnschall, N., Falb, T., Sommer, M., Findl, O., and Wedrich, A. Impact of intraocular lens characteristics on intraocular lens dislocation after cataract surgery. *British Journal of Ophthalmology* (2020).

[166] Michaeli, T., and Irani, M. Blind Deblurring Using Internal Patch Recurrence. In *Computer Vision – ECCV 2014* (Cham, 2014), D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds., Springer International Publishing, pp. 783–798.

[167] Misra, I., and Maaten, L. v. d. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2020).

[168] Misra, I., Zitnick, C. L., and Hebert, M. Shuffle and learn: Unsupervised learning using temporal order verification. In *Computer Vision – ECCV 2016* (Cham, 2016), B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., Springer International Publishing, pp. 527–544.

[169] Moglia, A. Automated, objective and predictive evaluation of technical skills in robot-assisted surgery. *Journal of Robotic Surgery 13*, 1 (Feb 2019), 189–190.

[170] Morita, S., Tabuchi, H., Masumoto, H., Tanabe, H., and Kamiura, N. Real-time surgical problem detection and instrument tracking in cataract surgery. *Journal of Clinical Medicine 9*, 12 (2020).

[171] Morita, S., Tabuchi, H., Masumoto, H., Yamauchi, T., and Kamiura, N. Real-time extraction of important surgical phases in cataract surgery videos. *Scientific Reports 9* (12 2019).

[172] Münzer, B., Schoeffmann, K., Böszörmenyi, L., Smulders, J. F., and Jakimowicz, J. J. Investigation of the impact of compression on the perceptional quality of laparoscopic videos. In *2014 IEEE 27th International Symposium on Computer-Based Medical Systems* (2014), pp. 153–158.

[173] NeginGhamsarian, Taschwer, M., and Schoeffmann, K. Deblurring cataract surgery videos using a multi-scale deconvolutional neural network. *CoRR abs/1504.06852* (2020).

[174] Newell, A., and Deng, J. Associative embedding: End-to-end learning for joint detection and grouping. *CoRR abs/1611.05424* (2016).

[175] Ni, Z.-L., Bian, G.-B., Wang, G.-A., Zhou, X.-H., Hou, Z.-G., Chen, H.-B., and Xie, X.-L. Pyramid attention aggregation network for semantic segmentation of surgical instruments. *Proceedings of the AAAI Conference on Artificial Intelligence 34*, 07 (Apr. 2020), 11782–11790.

[176] Ni, Z.-L., Bian, G.-B., Wang, G.-A., Zhou, X.-H., Hou, Z.-G., Xie, X.-L., Li, Z., and Wang, Y.-H. Barnet: Bilinear attention network with adaptive receptive fields for surgical instrument segmentation. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20* (7 2020), C. Bessiere, Ed., International Joint Conferences on Artificial Intelligence Organization, pp. 832–838. Main track.

[177] Ni, Z.-L., Bian, G.-B., Zhou, X.-H., Hou, Z.-G., Xie, X.-L., Wang, C., Zhou, Y.-J., Li, R.-Q., and Li, Z. Raunet: Residual attention u-net for semantic segmentation of cataract surgical instruments. In *Neural Information Processing* (Cham, 2019), T. Gedeon, K. W. Wong, and M. Lee, Eds., Springer International Publishing, pp. 139–149.

[178] Nishio, S., Hossain, B., Yagi, N., Nii, M., Hiranaka, T., and Kobashi, S. Surgical phase recognition method with a sequential consistency for caos-ai navigation system. In *2020 IEEE 2nd Global Conference on Life Sciences and Technologies (LifeTech)* (2020), pp. 8–10.

[179] Noroozi, M., Chandramouli, P., and Favaro, P. Motion Deblurring in the Wild. In *Pattern Recognition* (Cham, 2017), V. Roth and T. Vetter, Eds., Springer International Publishing, pp. 65–77.

[180] Noroozi, M., and Favaro, P. Unsupervised learning of visual representations by solving jigsaw puzzles. *CoRR abs/1603.09246* (2016).

[181] Padoy, N., Blum, T., Ahmadi, S.-A., Feussner, H., Berger, M.-O., and Navab, N. Statistical modeling and recognition of surgical workflow. *Medical Image Analysis 16*, 3 (2012), 632–641. Computer Assisted Interventions.

[182] Pakhomov, D., Premachandran, V., Allan, M., Azizian, M., and Navab, N. Deep residual learning for instrument segmentation in robotic surgery. In *Machine Learning in Medical Imaging* (Cham, 2019), H.-I. Suk, M. Liu, P. Yan, and C. Lian, Eds., Springer International Publishing, pp. 566–573.

[183] Pakhomov, D., Shen, W., and Navab, N. Towards unsupervised learning for instrument segmentation in robotic surgery with cycle-consistent adversarial networks, 2020.

[184] Pathak, D., Girshick, R., Dollar, P., Darrell, T., and Hariharan, B. Learning features by watching objects move. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (July 2017).

[185] Pathak, D., Girshick, R., Dollar, P., Darrell, T., and Hariharan, B. Learning features by watching objects move. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (July 2017).

[186] Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., and Efros, A. A. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2016).

[187] Pereira, S., Pinto, A., Amorim, J., Ribeiro, A., Alves, V., and Silva, C. A. Adaptive feature recombination and recalibration for semantic segmentation with fully convolutional networks. *IEEE Transactions on Medical Imaging 38*, 12 (2019), 2914–2925.

[188] Petscharnig, S., and Schöffmann, K. Learning laparoscopic video shot classification for gynecological surgery. *Multimedia Tools and Applications 77*, 7 (Apr 2018), 8061–8079.

[189] Pohlen, T., Hermans, A., Mathias, M., and Leibe, B. Full-resolution residual networks for semantic segmentation in street scenes. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), pp. 3309–3318.

[190] Poyser, M., Atapour-Abarghouei, A., and Breckon, T. P. On the impact of lossy image and video compression on the performance of deep convolutional neural network architectures. In *2020 25th International Conference on Pattern Recognition (ICPR)* (2021), pp. 2830–2837.

[191] Qayyum, A., Ahmad, I., Mumtaz, W., Alassafi, M. O., Alghamdi, R., and Mazher, M. Automatic segmentation using a hybrid dense network integrated with an 3d-atrous spatial pyramid pooling module for computed tomography (ct) imaging. *IEEE Access 8* (2020), 169794–169803.

[192] Qin, Y., Feyzabadi, S., Allan, M., Burdick, J. W., and Azizian, M. davincinet: Joint prediction of motion and surgical state in robot-assisted surgery, 2020.

[193] Qin, Y., Kamnitsas, K., Ancha, S., Nanavati, J., Cottrell, G., Criminisi, A., and Nori, A. Autofocus layer for semantic segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018* (Cham, 2018), A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López, and G. Fichtinger, Eds., Springer International Publishing, pp. 603–611.

[194] Qin, Y., Pedram, S. A., Feyzabadi, S., Allan, M., McLeod, A. J., Burdick, J. W., and Azizian, M. Temporal segmentation of surgical sub-tasks through deep learning with multiple data sources. In *2020 IEEE International Conference on Robotics and Automation (ICRA)* (2020), pp. 371–377.

[195] Qiu, H., Li, Z., Yang, Y., Xin, C., and Bian, G.-B. Real-time iris tracking using deep regression networks for robotic ophthalmic surgery. *IEEE Access 8* (2020), 50648–50658.

[196] Quellec, G., Lamard, M., Cochener, B., and Cazuguel, G. Real-time segmentation and recognition of surgical tasks in cataract surgery videos. *IEEE Transactions on Medical Imaging 33* (12 2014), 2352–60.

[197] Rabbani, M. JPEG2000: Image Compression Fundamentals, Standards and Practice. *Journal of Electronic Imaging 11*, 2 (2002).

[198] Radford, A., Metz, L., and Chintala, S. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. *CoRR abs/1511.06434* (2015).

[199] Ramakrishnan, S., Pachori, S., Gangopadhyay, A., and Raman, S. Deep Generative Filter for Motion Deblurring. *CoRR abs/1709.03481* (2017).

[200] Redmon, J., and Farhadi, A. Yolo9000: Better, faster, stronger. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), pp. 6517–6525.

[201] Ren, S., He, K., Girshick, R., and Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence 39*, 6 (2017), 1137–1149.

[202] Ren, S., He, K., Girshick, R. B., and Sun, J. Faster R-CNN: towards real-time object detection with region proposal networks. *CoRR abs/1506.01497* (2015).

[203] Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* (Cham, 2015), N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds., Springer International Publishing, pp. 234–241.

[204] Rossetto, L., Gasser, R., Lokoč, J., Bailer, W., Schoeffmann, K., Muenzer, B., Souček, T., Nguyen, P. A., Bolettieri, P., Leibetseder, A., and Vrochidis, S. Interactive video retrieval in the age of deep learning – detailed evaluation of vbs 2019. *IEEE Transactions on Multimedia 23* (2021), 243–256.

[205] Roy, A. G., Navab, N., and Wachinger, C. Recalibrating fully convolutional networks with spatial and channel "squeeze and excitation" blocks. *IEEE Transactions on Medical Imaging 38*, 2 (2019), 540–549.

[206] S, P. T., and Fleuret, F. Uncertainty reduction for model adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2021), pp. 9613–9623.

[207] Sarikaya, D., Corso, J. J., and Guru, K. A. Detection and localization of robotic tools in robot-assisted surgery videos using deep neural networks for region proposal and detection. *IEEE Transactions on Medical Imaging 36*, 7 (2017), 1542–1549.

[208] Sarikaya, D., et al. Atlas dione dataset. `https://www.roswellpark.org/education/professional-training/atlas-program/research-development/dione-dataset/`, 2017.

[209] Schoeffmann, K., Taschwer, M., Sarny, S., Münzer, B., Primus, M. J., and Putzgruber, D. Cataract-101: Video Dataset of 101 Cataract Surgeries. In *Proceedings of the 9th ACM Multimedia Systems Conference* (New York, NY, USA, 2018), MMSys '18, ACM, pp. 421–425.

[210] Schuler, C. J., Burger, H. C., Harmeling, S., and Schölkopf, B. A Machine Learning Approach for Non-blind Image Deconvolution. In *2013 IEEE Conference on Computer Vision and Pattern Recognition* (June 2013), pp. 1067–1074.

[211] Sermanet, P., Lynch, C., Chebotar, Y., Hsu, J., Jang, E., Schaal, S., Levine, S., and Brain, G. Time-contrastive networks: Self-supervised learning from video. In *2018 IEEE International Conference on Robotics and Automation (ICRA)* (2018), pp. 1134–1141.

[212] Shi, W., Caballero, J., Huszar, F., Totz, J., Aitken, A. P., Bishop, R., Rueckert, D., and Wang, Z. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2016).

[213] Shotton, J., Johnson, M., and Cipolla, R. Semantic texton forests for image categorization and segmentation. In *2008 IEEE Conference on Computer Vision and Pattern Recognition* (2008), pp. 1–8.

[214] Shvets, A. A., Rakhlin, A., Kalinin, A. A., and Iglovikov, V. I. Automatic instrument segmentation in robot-assisted surgery using deep learning. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)* (2018), pp. 624–628.

[215] Sokolova, N., Schoeffmann, K., Taschwer, M., Putzgruber-Adamitsch, D., and El-Shabrawi, Y. Evaluating the generalization performance of instrument classification in cataract surgery videos. In *MultiMedia Modeling* (Cham, 2020), Y. M. Ro, W.-H. Cheng, J. Kim, W.-T. Chu, P. Cui, J.-W. Choi, M.-C. Hu, and W. De Neve, Eds., Springer International Publishing, pp. 626–636.

[216] Sokolova, N., Taschwer, M., Sarny, S., Putzgruber-Adamitsch, D., and Schoeffmann, K. Pixel-based iris and pupil segmentation in cataract surgery videos using mask r-cnn. In *2020 IEEE 17th International Symposium on Biomedical Imaging Workshops (ISBI Workshops)* (2020), pp. 1–4.

[217] Stauder, R., Okur, A., Peter, L., Schneider, A., Kranzfelder, M., Feussner, H., and Navab, N. Random forests for phase detection in surgical workflow analysis. In *Information Processing in Computer-Assisted Interventions* (Cham, 2014), D. Stoyanov, D. L. Collins, I. Sakuma, P. Abolmaesumi, and P. Jannin, Eds., Springer International Publishing, pp. 148–157.

[218] Stauder, R., Ostler, D., Kranzfelder, M., Koller, S., Feussner, H., and Navab, N. The TUM lapchole dataset for the M2CAI 2016 workflow challenge. *CoRR abs/1610.09278* (2016).

[219] Su, S., Delbracio, M., Wang, J., Sapiro, G., Heidrich, W., and Wang, O. Deep Video Deblurring for Hand-Held Cameras. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (July 2017), pp. 237–246.

[220] Sudhakaran, S., Escalera, S., and Lanz, O. Lsta: Long short-term attention for egocentric action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2019).

[221] Sudhakaran, S., Escalera, S., and Lanz, O. Gate-shift networks for video action recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), pp. 1099–1108.

[222] Sullivan, G. J., Ohm, J., Han, W., and Wiegand, T. Overview of the high efficiency video coding (hevc) standard. *IEEE Transactions on Circuits and Systems for Video Technology 22*, 12 (2012), 1649–1668.

[223] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015), pp. 1–9.

[224] Sznitman, R., Ali, K., Richa, R., Taylor, R. H., Hager, G. D., and Fua, P. Data-driven visual tracking in retinal microsurgery. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2012* (Berlin, Heidelberg, 2012), N. Ayache, H. Delingette, P. Golland, and K. Mori, Eds., Springer Berlin Heidelberg, pp. 568–575.

[225] T. Zemcik, L. Kratochvila, S. B. O. B. P. Z. K. H. Performance evaluation of cnn based pedestrian and cyclist detectors on degraded images. *International Journal of Image Processing (IJIP) 15* (2021).

[226] Tajbakhsh, N., Hu, Y., Cao, J., Yan, X., Xiao, Y., Lu, Y., Liang, J., Terzopoulos, D., and Ding, X. Surrogate supervision for medical image analysis: Effective deep learning from limited quantities of labeled data. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)* (2019), pp. 1251–1255.

[227] Tao, L., Elhamifar, E., Khudanpur, S., Hager, G. D., and Vidal, R. Sparse hidden markov models for surgical gesture classification and skill evaluation. In *Information Processing in Computer-Assisted Interventions* (Berlin, Heidelberg, 2012), P. Abolmaesumi, L. Joskowicz, N. Navab, and P. Jannin, Eds., Springer Berlin Heidelberg, pp. 167–177.

[228] Tian, Y., Krishnan, D., and Isola, P. Contrastive multiview coding. In *Computer Vision – ECCV 2020* (Cham, 2020), A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds., Springer International Publishing, pp. 776–794.

[229] Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., and Paluri, M. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2018).

[230] Travis, E., Woodhouse, S., Tan, R., Patel, S., Donovan, J., and Brogan, K. Operating theatre time, where does it all go? a prospective observational study. *BMJ 349* (2014).

[231] Trikha, S., Turnbull, A., Morris, R., Anderson, D., and Hossain, P. The journey to femtosecond laser-assisted cataract surgery: New beginnings or a false dawn? *Eye (London, England) 27* (02 2013).

[232] Twinanda, A. P., Shehata, S., Mutter, D., Marescaux, J., de Mathelin, M., and Padoy, N. EndoNet: A Deep Architecture for Recognition Tasks on Laparoscopic Videos. *IEEE Transactions on Medical Imaging 36*, 1 (Jan 2017), 86–97.

[233] Twinanda, A. P., Yengera, G., Mutter, D., Marescaux, J., and Padoy, N. RSDNet: Learning to Predict Remaining Surgery Duration from Laparoscopic Videos Without Manual Annotations. *CoRR abs/1802.03243* (2018).

[234] Vasiljevic, I., Chakrabarti, A., and Shakhnarovich, G. Examining the impact of blur on recognition by convolutional networks. *CoRR abs/1611.05760* (2016).

[235] Vasiljevic, I., Chakrabarti, A., and Shakhnarovich, G. Examining the Impact of Blur on Recognition by Convolutional Networks. *CoRR abs/1611.05760* (2016).

[236] Vedula, S. S., Ishii, M., and Hager, G. D. Objective assessment of surgical technical skill and competency in the operating room. *Annual review of biomedical engineering 19* (Jun 2017), 301–325. 28375649[pmid].

[237] Wang, H., Wang, Z., Du, M., Yang, F., Zhang, Z., Ding, S., Mardziel, P., and Hu, X. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (2020), pp. 111–119.

[238] Wang, H., Wu, X., Huang, Z., and Xing, E. P. High-frequency component helps explain the generalization of convolutional neural networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2020).

[239] Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., and Van Gool, L. Temporal segment networks for action recognition in videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence 41*, 11 (2019), 2740–2755.

[240] Wang, W., Yan, W., Müller, A., and He, M. A Global View on Output and Outcomes of Cataract Surgery With National Indices of Socioeconomic Development. *Investigative Ophthalmology & Visual Science 58*, 9 (07 2017), 3669–3676.

[241] Wang, X., and Gupta, A. Unsupervised learning of visual representations using videos. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (December 2015).

[242] Wang, X., Wang, L., Zhong, X., Bai, C., Huang, X., Zhao, R., and Xia, M. Pai-net: A modified u-net of reducing semantic gap for surgical instrument segmentation. *IET Image Processing n/a*, n/a.

[243] Wang, Z., and Majewicz Fey, A. Deep learning with convolutional neural network for objective skill evaluation in robot-assisted surgery. *International Journal of Computer Assisted Radiology and Surgery 13*, 12 (Dec 2018), 1959–1970.

[244] Wei, D., Lim, J., Zisserman, A., and Freeman, W. T. Learning and using the arrow of time. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018), pp. 8052–8060.

[245] Wiegmann, D. A., ElBardissi, A. W., Dearani, J. A., Daly, R. C., and Sundt, T. M. Disruptions in surgical flow and their relationship to surgical errors: An exploratory investigation. *Surgery 142*, 5 (2007), 658–665.

[246] Wojke, N., Bewley, A., and Paulus, D. Simple online and realtime tracking with a deep association metric. In *2017 IEEE International Conference on Image Processing (ICIP)* (2017), IEEE, pp. 3645–3649.

[247] Wu, Z., Xiong, Y., Yu, S. X., and Lin, D. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2018).

[248] Xu, D., Xiao, J., Zhao, Z., Shao, J., Xie, D., and Zhuang, Y. Self-supervised spatiotemporal learning via video clip order prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2019).

[249] Xu, M., Deng, X., Li, S., and Wang, Z. Region-of-interest based conversational hevc coding with hierarchical perception model of face. *IEEE Journal of Selected Topics in Signal Processing 8*, 3 (2014), 475–489.

[250] Yang, J., Zhu, J., Wang, H., and Yang, X. Dilated multiresunet: Dilated multiresidual blocks network based on u-net for biomedical image segmentation. *Biomedical Signal Processing and Control 68* (2021), 102643.

[251] Yengera, G., Mutter, D., Marescaux, J., and Padoy, N. Less is More: Surgical Phase Recognition with Less Annotations through Self-Supervised Pre-training of CNN-LSTM Networks. *CoRR abs/1805.08569* (2018).

[252] Yi, F., and Jiang, T. Hard frame detection and online mapping for surgical phase recognition. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019* (Cham, 2019), D. Shen, T. Liu, T. M. Peters, L. H. Staib, C. Essert, S. Zhou, P.-T. Yap, and A. Khan, Eds., Springer International Publishing, pp. 449–457.

[253] Yoon, J., Lee, J., Park, S., Hyung, W. J., and Choi, M.-K. Semi-supervised learning for instrument detection with a class imbalanced dataset. In *Interpretable and Annotation-Efficient Learning for Medical Image Computing* (Cham, 2020), J. Cardoso, H. Van Nguyen, N. Heller, P. Henriques Abreu, I. Isgum, W. Silva, R. Cruz, J. Pereira Amorim, V. Patel, B. Roysam, K. Zhou,

S. Jiang, N. Le, K. Luu, R. Sznitman, V. Cheplygina, D. Mateus, E. Trucco, and S. Abbasi, Eds., Springer International Publishing, pp. 266–276.

[254] Yu, F., and Koltun, V. Multi-scale context aggregation by dilated convolutions. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings* (2016), Y. Bengio and Y. LeCun, Eds.

[255] Yu, F., Silva Croso, G., Kim, T. S., Song, Z., Parker, F., Hager, G. D., Reiter, A., Vedula, S. S., Ali, H., and 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 (04 2019), e191860–e191860.

[256] Yu, T., Mutter, D., Marescaux, J., and Padoy, N. Learning from a tiny dataset of manual annotations: a teacher/student approach for surgical phase recognition. *CoRR abs/1812.00033* (2018).

[257] Yuniartha, D. R., Masruroh, N. A., and Herliansyah, M. K. An evaluation of a simple model for predicting surgery duration using a set of surgical procedure parameters. *Informatics in Medicine Unlocked 25* (2021), 100633.

[258] Zafar, S., Vedula, S., and Sikder, S. Objective assessment of technical skill targeted to time in cataract surgery. *Journal of Cataract & Refractive Surgery 46*, 5 (2020).

[259] Zang, D., Bian, G.-B., Wang, Y., and Li, Z. An extremely fast and precise convolutional neural network for recognition and localization of cataract surgical tools. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019* (Cham, 2019), D. Shen, T. Liu, T. M. Peters, L. H. Staib, C. Essert, S. Zhou, P.-T. Yap, and A. Khan, Eds., Springer International Publishing, pp. 56–64.

[260] Zappella, L., Béjar, B., Hager, G., and Vidal, R. Surgical gesture classification from video and kinematic data. *Medical Image Analysis 17*, 7 (2013), 732 – 745. Special Issue on the 2012 Conference on Medical Image Computing and Computer Assisted Intervention.

[261] Zeiler, M. D., and Fergus, R. Visualizing and Understanding Convolutional Networks. *CoRR abs/1311.2901* (2013).

[262] Zhai, Y., Zhang, G., Zheng, L., Yang, G., Zhao, K., Gong, Y., Zhang, Z., Zhang, X., Sun, B., and Wang, Z. Computer-aided intraoperative toric intraocular lens positioning and alignment during cataract surgery. *IEEE Journal of Biomedical and Health Informatics* (2021), 1–1.

[263] Zhang, K., Luo, W., Zhong, Y., Ma, L., Liu, W., and Li, H. Adversarial Spatio-Temporal Learning for Video Deblurring. *IEEE Transactions on Image Processing 28*, 1 (Jan 2019), 291–301.

[264] Zhang, M., Li, X., Xu, M., and Li, Q. Automated semantic segmentation of red blood cells for sickle cell disease. *IEEE Journal of Biomedical and Health Informatics 24*, 11 (2020), 3095–3102.

[265] Zhang, P., Wang, F., and Zheng, Y. Self supervised deep representation learning for fine-grained body part recognition. In *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)* (2017), pp. 578–582.

[266] Zhang, R., Isola, P., and Efros, A. A. Colorful image colorization. In *Computer Vision – ECCV 2016* (Cham, 2016), B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., Springer International Publishing, pp. 649–666.

[267] Zhang, R., Isola, P., and Efros, A. A. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (July 2017).

[268] Zhao, H., Shi, J., Qi, X., Wang, X., and Jia, J. Pyramid scene parsing network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (July 2017).

[269] Zhao, Z., Cai, T., Chang, F., and Cheng, X. Real-time surgical instrument detection in robot-assisted surgery using a convolutional neural network cascade. *Healthcare Technology Letters 6*, 6 (2019), 275–279.

[270] Zhao, Z., Voros, S., Weng, Y., Chang, F., and Li, R. Tracking-by-detection of surgical instruments in minimally invasive surgery via the convolutional neural network deep learning-based method. *Computer Assisted Surgery 22*, sup1 (2017), 26–35. PMID: 28937281.

[271] Zhou, Z., He, Z., and Jia, Y. Afpnet: A 3d fully convolutional neural network with atrous-convolution feature pyramid for brain tumor segmentation via mri images. *Neurocomputing 402* (2020), 235–244.

[272] Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N., and Liang, J. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Transactions on Medical Imaging 39*, 6 (2020), 1856–1867.

[273] Zhu, J., Li, Y., Hu, Y., Ma, K., Zhou, S. K., and Zheng, Y. Rubik's cube+: A self-supervised feature learning framework for 3d medical image analysis. *Medical Image Analysis 64* (2020), 101746.

[274] Zhu, Q., Du, B., Turkbey, B., Choyke, P. L., and Yan, P. Deeply-supervised cnn for prostate segmentation. In *2017 International Joint Conference on Neural Networks (IJCNN)* (2017), pp. 178–184.

[275] Zhu, Y., and Song, L. Natural scene text image compression using jpeg2000 roi coding. In *Pattern Recognition* (Berlin, Heidelberg, 2014), S. Li, C. Liu, and Y. Wang, Eds., Springer Berlin Heidelberg, pp. 481–490.

[276] Zhu, Z., Mittendorf, A., Shropshire, E., Allen, B., Miller, C., Bashir, M. R., and Mazurowski, M. A. 3d pyramid pooling network for abdominal mri series classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020), 1–1.

[277] Zhuang, X., Li, Y., Hu, Y., Ma, K., Yang, Y., and Zheng, Y. Self-supervised feature learning for 3d medical images by playing a rubik's cube. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019* (Cham, 2019), D. Shen, T. Liu, T. M. Peters, L. H. Staib, C. Essert, S. Zhou, P.-T. Yap, and A. Khan, Eds., Springer International Publishing, pp. 420–428.

[278] Zia, A., and Essa, I. Automated surgical skill assessment in rmis training. *International Journal of Computer Assisted Radiology and Surgery 13*, 5 (May 2018), 731–739.

[279] Zia, A., Sharma, Y., Bettadapura, V., Sarin, E. L., and Essa, I. Video and accelerometer-based motion analysis for automated surgical skills assessment. *International Journal of Computer Assisted Radiology and Surgery 13*, 3 (Mar 2018), 443–455.

[280] Zisimopoulos, O., Flouty, E., Luengo, I., Giataganas, P., Nehme, J., Chow, A., and Stoyanov, D. DeepPhase: Surgical Phase Recognition in CATARACTS Videos. *CoRR abs/1807.10565* (2018).