공학석사 학위논문

의료비디오의 인식, 인페인팅, 정량화를 위한 딥러닝 연구 개발 및 유양돌기 절제술과 대장내시경의 적용

Deep Learning Research and Development for Recognition, Inpainting, and Quantification in Medical Videos, and Its Application in Mastoidectomy and Colonoscopy

울산대학교대학원

의과학과

박강길

의료비디오의 인식, 인페인팅, 정량화를 위한 딥러닝 연구 개발 및 유양돌기 절제술과 대장내시경의 적용

지도교수 김남국

이 논문을 공학석사 학위 논문으로 제출함

2024년 02월

울산대학교대학원

의과학과

박강길

박강길의 공학석사학위 논문을 인준함

심사위원	정 종 우	(인)
심사위원	김 남 국	(인)
심사위원	이 준 구	(인)

울산대학교대학원

2024 년 02 월

Abstract

This study explores the deep learning technologies for medical video analysis. Applying these technologies in the medical field poses challenges, including difficulties in data collection and the unique nature of medical video. This paper addresses these challenges by applying advanced deep learning techniques such as Knowledge Distillation, Implicit Neural Representation, and Deep Metric Learning to the analysis of medical video data. We particularly focus on Surgical Phase Recognition and Video Inpainting techniques for tympanomastoidectomy (TM), and methods for quantifying colonoscopy video data.

In this research, three experiments were conducted to analyze medical videos. The first study proposed a TM surgical phase recognition learning method using Teacher-Student learning, showing improved performance in situations of class imbalance and data scarcity. The second study introduced a method using implicit neural representation for high-resolution video inpainting without the need for large-scale video data collection, demonstrating enhanced visual performance compared to state-of-the-art model. The third study proposed a method integrating quality-aware metric learning with a noise-robust phase recognition model, considering the characteristics of the colon, leading to more accurate quantification in colonoscopy.

We believe that the methodologies developed in this study have the potential to be applied not only to TM surgery and colonoscopy videos but also to a broader range of medical video analysis areas.

Contents

Contents of Tables

Contents of Figures

Introduction

Background

In recent years, the rapid advancement of deep learning-based video analysis technology has made a significant impact across various industries. A multitude of studies in the field of natural video analysis have showcased remarkable improvements in performance through the application of deep learning techniques [1, 2]. However, applying these advancements to the medical domain presents some challenges. The first major challenge is data collection. Deep learning relies heavily on data to learn and extract necessary information; hence, achieving high performance necessitates a substantial amount of data. Nevertheless, the collection of medical data poses considerable challenges, particularly due to patient privacy concerns. Moreover, medical data typically requires the input and knowledge of experts, making it a high-cost endeavor. A second challenge arises from the fundamental differences between natural and medical images. The information contained within medical videos is vastly different from that in natural videos, often rendering methods effective in natural video analysis less suitable for medical applications. This discrepancy necessitates a more delicate and tailored approach for medical video analysis.

This paper seeks to address these challenges by employing advanced deep learning technologies such as *Knowledge Distillation*, *Implicit Neural Representation*, and *Deep Metric Learning* in the analysis of medical video data. In particularly, we focus on the application of Surgical Phase Recognition and Video Inpainting techniques for tympanomastoidectomy [3], an otologic surgery. Additionally, we discuss methods for quantifying colonoscopy video data.

Surgical Phase Recognition

In the realm of medical video analysis, surgical phase recognition has emerged as a highly specialized and effective application of video recognition technique. Early studies on surgical phase recognition primarily relied on features such as pixel intensity and gradient changes [4], and spatial-temporal characteristics [5], in addition to color, texture, and shape elements [6]. Recently, with the development of deep learning technology, research on surgical phase recognition using neural networks has been active. Numerous studies have adopted a structure where Convolutional Neural Networks (CNN) [7, 8] are utilized for the extraction of spatial information, complemented by the use of temporal relation models like Long Short-Term Memory (LSTM) [9], Temporal Convolutional Network (TCN) [10], Multi-Stage Temporal Convolutional Network (MS-TCN) [11], Transformer [12] for effectively integrating both spatial and temporal data, thereby enhancing the accuracy of surgical phase recognition. Jin *et al.* proposed SV-RCNet, a unified framework combining ResNet and LSTM modules for effective spatial-temporal features learning [13]. Czempiel *et al.* introduced MS-TCN, a novel Multi-Stage Temporal Convolutional Network, marking its first application in surgical phase recognition [14]. Jin *et al.* developed TMRNet, designed to link long-range and multi-scale temporal patterns by utilizing memory bank and LSTM [15]. Yi et al. proposed a novel approach in surgical phase recognition by exploring different multi-stage architectures with MS-TCN [16]. Ding *et al.* introduced SAHC, a segment-attentive hierarchical consistency network, focused on learning segment-level semantics by integrating MS-TCN and Transformer [17].

Knowledge Distillation

Knowledge distillation (KD) is a method developed for model lightweighting, aimed at creating smaller models that mimic larger ones [18]. In this approach, the larger model is referred to as the teacher, and the smaller model as the student. The process involves initially pre-training the teacher network, and then training the student using the final prediction values of the trained teacher network as targets. This teacher-student architecture is not only beneficial for the propose of model lightweighting but has also shown impressive performance in various fields including self-supervised [19, 20, 21] and semi-supervised learning [22, 23, 24]. Recently, this structure has been adopted in the field of surgical video, indicating its broad applicability and effectiveness. Yu *et al.* propose a teacher/student approach for enhancing surgical phase recognition with limited annotated data [25]. Teevno *et al.* introduced a semi-supervised learning framework for surgical tool detection, utilizing a teacher-student model to effectively manage data scarcity and imbalance [26]. Yamlahi *et al.* proposed the use of self-distillation in surgical scene understanding, introducing a heterogeneous ensemble of three models [27].

Video Inpainting

Video inpainting is a video processing technique that fills in missing or damaged areas in each frame of a video with plausible visual content. It is a technique that can be used in various ways, such as removing unnecessary moving objects, subtitles, or watermarks from videos. Recent advancements in deep learning technologies have led to the release of high-performance video inpainting models. Chang *et al.* introduced a deep learning-based free-form video inpainting model, incorporating 3D gated convolutions and a Temporal PatchGAN loss [28]. Zeng *et al.* proposed Spatial-Temporal Transformer Network (STTN) for video inpainting, addressing spatial and temporal inconsistencies [29]. Li *et al.* introduced E2FGVI, an end-to-end flow-guided video inpainting framework that unifies flow completion, feature propagation, and content hallucination into a single, efficient process [30].

Implicit Neural Representation

Implicit neural representation (INR) is a method of parameterizing various signals using a neural network. INR, with its continuous and memory-efficient properties, has applications in a wide range of fields such as video analysis and 3D rendering [31]. Chen *et al.* introduced a novel Video Implicit Neural Representation (VideoINR) for Space-Time Video Super-Resolution (STVSR), which enables video decoding at arbitrary spatial resolutions and frame rates, thus providing enhanced flexibility over fixed up-sampling scales [32]. Kim *et al.* developed a novel neural video representation (NVP) with learnable positional features, significantly enhancing video encoding efficiency and quality while reducing training time and parameter count [33]. Attal *et al.* introduced HyperReel, an advanced 6-DoF video representation enabling high-quality, real-time rendering at high resolutions with improved visual fidelity and memory efficiency, without relying on specialized CUDA code [34]. Park *et al.* proposed DeepSDF, a groundbreaking approach for 3D shape representation using a learned Signed Distance Function, enabling enhanced shape reconstruction and interpolation with a notably smaller model size [35]. Mildenhall *et al.* introduces a novel method for creating photorealistic views of complex scenes from limited input, utilizing a deep network-based volumetric scene function and differentiable volume rendering [36].

Deep Metric Learning

Deep metric learning, a methodology that employs deep learning models to learn embedding functions, effectively quantifies data similarity. Recently, it has shown significant advances and superior performance in face recognition. ArcFace, a key method in this area, uses an angular margin loss to enhance face differentiation [37]. However, its approach of applying a uniform margin doesn't account for varying image qualities within the same class, which can lead to unstable distributions and impact practical recognition tasks. Consequently, recent research in metric learning is increasingly focused on integrating image quality considerations to improve the models' stability and accuracy. Meng *et al*. proposed MagFace, a methodology that uses the magnitude of feature embeddings as an indicator of image quality, enhancing face recognition accuracy by adapting to the difficulty of samples [38]. Kim *et al*. introduced AdaFace, a loss function that adapts to image quality by varying its focus on sample difficulty, using feature norms to improve face recognition [39].

Objectives

In this study, we aim to investigate the application of advanced deep learning methods to medical video analysis, evaluating their efficacy and practicality within a clinical environment. Our objectives are threefold:

A. Surgical phase recognition using Teacher-Student Learning

B. Surgical video inpainting using Implicit Neural Representation

C. Colonoscopy video quantification using Deep Metric Learning and Phase Recognition

Surgical phase recognition using Teacher-Student Learning

1. Background and Objective

Tympanomastoidectomy (TM) is the otologic surgical management of chronic otitis media with or without cholesteatoma and is a procedure on the temporal bone performed using a microscope, a high-speed drill with appropriately sized cutting and diamond burrs, and otologic instruments [3]. The curriculum for learning these surgeries is to watch recorded videos, but for beginners and trainees, simply watching recorded surgery videos can be difficult to understand due to the complexity of the surgery. This situation highlights the need for improved educational tools in surgical training that clearly show the details of each surgical phase.

Advances in deep learning can significantly improve surgical phase recognition, providing advanced tools for both educational and clinical purposes. Surgical phase recognition primarily depends on fully annotated videos, requiring experienced surgeons to meticulously label each phase [13, 14, 15, 16, 17]. However, due to the complexity of TM and the ambiguity of surgical phases, labeling for surgical phase is very limited, and these problems cause some challenges such as class imbalance and data scarcity.

In this study, to solve these challenges, we propose a surgical phase recognition learning method using Teacher-Student learning, an approach based on Knowledge Distillation (KD) [18]. Our method utilizes a 'Teacher' network to generate pseudo-labels for unlabeled segments of surgical videos, thereby expanding the usable dataset for training. This approach could alleviate class imbalance and data scarcity problems. And by using pseudo-labels in the form of soft labels, we expressed the ambiguity of the original TM phase classification, allowing the model to learn the characteristics of the actual TM procedure. We explore the efficacy of our proposed learning method through experiments using various state-of-the-arts surgical phase recognition model.

2. Material and Method

Dataset. In this study, we utilized a total a 65 TM surgical procedure videos featuring 38 patients, recorded at either 25 or 30 FPS. Each video frame has a resolution of 1920 × 1080. The videos encompass 6 distinct phases: "Cortical drilling", "Drilling of the mastoid", "Drilling the antrum", "Drilling near the facial nerve", "Posterior tympanotomy", and "No drilling". Labels were assigned only to segments of the surgical videos that could be clearly identified for the five drilling-related classes. For the "No drilling" phase, we utilized a binary classification model to determine the presence or absence of drilling in each segment and then verified the accuracy of these predictions before using them in our study [40]. For training, we used 42 videos from 26 patients, 7 videos from 4 patients for validation, and 16 videos from 8 patients for testing.

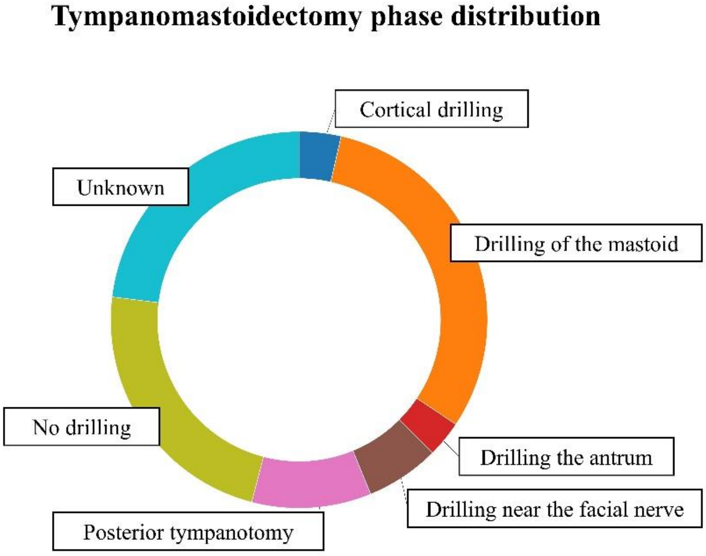


Figure 1. Distribution of surgical phases in TM procedure videos.

Surgical phase recognition. In our study, we employ the commonly used structure in surgical phase recognition: the feature extraction and temporal relation modeling method. This approach first learns a model through frame-wise classification and then trained model is used to extract spatial information features from surgical video. Subsequently, features extracted from the trained feature extractor network are utilized in the temporal relation model to simultaneously leverage spatial and temporal information for more accurate phase prediction [13, 14, 15, 16, 17]. The effectiveness of these approach largely depends on how well the feature extractor model trains the spatial information, making the initial frame-wise classification process crucial [41]. However, in the case of TM surgery, the number of usable frames in the entire video is limited, and class imbalance is a significant issue, leading to a high risk of overfitting [42, 43]. **Figure 1** shows the phase distribution in TM procedure videos, illustrating class imbalance and a considerable "Unknown" segment, indicating unlabeled data, which could complicate model training. To address these challenges, we propose a learning method to further strengthen the feature extractor model by using Teacher-Student learning to make "Unknown" segments available in the training process.

Teacher-Student learning for tympanomastoidectomy phase recognition. In our Teacher-Student learning, the feature extraction model undergoes a two-stage training process. The first step involves training the teacher network using labeled data segments with Cross-Entropy loss. Once the teacher network is sufficiently trained, it generates soft labels for both the labeled and unlabeled data. And then, we leverage these soft labels to train a student network by applying a modified Knowledge Distillation loss function. Typically, Knowledge Distillation loss involves computing the KL divergence Loss for the soft labels and the Cross-Entropy Loss for the hard labels, with the sum of these losses forming the final loss [18]. Our modified loss function reflects the ambiguity of surgical phase, a characteristic of TM video. For unlabeled segments where hard labeling is not possible due to ambiguity, the loss is calculated using only the KL divergence with soft labels as the target. For labeled segments, Cross-Entropy loss and KL Divergence loss were calculated and added in the same way as Knowledge Distillation. By calculating and adding cross entropy loss using existing hard labels, the model could learn that labeled segments are clearly distinguishable surgical phases from unlabeled segments. The loss function L is defined as follows:

$$L = \sum_{(x,y,l)\in\mathbb{D}} L_{KL}\big(S(x,\theta_S,\tau),\ T(x,\theta_T,\tau)\big) + I(l)L_{CE}(S(x,\theta_S),y)$$

where L_{KL} denotes the KL divergence loss which measures the discrepancy between the predictions of the student network S and the teacher network T, with τ representing the temperature parameter that softens the outputs for a more effective knowledge transfer. The indicator function $I(l)$ determines whether a data is labeled and accordingly applies the Cross-Entropy loss L_{CE}, which computes the loss for the labeled data only. The student model, trained with this loss function, is utilized as the feature extractor network. An overview of this methodological approach is depicted in **Figure 2**.

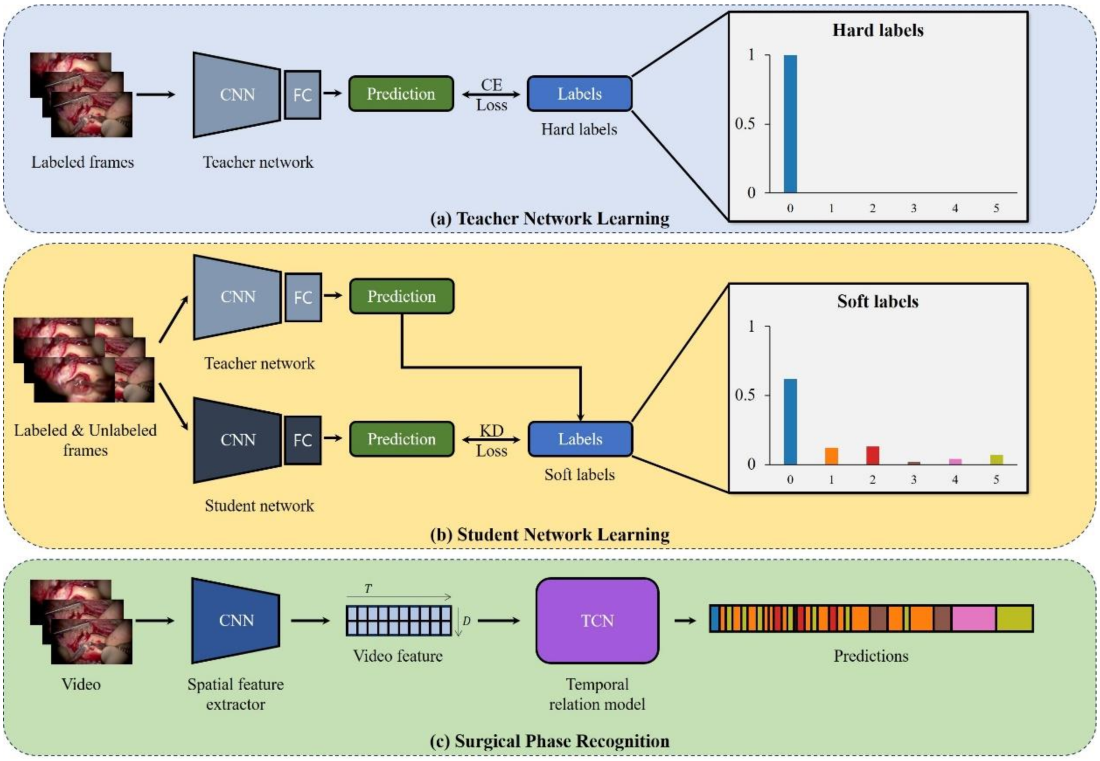


Figure 2. Overview of the Teacher-Student learning for surgical phase recognition. (a) Teacher Network Learning: Trains with labeled frames using a CNN, yielding predictions assessed by Cross-Entropy loss. (b) Student Network Learning: Combines labeled and unlabeled frames, using predictions from the teacher network to train the student network via modified Knowledge Distillation loss. (c) Surgical Phase Recognition: Processes video through a CNN to extract spatial features, then uses a temporal relation model to generate final phase predictions.

3. Experiments and Results

Metrics. Evaluation of phase prediction in TM videos was restricted to labeled segments, utilizing conventional phase recognition metrics such as F1 Score, Accuracy, Precision, Recall, and Jaccard Index. Accuracy is indicative of a video-level assessment, reflecting the percentage of correctly identified frames across the full duration of the video. Considering the imbalanced distribution of phases, F1 Score, Precision, Recall, and Jaccard Index metrics were adopted for video-level evaluation. We appraised the performance of our proposed method by applying it to well-known or state-of-the-art surgical phase recognition models, comparing the efficacy of training the feature extractor model solely with labeled data against training it through teacher-student learning with both labeled and unlabeled data. Additionally, to examine the influence of soft labels, we compared results between pseudo-labels transformed into hard labels and those constituted as soft labels.

Implementation details. Our experiments were executed utilizing the PyTorch framework on 4 NVIDIA GeForce RTX 3090 GPUs. For the feature extractor models, both the teacher and student networks were initialized with ImageNet pre-trained ResNeSt50 [8] architectures. We employed the AdamW [44] optimizer with parameters set to betas (0.9, 0.999), 1e-08 eps, and 0.05 weight decay. The learning rate was scheduled using the CosineAnnealingWarmRestarts method [45]. Owing to GPU memory limitations, images were resized to a resolution of 256 × 256 pixels. To prevent overfitting and to enhance the dataset's variety, we applied eight types of image augmentations provided by the Albumentations library [46], including ShiftScaleRotate, ColorJitter, RandomBrightnessContrast, GaussNoise, ISONoise, MedianBlur, MotionBlur, and HorizontalFlip.

Table 1. Comparison of video-level surgical phase recognition performance.

Method		F1 score (%)	Accuracy (%)	Precision (%)	Recall (%)	Jaccard (%)
MS-TCN [5]	Labeled	69.4 ± 16.2	87.5 ± 6.1	77.0 ± 15.0	74.9 ± 14.5	60.9 ± 16.7
	+Pseudo (Hard label)	75.1 ± 15.6	89.0 ± 7.1	82.1 ± 13.0	80.1 ± 13.0	67.7 ± 17.8
	+Pseudo (Soft label)	**77.6 ± 12.6**	**89.8 ± 4.7**	**85.0 ± 9.2**	**81.8 ± 10.8**	**69.9 ± 14.3**
Not-End [10]	Labeled	70.9 ± 14.2	87.6 ± 4.7	76.8 ± 14.6	78.6 ± 15.2	63.3 ± 13.7
	+Pseudo (Hard label)	75.2 ± 18.4	90.3 ± 7.1	81.9 ± 13.9	80.0 ± 16.1	68.6 ± 20.3
	+Pseudo (Soft label)	**79.8 ± 10.9**	**90.8 ± 4.0**	**84.6 ± 9.6**	**83.5 ± 9.7**	**71.5 ± 13.1**
SAHC [11]	Labeled	75.4 ± 13.8	88.9 ± 5.7	79.4 ± 13.5	80.7 ± 12.2	67.4 ± 15.5
	+Pseudo (Hard label)	73.2 ± 9.6	90.6 ± 3.6	78.2 ± 11.3	76.2 ± 8.6	65.2 ± 11.3
	+Pseudo (Soft label)	**80.2 ± 13.8**	**91.0 ± 4.5**	**81.7 ± 14.0**	**86.2 ± 11.1**	**73.5 ± 15.9**

Table 2. Comparative analysis of phase-level evaluation metrics for surgical videos.

Method		Cortical drilling	Drilling of the mastoid	Drilling the antrum	Drilling the facial nerve	Posterior tympanotomy	No drilling
Precision (%)	Labeled	83.5 ± 27.7	92.4 ± 5.7	91.5 ± 19.8	42.7 ± 30.4	77.5 ± 35.0	**97.9 ± 2.0**
	+Pseudo (Hard label)	**85.2 ± 25.5**	87.5 ± 5.0	66.1 ± 43.7	**56.4 ± 39.0**	85.2 ± 16.0	97.0 ± 2.8
	+Pseudo (Soft label)	79.2 ± 28.1	**95.4 ± 5.8**	**91.8 ± 20.2**	48.5 ± 34.3	**89.3 ± 19.4**	96.5 ± 2.6
Recall (%)	Labeled	89.7 ± 12.4	95.3 ± 4.8	72.9 ± 23.9	**87.5 ± 30.5**	58.3 ± 33.2	91.5 ± 6.3
	+Pseudo (Hard label)	84.9 ± 13.0	**96.2 ± 3.4**	38.4 ± 36.1	75.8 ± 33.2	**74.0 ± 26.9**	92.7 ± 8.6
	+Pseudo (Soft label)	**98.3 ± 3.5**	93.0 ± 5.2	**92.1 ± 11.8**	85.2 ± 34.8	67.4 ± 27.5	**94.2 ± 4.8**
Jaccard (%)	Labeled	73.7 ± 25.3	88.2± 6.2	64.5 ± 22.7	42.3 ± 30.9	53.1 ± 32.8	89.9 ± 7.0
	+Pseudo (Hard label)	71.8 ± 23.2	84.5 ± 4.3	33.1 ± 28.6	**48.4 ± 32.2**	**67.1 ± 27.9**	90.5 ± 9.6
	+Pseudo (Soft label)	**77.7 ± 27.4**	**88.9 ± 6.4**	**83.9 ± 20.4**	48.2 ± 34.0	61.0 ± 26.9	**91.1 ± 5.2**

Surgical phase recognition results. Table 1 displays the video-level quantitative results, with mean and standard deviation metrics, for evaluating TM phase recognition. The performance of three surgical phase recognition models is shown when trained exclusively on labeled data, when incorporating hard pseudo-labels for unlabeled data, and when utilizing soft pseudo-labels. Training with soft pseudo-labels consistently outperformed the other methods in terms of F1 score, accuracy, precision, recall, and Jaccard index across all models. The SAHC model, in particularly, achieved the highest metrics when enhanced with soft pseudo-labels, with an F1 score of 80.2 ± 13.8, accuracy of 91.6 ± 4.5, recall of 86.2 ± 11.1 and Jaccard index of 73.5 ± 15.9. **Table 2** provides phase-level precision, recall, and Jaccard index evaluations within TM surgical videos. Similarly, the application of soft pseudo-labels during training exhibited superior overall performance compared to the other two approaches. For qualitative analysis, **Figure 3** illustrates the results on TM surgical videos using color-coded ribbons to represent three learning methods: labeled data, hard pseudo-labels and soft pseudo-labels. It is observable that the application of soft pseudo-labels generally yields results that more closely align with the ground truth, particularly for the class represented by the red color segment, “Drilling the antrum”, which is more accurately predicted with this method compared to the others. However, as evident in the results for videos (c) and (d), there is confusion between the brown color segment, “Drilling the facial nerve”, and the pink color segment, “Posterior tympanotomy”. We will discuss this confusion in discussion section. These results can also be confirmed through confusion matrices (**Figure 4**).

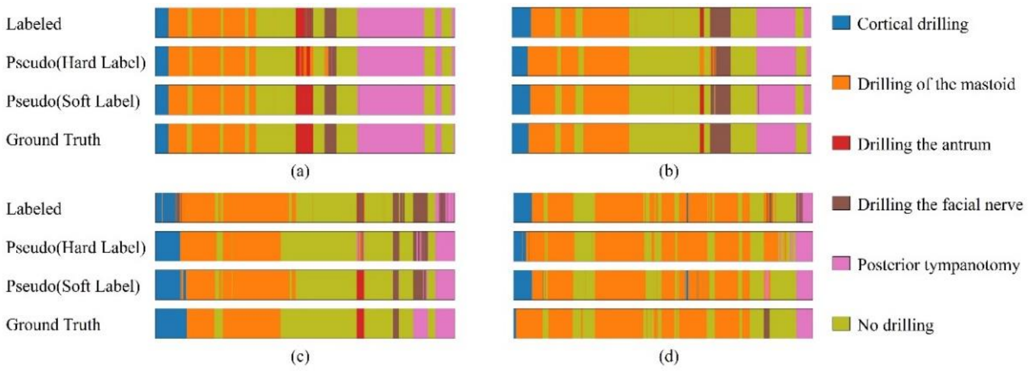
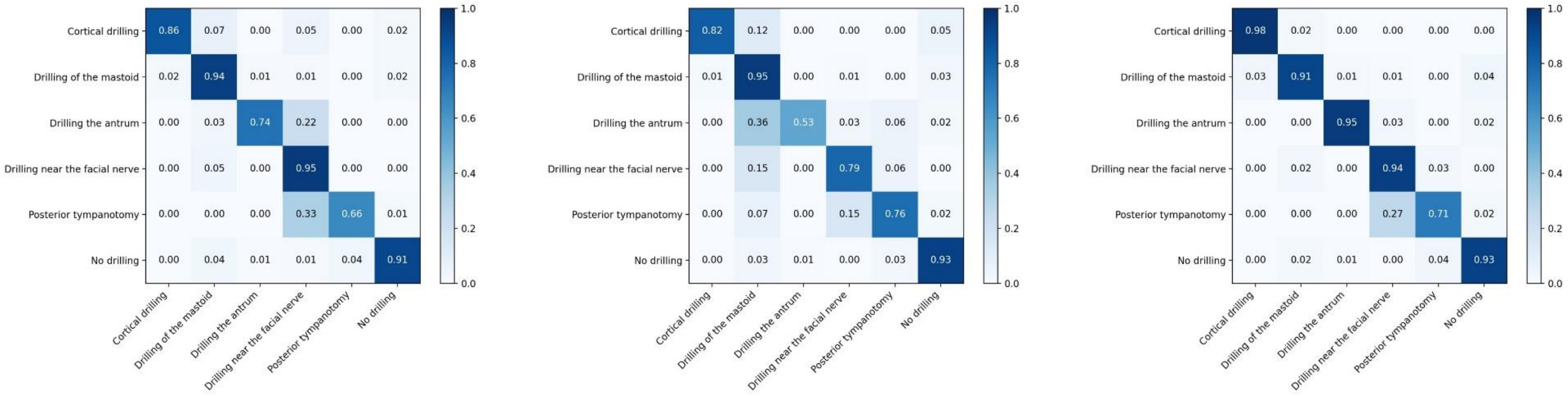


Figure 3. Visualization of four TM phase recognition results using color-coded ribbons for labeled, hard pseudo-label, soft pseudo-label learning methods and the ground truth (a-d).



	Cortical drilling	Drilling of the mastoid	Drilling the antrum	Drilling near the facial nerve	Posterior tympanotomy	No drilling
Cortical drilling	0.86	0.07	0.00	0.05	0.00	0.02
Drilling of the mastoid	0.02	0.94	0.01	0.01	0.00	0.02
Drilling the antrum	0.00	0.03	0.74	0.22	0.00	0.00
Drilling near the facial nerve	0.00	0.05	0.00	0.95	0.00	0.00
Posterior tympanotomy	0.00	0.00	0.00	0.33	0.66	0.01
No drilling	0.00	0.04	0.01	0.01	0.04	0.91

(a) Labeled

	Cortical drilling	Drilling of the mastoid	Drilling the antrum	Drilling near the facial nerve	Posterior tympanotomy	No drilling
Cortical drilling	0.82	0.12	0.00	0.00	0.00	0.05
Drilling of the mastoid	0.01	0.95	0.00	0.01	0.00	0.03
Drilling the antrum	0.00	0.36	0.53	0.03	0.06	0.02
Drilling near the facial nerve	0.00	0.15	0.00	0.79	0.06	0.00
Posterior tympanotomy	0.00	0.07	0.00	0.15	0.76	0.02
No drilling	0.00	0.03	0.01	0.00	0.03	0.93

(b) Pseudo (Hard label)

	Cortical drilling	Drilling of the mastoid	Drilling the antrum	Drilling near the facial nerve	Posterior tympanotomy	No drilling
Cortical drilling	0.98	0.02	0.00	0.00	0.00	0.00
Drilling of the mastoid	0.03	0.91	0.01	0.01	0.00	0.04
Drilling the antrum	0.00	0.00	0.95	0.03	0.00	0.02
Drilling near the facial nerve	0.00	0.02	0.00	0.94	0.03	0.00
Posterior tympanotomy	0.00	0.00	0.00	0.27	0.71	0.02
No drilling	0.00	0.02	0.01	0.00	0.04	0.93

(c) Pseudo (Soft label)

Figure 4. Confusion matrices for surgical phase recognition using different learning methods: (a) Labeled, (b) Hard-pseudo label, and (c) Soft-pseudo label. Each matrix compares the predicted phases to the true phases.

4. Discussion

In this study, we introduced a Teacher-Student learning framework for surgical phase recognition in TM video, aimed at alleviating the challenges associated with class imbalance and data scarcity. In Tympanomastoidectomy (TM) surgical videos, the location of the “Drilling of the antrum” and the “Drilling of the facial nerve” can be confusing because the locations are interchanged depending on whether the right or left ear surgery. Moreover, "Drilling the antrum" has the least amount of data in the class imbalance situation, so there is a risk of overfitting with "Drilling the facial nerve", which has a relatively large amount of data (**Figure 1**) [47]. In this situation, utilizing teacher-student learning structure that can create pseudo-labels for unlabeled data and a soft label that can reflect the ambiguity of surgical phases characteristic of unlabeled TM video segments can be effective for TM surgical phase recognition (**Table 1** and **2**). Training exclusively on labeled data has been observed to lead to misclassifications, confusing “Drilling the antrum” with “Drilling the facial nerve”. Training with hard pseudo-labels appears to mitigate the confusion between “Drilling the antrum” and “Drilling the facial nerve”, yet it also introduces a new misclassification trend of mistaking “Drilling the antrum” for “Drilling of the mastoid”. The reasons for this trend can be class bias due to class imbalance and scarcity of data. The main reason for the unlabeled segments of TM surgery is the blurred boundary between “Drilling of the mastoid” and other surgical procedures, and **Figure 1** shows that 'Drilling of the mastoid' occupies the most area. In this situation, there is a possibility of class bias in the process of hard labelling the prediction of the model trained with labeled data. On the other hand, training with soft pseudo-label approach shows the robustness, particularly for “Drilling the antrum” phase (**Figure 3** and **4**). Soft pseudo-labels are thought to reduce the problem of overfitting and misclassification due to the nature of TM video, which can exhibit phase ambiguity. Misclassifications of the “Posterior tympanotomy” phase as “Drilling the facial nerve” were also noted, which is reasonable considering these phases occur in succession during TM surgery.

We believe this Teacher-Student learning method for surgical phase recognition extends beyond TM video and can be applied to any surgical video with similar characteristics.

Surgical video inpainting using Implicit Neural Representation

1. Background and Objective

Video inpainting, the process of filling spatiotemporal gaps in videos, has seen significant advancements with the advent of deep learning. This technique holds potential for various applications, such as removing unwanted objects, restoring damaged footage, and retargeting videos. However, applying these inpainting models to specific domains like medical videos, especially surgical videos, poses some challenges.

Dataset Collection: Various deep learning-based video inpainting models, trained on extensive video datasets, outperform their predecessors [28, 29, 20]. Yet, the process of collection these large datasets is labor-intensive and time-consuming. More critically, the performance of these models may degrade when applied to videos from domains different from their training datasets [48]. This issue is even more difficult to collect large amounts of video data in the medical field due to patient privacy concerns.

High-Resolution Video Inpainting: Video inpainting inherently necessitates the consideration of both spatial and temporal consistency. To leverage this property, many video inpainting methods refer to multiple adjacent frames simultaneously for a more coherent inpainting result. However, referencing multiple frames simultaneously consumes substantial memory, prompting most methods to downscale the original video, which inevitably compromises its quality. However, low-resolution videos can obscure surgical landmarks and hinder understanding of surgical procedures, a video inpainting technology that can maintain the original resolution is needed for application to surgical videos.

Although some recent papers have introduced “internal learning” for video inpainting: a method that enables training directly on the test video without the need for a large dataset, these methods still struggle with handling high-quality videos in one go. They typically involve complex and resource-intensive multi-stage training processes for high-resolution video inpainting [48, 49].

In response to these challenges, our approach leverages Implicit Neural Representation (INR). INR, a form of internal learning, allows training without vast datasets. This technology has the characteristics of being resolution-agnostic and representing continuous signals well [31].

Using the capabilities of INR, we present a method that enables high-resolution inpainting in a single training step and at the same time does not require massive video datasets. In this study, we experiment with inpainting surgical tools on surgical videos and compare the visual results with state-of-the-art video inpainting methods. Additionally, through surveys with trainee surgeons, we evaluate the impact of our technology on surgical education.

2. Material and Method

Dataset. In this study, we used five distinct tympanomastoidectomy (TM) surgical procedure video clips, each with a duration of approximately 10 seconds and a resolution of 1920×1080. These video clips encompassed various surgical procedures including “Drilling of the mastoid cortex”, “Drilling of the mastoid air cells around the antrum and tegmen mastoideum (superficial mastoidectomy)”, “Drilling into the auditus ad antrum”, “Drilling over the tegmen mastoideum”, and “Drilling over the sigmoid sinus”. The masks used for inpainting the surgical tools were created manually.

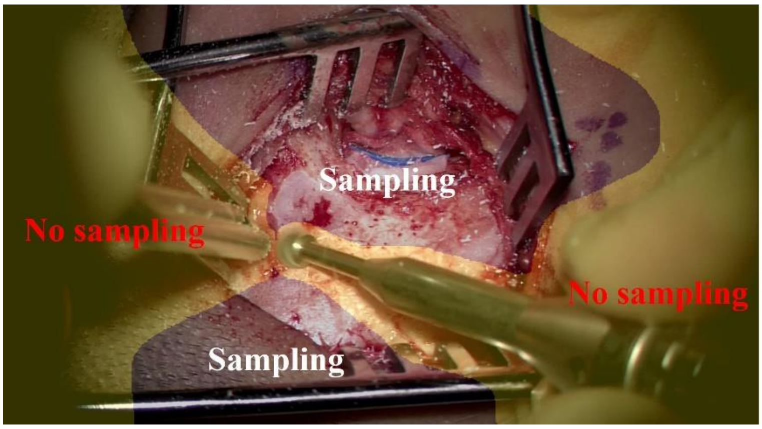


Figure 5. Illustration of mask-guided ray casting in a surgical setting. The image shows the application of the tool mask, where 'Sampling' indicates the regions being used and 'No sampling' denotes the masked areas being not used in training procedure.

Video inpainting using Implicit Neural Representation. In this study, we utilize the HyperReel [34] model, an advancement over the Neural Radiance Field (NeRF) [36] approach for the task of novel view synthesis. While the traditional NeRF method is designed solely for rendering static objects, HyperReel model is distinct in its capability to process sparse view video inputs concurrently. This incorporation of a temporal dimension in HyperReel enables the effective synthesis of video content. The method utilizes a function $F_\theta : (x, d, t) \rightarrow (c, \sigma)$, where position x, direction d along a ray, and time t are mapped to RGB radiance c and opacity σ. HyperReel uses rendering techniques to render the radiance field for a given query view in both space and time. The rendered color of a pixel corresponding to a ray, defined as $r(s) = o + sd$, with the origin o and direction d defined by the camera pose and intrinsic, can be calculated by integrating the radiance multiplied by the accumulated opacity:

$$C^{(t)}(r) = \int_{s_n}^{s_f} T(s)\sigma(r(s), t)c(r(s), d, t))ds$$

where s_n and s_f denote the bounds of the volume depth range, and the accumulated opacity $T(s) = exp(-\int_{s_n}^{s} \sigma(r(p), t))dp)$. In our research, we focused on single-view videos captured from a fixed camera positioned above the surgical field. To facilitate the training process, we fixed the camera parameters to arbitrary constants, allowing us to conduct the experiments in a single-view setting.

Mask-guided ray casting. To exclude surgical tools from surgical videos and exclusively reconstruct tissue, we employed mask-guided ray casting, a technique introduced in EndoNeRF [50]. This method involves using a tool mask to selectively sample only the unmasked regions of the video (as illustrated in **Figure 5**). By focusing on these specific areas, the technique enables the model to train on relevant information while omitting extraneous details. Subsequently, the areas obscured by the mask are naturally inpainted by the model, which has learned from the unmasked regions. The overview of our approach is illustrated in **Figure 6**.

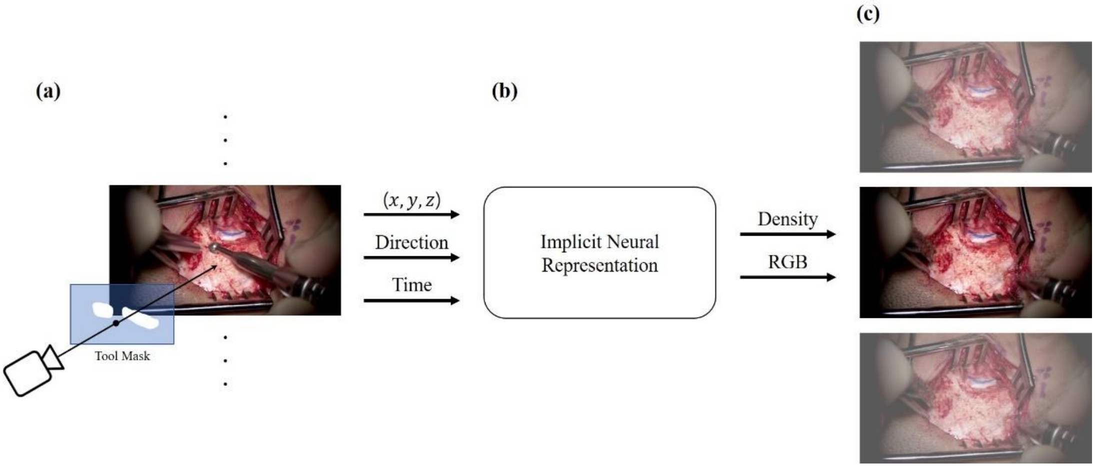


Figure 6. Overview of the video inpainting pipeline using INR. (a) Sampling the unmasked region using a tool mask, (b) using the sampled location and temporal information to predict the corresponding RGB and density values through the INR model, and (c) the actual inpainting process where the predicted information is used to reconstruct the occluded regions.

3. Experiments and Results

Evaluation of the video inpainting. We conducted a qualitative comparison between our outcomes and those produced by E2FGVI [24], a state-of-the-art video inpainting model. Furthermore, to assess the educational impact of the inpainted videos with surgical tools removed, we conducted a survey-based evaluation. We enrolled 21 trainees of Asan Medical Center (AMC) (6 medical students, 12 residents in Otorhinolaryngology-Head and Neck Surgery, and 3 fellows in the Otology Neurotology program) and asked them to watch processed TM videos and complete questionnaires. We divided subjects into 4 subgroups: student, low-grade resident (R1, R2), high-grade resident (R3, R4), and fellow. We analyzed the results of the questionnaires comprehensively according to the subgroups.

Questionnaire. The questionnaire comprised 6 questions: Did the processed surgical video help in (1) visual comfort, (2) identifying bleeding focus, (3) identifying the sigmoid sinus, (4) identifying the tegmen mastoideum, (5) understanding the surgical process, and (6) performing the surgical procedure when compared to the original unprocessed video. Each question was scored on a 5-point discrete scale from +2 to -3; +2 (helpful), +1 (slightly helpful), 0 (indifferent), -1 (slightly bothersome), and -2 (bothersome). Questionnaire scores were presented as mean ± standard deviation. The statistical analysis was performed using SPSS software version 24.0 (IBM Corp., Armonk, NY, USA). A p-value of less than 0.05 was considered statistically significant.

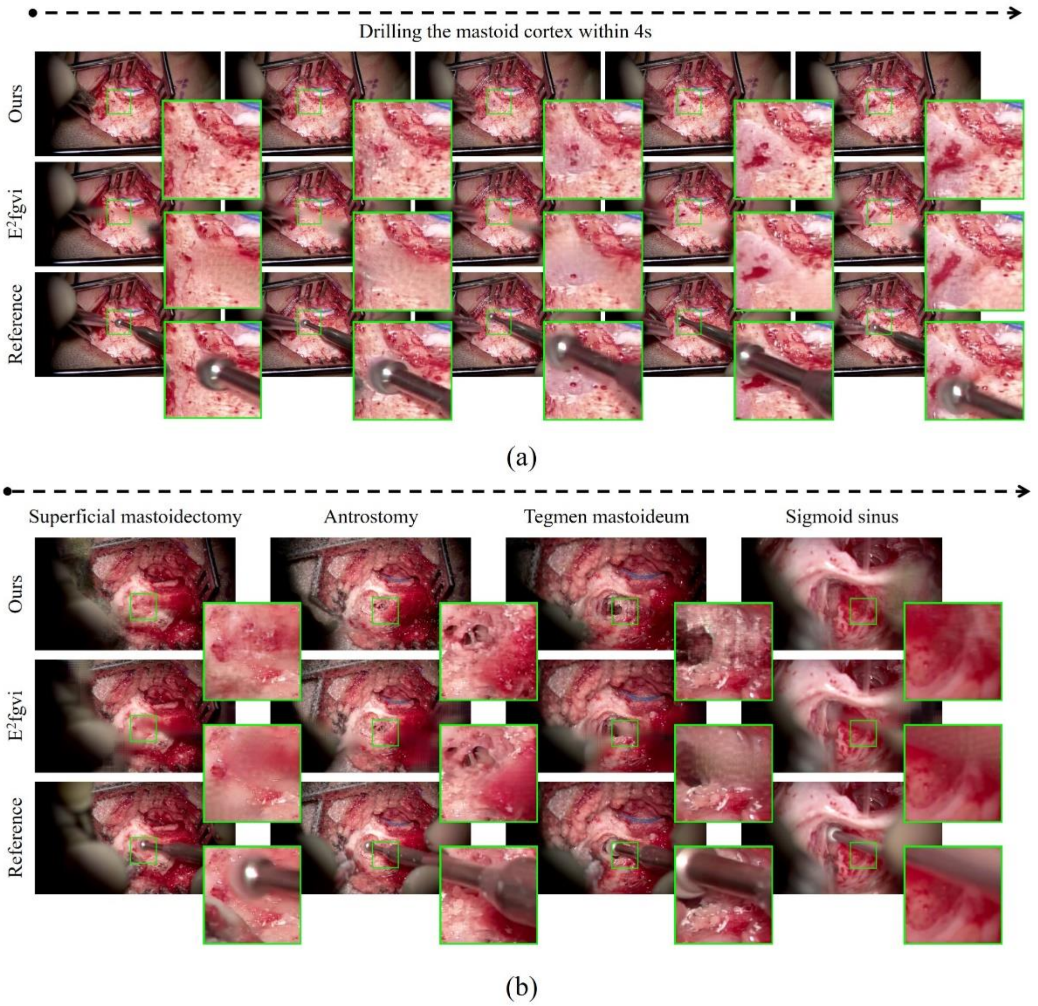


Figure 7. Qualitative result of 5 TM surgical video inpainting. (a) Results on the case "Drilling the mastoid cortex" within 4s, and (b) results on other 4 surgical procedure.

Qualitative Results. For qualitative evaluation, **Figure 7** displays the qualitative results of our video inpainting technique in comparison to a state-of-the-art (SOTA) method. In **Figure 7** (a), "Drilling the mastoid cortex", the sequence displays consecutive frames within a 4 second. Our approach benefits from a continuous representation of the scene, which can be observed through its ability to seamlessly reconstruct the flow of bleeding over time. In contrast, the frames processed by the current SOTA method show noticeable blurring where tools were previously positioned, and a less continuous portrayal of bleeding, indicating a deficiency in maintaining the temporal consistency of the video. In **Figure 7** (b), our method's inpainting capabilities are validated through its application to a frame from videos of various surgical procedures: "Superficial mastoidectomy", "Antrostomy", "Tegmen mastoideum", and "Sigmoid sinus". As in Figure 7 (a), the results of the comparison method confirm that the texture of the area previously obscured by the surgical tool has become blurred. On the other hand, the results of our method can be seen that the texture is reconstructed in more detail.

Table 3. Questionnaire scores for the video inpainting.

	Visual comfort	Bleeding focus	Sigmoid sinus	Tegmen mastoideum	Understandi ng	Performing	Average
Student (n=6)	0.97±0.96	0.80±0.42	0.63±0.56	0.70±0.45	0.83±0.70	0.80±0.95	**0.79±0.58**
Low-grade resident (n=6)	0.43±0.61	0.63±0.27	0.30±0.28	0.20±0.25	0.43±0.37	0.70±0.37	**0.45±0.26**
High-grade resident (n=6)	0.00±0.55	0.87±0.24	0.47±0.27	0.27±0.21	0.50±0.30	0.73±0.39	**0.47±0.21**
Fellow (n=3)	-0.20±0.60	0.67±0.58	-0.47±0.12	-0.07±0.61	0.13±0.31	0.20±0.60	**0.04±0.33**
Average	**0.37±0.80**	**0.75±0.35**	**0.33±0.50**	**0.32±0.43**	**0.52±0.49**	**0.67±0.61**	**0.50±0.43**

Questionnaire Results. For evaluating video inpainting technique, 5 videos processed with our method were evaluated. According to the average score over the 5 videos, medical students valued the videos as the most helpful (0.79 ± 0.58), followed by high-grade residents (0.47 ± 0.21), low-grade residents (0.45 ± 0.26), and fellows (0.04 ± 0.33) (**Table 3**). The fellow group evaluated videos as less helpful than the other subgroups. However, there was no significant difference among the four subgroups (p = 0.185, **Figure 8**). When 5 different procedures were analyzed in each of 5 videos that showed different procedures, every video had a score of more than 0.50, and a video that mainly showed drilling over the sigmoid sinus had the highest score (0.60 ± 0.45) (**Table 4**).

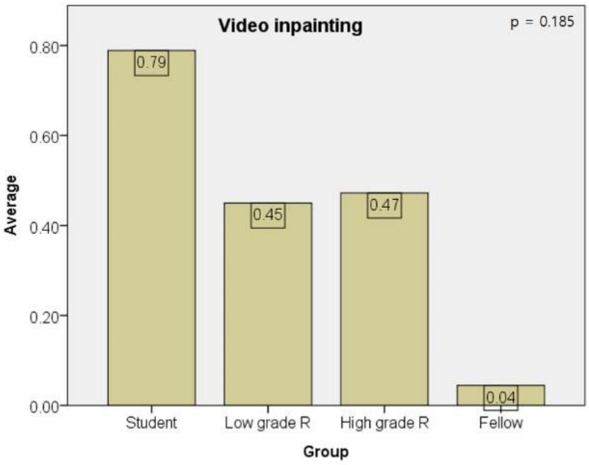


Figure 8. Bar graph showing the average results of a survey on video inpainting effectiveness across different groups: Students, Low-grade Residents (Low grade R), High-grade Residents (High grade R), and Fellows.

Table 4. Average questionnaire scores of the video inpainting for five different surgical procedures.

Video drilling	Mastoid cortex	Superficial mastoidectomy	Antrum	Tegmen mastoideum	Sigmoid sinus	Average
Student (n=6)	0.94±0.54	0.67±0.82	0.64±0.81	0.81±0.69	0.89±0.48	**0.79±0.58**
Low-grade resident (n=6)	0.44±0.33	0.53±0.27	0.44±0.25	0.33±0.41	0.50±0.24	**0.45±0.26**
High-grade resident (n=6)	0.36±0.29	0.39±0.25	0.47±0.32	0.42±0.27	0.72±0.31	**0.47±0.21**
Fellow (n=3)	0.00±0.44	0.11±0.35	0.06±0.38	0.06±0.51	0.00±0.44	**0.04±0.33**
Average	**0.50±0.49**	**0.47±0.50**	**0.45±0.51**	**0.45±0.52**	**0.60±0.45**	**0.50±0.43**

4. Discussion

In this study, we introduce a method for inpainting high-resolution surgical videos that does not require extensive data collection. The qualitative results suggest that our approach offers significant improvements over existing state-of-the-art methods, particularly in reconstructing continuous and temporally consistent scenes, such as the flow of bleeding during surgery (**Figure 7**). By improving the visual clarity and fidelity of video inpainting results, we could provide better quality training materials to trainees.

The results of the survey among trainee surgeons at Asan Medical Center (AMC) suggest a significant divergence in the perceived educational value of AI enhanced videos among trainees with varying levels of experience in otologic surgery. Notably, trainees at earlier stages of their medical education: medical students and low-grade residents reported a benefit from the use of video inpainting techniques. The AI processed videos particularly excelled in aiding the identification of bleeding focus, a critical aspect of surgical procedures. This benefit is likely attributed to the novice trainees' lesser familiarity with ear anatomy and procedural nuances, which necessitates a clearer visual field to bolster their comprehension.

Conversely, the otology fellows, who boast several years of hands-on experience, voiced a markedly different perspective. The inpainting process was perceived as having a negligible or non-beneficial impact on their understanding of the surgical techniques. The incomplete removal of surgical tools occasionally led to a distraction, as it interfered with the fellows' ability to accurately discern anatomical landmarks. Additionally, fellows reported that image distortions resulting from the inpainting process caused visual discomfort, further detracting from the utility of the AI-modified videos for their advanced training needs.

This discrepancy highlights the need to tailor training tools according to the level of experience and expertise of the trainee. For those at the initial stages of learning otologic surgery, AI powered video inpainting may serve as a valuable educational tool, enhancing their visual experience and potentially accelerating their understanding of complex surgical procedures. Meanwhile, for seasoned surgeons, traditional videos with unaltered views of the surgical field might continue to be preferred for their realism and fidelity to the actual operative environment. If these points are reflected in the optimization of the surgical education system and AI technology is developed, the quality of education is expected to be further improved.

Colonoscopy quantification using Deep Metric Learning and Phase Recognition

1. Background and Objective

Colonoscopy withdrawal time, one of the evaluation items for colonoscopy, has recently been attracting attention as a key quality marker through studies showing a correlation between withdrawal time and polyp detection rates [51, 52, 53, 54]. However, there are several problems with measuring withdrawal time in clinical practice. First, colonoscopy is a very delicate procedure that requires a high level of physician concentration, and measuring withdrawal time during the procedure can be a distraction. Second, the measurement of withdrawal time is different for each doctor. Therefore, it is essential to automate the quantification of withdrawal time through video analysis to allow objective assessment and allow physicians to focus solely on the clinical aspects of colonoscopy.

The advancement of artificial intelligence (AI), particularly in deep learning, has led to the development of various technologies for video analysis, with numerous attempts to apply these to medical videos [55, 56, 57]. A notable example is surgical phase recognition [13, 14, 15, 16, 17]. This method employs sophisticated computational algorithms to precisely detect and categorize various phases of surgical operations. However, applying such technologies to colonoscopy videos presents challenges due to factors like peristaltic motion, light reflection, foreign substances, and variations in illumination and color. These elements may pose challenges for deep learning models in accurately learning the anatomical structure of the colon. To overcome these issues, we proposed a method for measuring withdrawal time more accurately through phase recognition utilizing quality-aware metric learning [38, 39]. We evaluated our method by comparing it with withdrawal times calculated by physicians.

2. Materials and Methods

Dataset. This study utilized a collection of colonoscopy videos acquired from Asan Medical Center over the period from October to November 2020. A total of 79 colonoscopy procedures were included in this dataset. The videos were annotated with the start and end points of a specific section in colonoscopy. The labeled sections are follows:

- Cecum Identification Segment (Cecum): This segment covers the timeframe in which the cecum, the beginning of the large intestine, is detected in the video.

- Instrument Withdrawal Segment (Exit): This segment refers to the section where the endoscope is removed from the patient’s body.
- Procedure Segment (Surgery): During this segment, various sections were labeled and include “Forcep”, “Clip”, “Endoscopic Submucosal Dissection (ESD)”, “Injection”, “Polyp Removal”, “Coagulation (Coag)”, Precutting”, “Tattooing”, “Snaring”, and “Endoscopic Mucosal Resection (EMR)”.
- Observation Segment (Observation): This segment includes all video footage not encompassed by the previously defined categories. It primarily consists of standard inspection phases where the endoscope is navigated through the colon without performing any specific sections.

The annotations were made by qualified endoscopist to ensure accuracy and reliability of the labeled data.

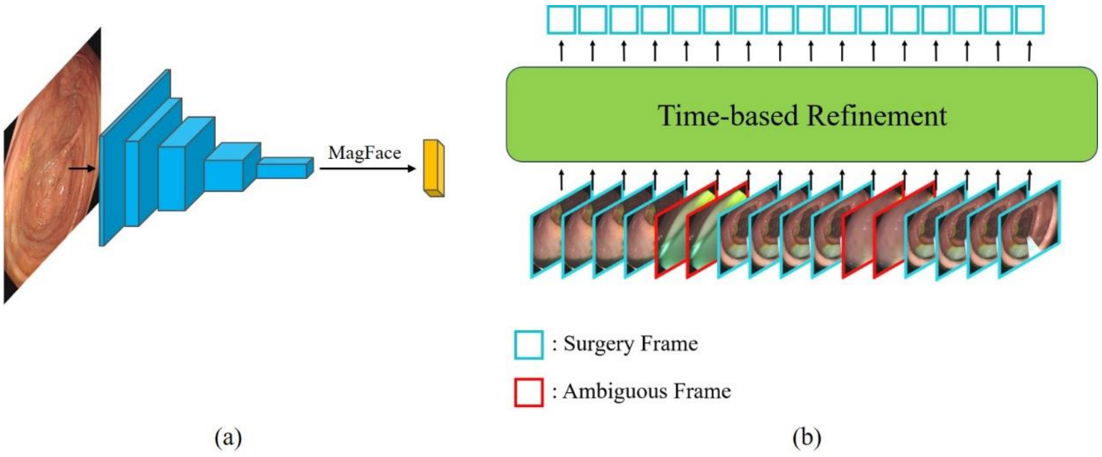


Figure 9. Schematics of the model architecture. (a) Frame-by-frame analysis using MagFace to evaluate image quality. (b) Utilization of a temporal relation model to refine intermediate ambiguous frames for improved final prediction accuracy.

Training strategy. In our study, there are two key points to perform colonoscopy withdrawal time evaluation: quality-aware metric learning, and video phase recognition. The overview of our approach is illustrated in **Figure 9**.

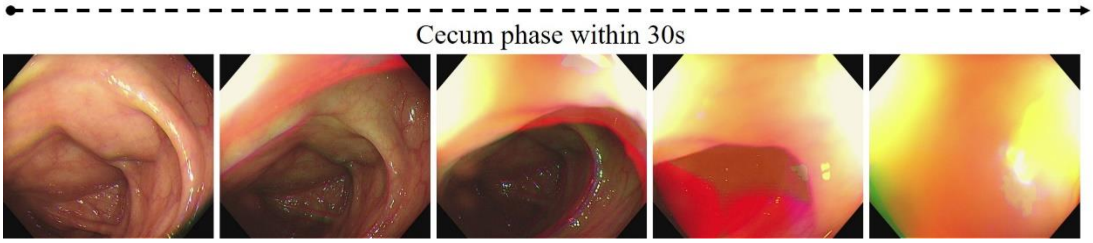


Figure 10. Sequence of cecum phase images captured over a 30 second interval.

Overcoming image quality variability. In the acquisition of colonoscopic imagery, factors such as the inherent peristaltic movement of the colon and various image quality degradations including washing effects and light reflection present some challenges. As can be seen in **Figure 10**, even within the same cecum phase, images of various quality appear, ranging from images in which the cecum is easily recognized to images in which the cecum is difficult to recognize. The presence of low-quality images poses a risk of the model overfitting to incorrect information during deep learning training procedure [38, 39]. To address these issues, we utilize MagFace [38], one of the quality-aware metric learning methods, in the model learning process. MagFace is known for its capability to adjust the feature scale adaptively based on the quality of the input image, which aligns with our objective of mitigating the influence of low-quality images on the learning process. We used this method for frame-level classification training so that the model can learn discriminative features while recognizing image quality. Upon completion of the training phase, the trained backbone network is used as a feature extractor to extract spatial information from the colonoscopy video. These extracted features are then employed for time-series analysis to further process and interpret the colonoscopic video data. By implementing this strategy, we aimed to enhance the robustness of our model against the variability of image quality and ensure the learning is focused on features indicative of the colon’s structure, rather than artifacts and noise.

Video phase recognition for temporal consistency. Colonoscopy often produces images that make it difficult to determine which section of the colon corresponds to factors such as peristaltic motion, light reflection, foreign substances, and variations mentioned above. Due to these characteristics, the method of predicting colonoscopy video frame-wise may generate numerous errors and may not be appropriate for calculating accurate withdrawal time. To avoid errors that may occur when making frame-wise predictions, we use a video phase recognition method of predicting sections by considering temporal information. Video phase recognition is increasingly used in medical applications, particularly for segmenting surgical phases in operation videos, and it aims to partition a temporally untrimmed video by time, labeling each segmented part with predefined action labels [13, 14, 15, 16, 17]. In our study, we utilize a surgical phase recognition method known as SAHC [17]. SAHC employs Transformer to understand the relationship between segments at different temporal resolutions. It is particularly tailored for refining ambiguous frames that may disrupt accurate predictions mid-segment, thus enhancing the precision of segment identification and temporal localization. The characteristic of this model allows us to maintain the integrity of segment recognition despite intermittent low-quality or obstructive frames, contributing to a more reliable and accurate assessment of withdrawal times in colonoscopic procedures.

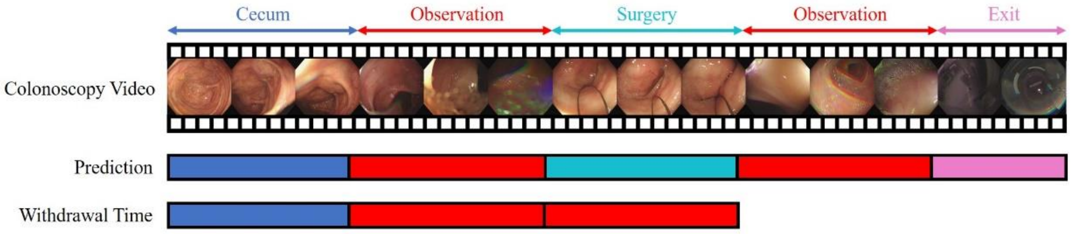


Figure 11. Overview of the automatic evaluation of colonoscopy withdrawal time. The system identifies key phases such as Cecum, Observation, Surgery, and Exit within the colonoscopy video. The first frame recognized as Cecum is designated as the starting point of the withdrawal time, and the final withdrawal time is calculated by subtracting the recognized Surgery and Exit phase within the section from the starting point until the end of the video.

Withdrawal times evaluation. To facilitate the assessment of colonoscopy, our study focuses on the automated calculation of withdrawal times from video segments identified. In our approach, withdrawal time is calculated from the moment the cecum is first identified, indicating the start of the withdrawal phase, to when the colonoscope is fully removed from the patient’s body. It is important to note that the section representing the surgery and the section where the endoscope leaves the body is not included in the withdrawal time calculation. This segment does not reflect the mucosal inspection phase and is therefore excluded from the overall withdrawal time calculation (**Figure 11**).

3. Experiments and Results

Metrics. Our evaluation of video phase recognition performance on colonoscopy videos used established metrics such as Accuracy, Precision, Recall, and the Jaccard index. To examine whether MagFace truly learns by considering image quality, we sampled images from the cecum section and evaluated the correlation between the magnitude of image features and image quality. Also, to validate the discriminative capability imparted by metric learning, we visualized class separability using t-SNE [58]. To assess the accuracy of withdrawal time, we calculated the Mean Absolute Error (MAE) in seconds, which accounts for the identified start of withdrawal, intervals of surgical procedures, the endoscope exit period, and the total computed withdrawal time. Based on the results, we analyzed withdrawal time errors and experimented with applying post-processing to measure withdrawal time more accurately. Lastly, a Bland-Altman plot was used for comparison before and after post-processing.

Comparison methods. In the feature extraction learning process, to check the effectiveness of MagFace, a quality recognition metric learning that considers image quality for colonoscopy images, it was compared with Softmax and ArcFace [37] classifiers that do not consider image quality, both of which were ImageNet pre-trained ResNeSt50 [8] was used. To confirm the effect of applying a method that considers temporal information to colonoscopy images, a comparison was made between frame-wise prediction and applying a temporal relation network. To examine the effectiveness of the SAHC model, which has robust characteristics against errors from ambiguous images that appear in the middle of the video, in temporal relation network learning, a comparative experiment was conducted with MS-TCN [11], a popular phase recognition model.

Implementation details. For implementation, our experiments were conducted on PyTorch, leveraging 4 NVIDIA GeForce RTX 3090 GPUs. The dataset comprised 79 videos, each recorded at 30 FPS with a resolution of 1920×1080. Preprocessing involved cropping to the colon-visible regions and resizing to 256×256. For video action segmentation tasks, we reduced the frame rate from 30fps to 6fps. Hyperparameters were set in accordance with those specified in the original method papers.

Table 5. Quantitative result of video phase recognition.

Method	Accuracy (%)	Precision (%)	Recall (%)	Jaccard (%)
Softmax	92.8 ± 3.2	79.5 ± 9.5	70.6 ± 10.5	63.7 ± 9.0
Softmax + MS-TCN	93.7 ± 3.1	87.0 ± 6.5	79.7 ± 9.2	71.9 ± 8.2
Softmax + SAHC	94.0 ± 3.2	88.4 ± 6.4	80.8 ± 8.4	73.3 ± 8.6
ArcFace	92.7 ± 3.4	81.4 ± 8.8	68.8 ± 9.2	63.3 ± 8.1
ArcFace + MS-TCN	94.0 ± 3.3	86.4 ± 6.9	81.3 ± 6.9	73.8 ± 6.7
ArcFace + SAHC	94.5 ± 3.4	89.1 ± 7.4	83.2 ± 7.1	76.0 ± 6.8
MagFace	93.4 ± 3.0	80.4 ± 8.9	71.4 ± 9.0	65.3 ± 8.8
MagFace + MS-TCN	95.2 ± 2.6	90.0 ± 6.9	82.0 ± 8.9	76.7 ± 8.9
MagFace + SAHC	**96.0 ± 2.3**	**90.1 ± 6.4**	**87.9 ± 6.9**	**81.0 ± 7.9**

Video phase recognition results. Table 5 presents the phase recognition performance on a dataset comprising 39 colonoscopy videos, evaluating different learning methods. Training the feature extractor model with MagFace outperformed the use of Softmax and ArcFace across all metrics. In comparisons among frame-wise, MS-TCN, and SAHC, the performance ranking was observed to be in the order of SAHC, MS-TCN, and then frame-wise for the Softmax, ArcFace, and MagFace methods, respectively. These findings are further confirmed by the color-coded ribbon diagrams in **Figure 12**, which visualizes the phase recognition outcomes for two videos. **Figure 12** shows that errors in frame-wise predictions are gradually corrected when applying the temporal-relation model, and the SAHC method exhibits the most significant error refinement in both (a) and (b). The combined use of MagFace and SAHC for phase recognition yielded the best results, with an accuracy of 96.0 ± 2.3, precision of 90.1 ± 6.4, recall of 87.9 ± 6.9, and a Jaccard index of 81.0 ± 7.9.

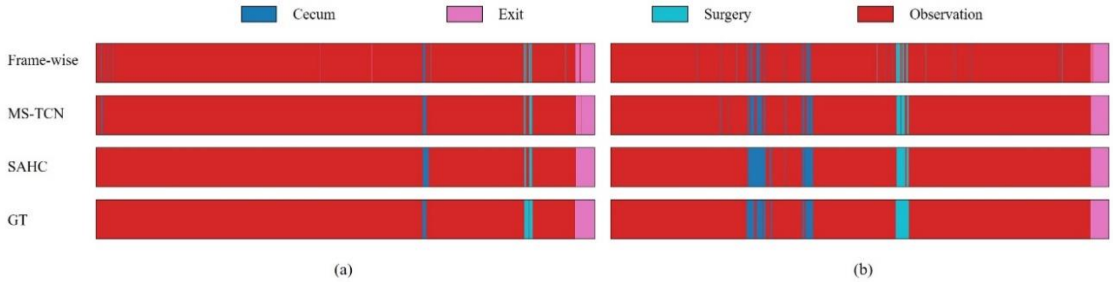


Figure 12. Color-coded ribbon visualizing phase recognition results, with different colors representing distinct phases of the colonoscopy procedure: Cecum (blue), Exit (pink), Surgery (cyan), and Observation (red).

Correlation between image magnitude and image quality. MagFace utilizes the magnitude of image features as an indicator of quality, with higher magnitudes signifying higher-quality images. **Figure 13** (a) shows a low-quality image with sizes between 10 and 20, which can be seen to have issues such as blurring and non-optimal lighting conditions. Conversely, **Figure 13** (b) displays images of higher quality, with magnitudes between 40 and 50, offering clearer and more detailed visual information.

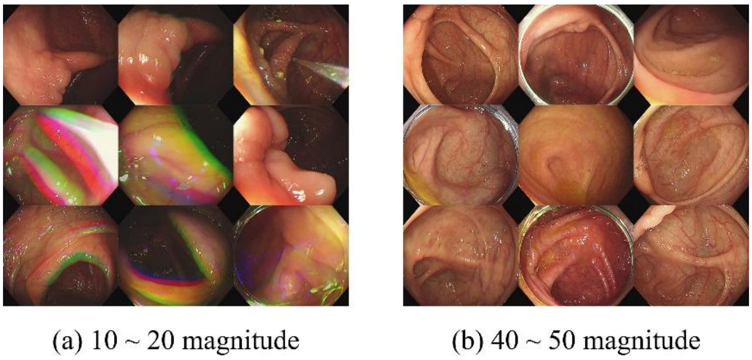
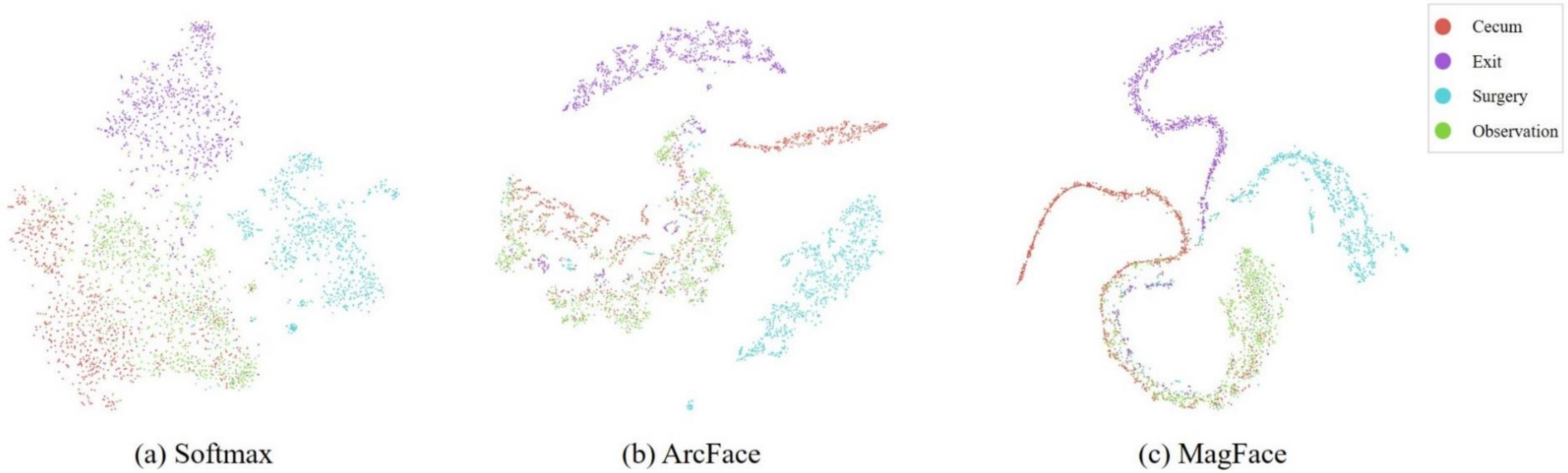


Figure 13. Display of cecum images from colonoscopy videos with MagFace's quality scores. (a) Exhibits images within a lower quality range, scored between 10 and 20 in magnitude, (b) presents images within a higher quality range, scored between 40 and 50 in magnitude.



(a) Softmax

(b) ArcFace

(c) MagFace

Figure 14. The t-SNE visualization of the feature spaces for four distinct phases in colonoscopy images, using a dataset of 1,000 images per phase to demonstrate the clustering capability of each method.

Visualization of feature clustering. Figure 14 presents t-SNE visualizations of high-dimensional feature spaces reduced to a 2D projection for ‘Cecum’, ‘Exit’, ‘Surgery’, and ‘Observation’ classes. Each figure corresponds to features extracted after learning using each learning method: (a) Softmax, (b) ArcFace, and (c) MagFace, with 1000 images per sampled class. (c) is (a) and (b) shows superior class separation, indicating a more distinct clustering of features.

Table 6. Quantitative result of withdrawal time evaluation.

Method	MAE			
	Start of Withdrawal	Surgery Duration	Endoscope Exit	Withdrawal Time
Softmax	163.98 ± 185.13	36.56 ± 39.75	4.05 ± 6.24	199.58 ± 181.27
Softmax + MS-TCN	53.5 ± 102.05	34.68 ± 40.07	4.9 ± 8.59	90.84 ± 105.12
Softmax + SAHC	33.69 ± 95.92	41.19 ± 44.83	2.36 ± 3.7	70.41 ± 102.95
ArcFace	147.62 ± 207.92	39.48 ± 44.25	3.15 ± 4.21	188.29 ± 201.44
ArcFace + MS-TCN	85.84 ± 195.38	24.76 ± 30.2	2.29 ± 3.24	111.08 ± 195.95
ArcFace + SAHC	42.06 ± 91.9	27.52 ± 36.5	**2.01 ± 2.9**	67.87 ± 104.86
MagFace	247.03 ± 182.39	30.58 ± 37.99	3.23 ± 4.74	278.4 ± 179.39
MagFace + MS-TCN	55.3 ± 114.93	24.73 ± 33.33	2.9 ± 4.42	80.52 ± 120.01
MagFace + SAHC	**22.36 ± 61.07**	**22.65 ± 30.62**	2.25 ± 3.52	**42.84 ± 66.47**

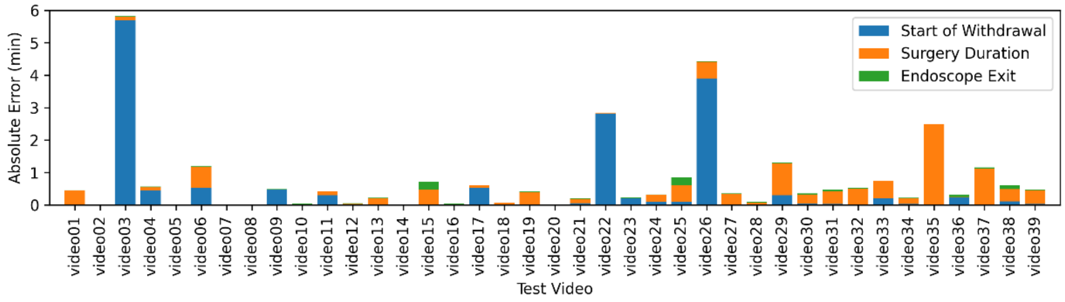


Figure 15. Stack bar graph showing the absolute errors in withdrawal time measurements across all 39 test videos. The errors are compared between the times predicted by the model that integrates MagFace with SAHC and the times measured by physicians. This visual representation highlights the 'Start of Withdrawal', 'Surgery Duration', and 'Endoscope Exit' for each video.

Withdrawal time results. Table 6 displays the results for the evaluation of withdrawal time. The table shows the differences between the withdrawal times measured by physicians and those predicted by the models. Similar to the video phase recognition results, it can be seen that errors are reduced when MagFace is used with a temporal relationship model. The combination of MagFace and SAHC shows the smallest error margin, with the start of withdrawal (Start of Withdrawal) showing an error of 22.36 ± 61.07, the surgery duration (Surgery Duration) an error of 22.65 ± 30.62, and the total withdrawal time (Withdrawal Time) an error of 42.84 ± 66.47, indicating the closest correlation to the withdrawal times calculated by physicians. **Figure 15** illustrates a stack bar graph representing the absolute error between the withdrawal time measurements of all 39 test videos by the model combining MagFace with SAHC and the measurements by physicians. This graph visually exhibits the errors for each test video concerning 'Start of Withdrawal', 'Surgery Duration', and 'Endoscope Exit'. Consistent with the results from **Table 6**, the overall errors are under one minute; however, some cases exhibit relatively large errors. **Figure 16** displays a color-coded ribbon visualization of phase recognition result for the video with the largest withdrawal time error, accompanied by a confidence score chart. A segment of misclassification is highlighted with a yellow circle, where corresponding confidence scores near 0.5 suggest uncertainty. These issues are addressed in the post-processing section that follows.

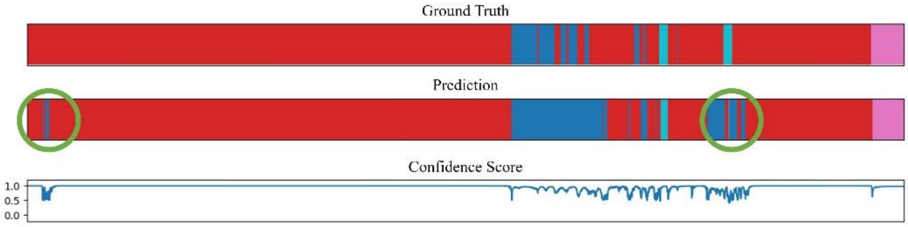


Figure 16. Color-coded ribbon visualization of the phase recognition for the video with the largest withdrawal time error, including confidence score chart. The yellow circle highlights a segment of misclassification, with corresponding confidence scores nearing 0.5, indicating uncertainty.

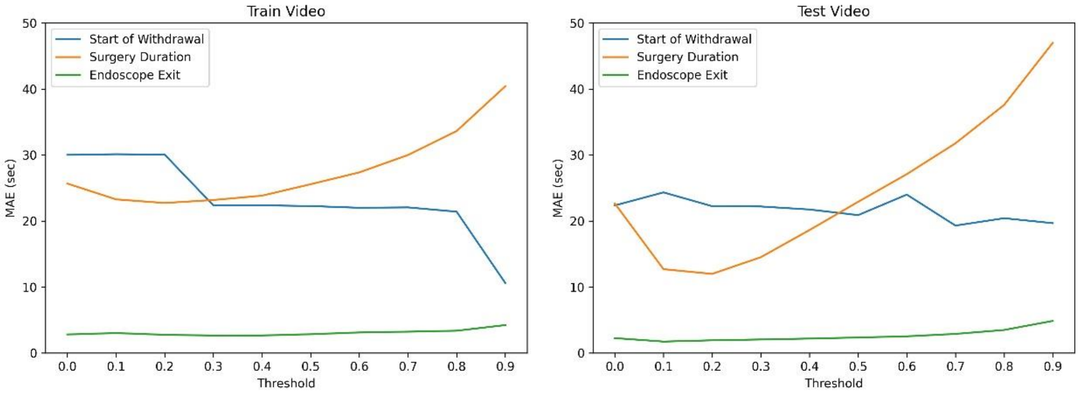


Figure 17. Mean Absolute Error (MAE) plot for 'Start of Withdrawal', 'Surgery Duration', and 'Endoscope Exit' across varying threshold levels in both training and test videos.

Post-processing. To address the errors identified in **Figures 15** and **16**, we conducted post-processing on the 'Start of Withdrawal', 'Surgery Duration', and 'Endoscope Exit' phases by applying a Gaussian filter to the confidence scores of each phase. Threshold values were determined using k-fold cross-validation on the training data to avoid overfitting when applied to the test data. **Figure 17** illustrates the effect of incremental threshold adjustments, ranging from 0 to 0.9, on the Mean Absolute Error (MAE) for both training and test datasets. For the 'Start of Withdrawal' phase, incrementally higher thresholds correlated with a consistent decrease in error across both sets of videos. The 'Surgery Duration' phase observed a decrease in error up to a threshold of 0.2, beyond which the error began to rise. The 'Endoscope Exit' phase experienced negligible changes in error across various thresholds. These optimal thresholds, identified through the training data analysis, were then applied to the test videos.

Evaluation of post-processing. Table 7 displays the MAE results for each phase post-processing: 'Start of Withdrawal' at 19.67 ± 43.47, 'Surgery Duration' at 13.29 ± 19.46, 'Endoscope Exit' at 2.07 ± 3.13, and overall 'Withdrawal Time' at 28.29 ± 49.07. All phases showed a reduction in error after post-processing. **Figure 18** illustrates the Bland-Altman plots comparing the predicted withdrawal times by the model to the ground truth, before and after post-processing. The mean difference between the predictions and actual times reduced from 0.65 minutes pre-post-processing to 0.06 minutes post-post-processing, indicating a improvement in prediction accuracy. Furthermore, the standard deviation decreased, which suggests a more consistent performance of the model. The majority of data points lie within the ±1.96 SD limit in both plots, with fewer outliers observed post-post-processing.

Table 7. Post-processing result using confidence score.

	MAE			
	Start of Withdrawal	Surgery Duration	Endoscope Exit	Withdrawal Time
w/o post-processing	22.36 ± 61.07	22.65 ± 30.62	2.25 ± 3.52	42.84 ± 66.47
w post-processing	**19.67 ± 43.47**	**13.29 ± 19.46**	**2.07 ± 3.13**	**28.29 ± 49.07**

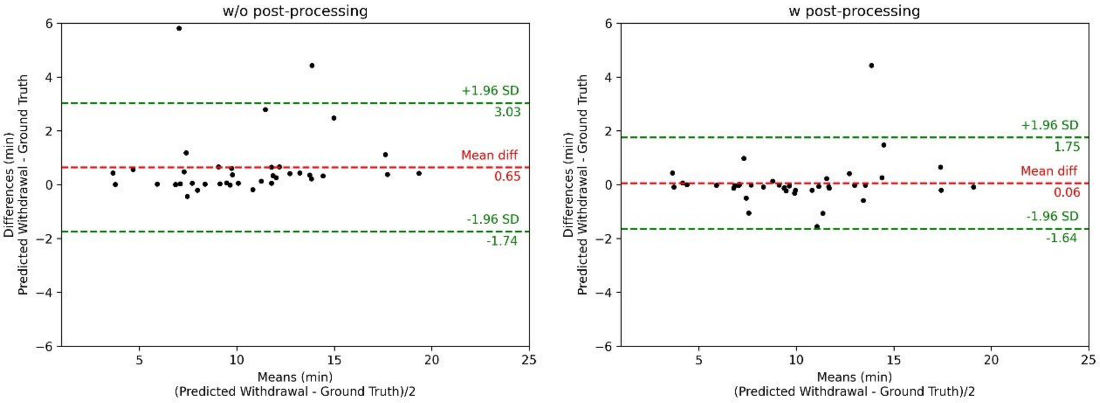


Figure 18. Bland-Altman plots comparing predicted and actual withdrawal times before and after post-processing.

4. Discussion

In this study, we introduced a method to quantify withdrawal time on colonoscopy video. By considering colonoscopy characteristics such as peristaltic movement of the large intestine and light reflection, we utilized quality-aware metric learning and phase recognition method that is robust to noise. Through **Figure 13**, we demonstrate that when applying MagFace to colonoscopy images, the feature magnitude indeed serves as a reliable indicator of image quality, enabling the model to learn differentially based on image quality. Furthermore, **Figure 12** reveals that the robust features of SAHC enable more precise phase prediction, even in the presence of ambiguous frames within colonoscopy video segments. It can be confirmed that the introduction of this method is effective in measuring phase recognition and withdrawal time for colonoscopy video (**Table 5** and **6**). These results suggest that considering disruptive elements (such as low-quality and ambiguous images) when training models is an effective strategy for analyzing colonoscopy videos. However, in **Figure 15**, you can see that large errors still appear in some cases. Although we have greatly improved the error problem by applying post-processing to the confidence scores of each phase (**Table 7** and **Figure 18**), we believe that more analysis and performance improvement are needed for actual clinical use. We anticipate that further refining these technologies for practical clinical application could enable more objective quality evaluations and ultimately enhance the quality of patient care.

Conclusion

In this study, three experiments were performed to analyze medical videos, especially tympanomastectomy (TM) surgical video and colonoscopy video. In the first study, we proposed a TM surgical phase recognition learning method utilizing Teacher-Student learning and confirmed that performance was improved in situations of class imbalance and scarcity of data In the second study, we proposed a method using implicit neural representation that does not require large-scale video data collection and enables high-resolution video inpainting without memory constraints, and showed improved performance in visual comparison with the state-of-the-art model. In the third study, we proposed a method that integrates quality-aware metric learning and a noise-robust phase recognition model by considering the characteristics of the colon, enabling more accurate quantification in colonoscopy. We believe that this methodology has the potential to be applied not only to TM surgery videos and colonoscopy videos, but also to a broader range of areas within medical video analysis.

References

[1] Oprea, Sergiu, et al. "A review on deep learning techniques for video prediction." IEEE Transactions on Pattern Analysis and Machine Intelligence 44.6 (2020): 2806-2826.

[2] Jiao, Licheng, et al. "New generation deep learning for video object detection: A survey." IEEE Transactions on Neural Networks and Learning Systems 33.8 (2021): 3195-3215.

[3] de Azevedo, Alexandre Fernandes, et al. "Tympanomastoidectomy: Comparison between canal wall-down and canal wall-up techniques in surgery for chronic otitis media." International archives of otorhinolaryngology 17.03 (2013): 242-245.

[4] Blum, Tobias, Hubertus Feußner, and Nassir Navab. "Modeling and segmentation of surgical workflow from laparoscopic video." Medical Image Computing and Computer-Assisted Intervention–MICCAI 2010: 13th International Conference, Beijing, China, September 20-24, 2010, Proceedings, Part III 13. Springer Berlin Heidelberg, 2010.

[5] Zappella, Luca, et al. "Surgical gesture classification from video and kinematic data." Medical image analysis 17.7 (2013): 732-745.

[6] Lalys, Florent, et al. "A framework for the recognition of high-level surgical tasks from video images for cataract surgeries." IEEE Transactions on Biomedical Engineering 59.4 (2011): 966-976.

[7] He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[8] Zhang, Hang, et al. "Resnest: Split-attention networks." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.

[9] Sherstinsky, Alex. "Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network." Physica D: Nonlinear Phenomena 404 (2020): 132306.

[10] Lea, Colin, et al. "Temporal convolutional networks for action segmentation and detection." proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017.

[11] Farha, Yazan Abu, and Jurgen Gall. "Ms-tcn: Multi-stage temporal convolutional network for action segmentation." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.

[12] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

[13] Jin, Yueming, et al. "SV-RCNet: workflow recognition from surgical videos using recurrent convolutional network." IEEE transactions on medical imaging 37.5 (2017): 1114-1126.

[14] Czempiel, Tobias, et al. "Tecno: Surgical phase recognition with multi-stage temporal convolutional networks." Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part III 23. Springer International Publishing, 2020.

[15] Jin, Yueming, et al. "Temporal memory relation network for workflow recognition from surgical video." IEEE Transactions on Medical Imaging 40.7 (2021): 1911-1923.

[16] Yi, Fangqiu, Yanfeng Yang, and Tingting Jiang. "Not end-to-end: Explore multi-stage architecture for online surgical phase recognition." Proceedings of the Asian Conference on Computer Vision. 2022.

[17] Ding, Xinpeng, and Xiaomeng Li. "Exploring segment-level semantics for online phase recognition from surgical videos." IEEE Transactions on Medical Imaging 41.11 (2022): 3309-3319.

[18] Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean. "Distilling the knowledge in a neural network." arXiv preprint arXiv:1503.02531 (2015).

[19] Xu, Guodong, et al. "Knowledge distillation meets self-supervision." European Conference on Computer Vision. Cham: Springer International Publishing, 2020.

[20] Abbasi Koohpayegani, Soroush, Ajinkya Tejankar, and Hamed Pirsiavash. "Compress: Self-supervised learning by compressing representations." Advances in Neural Information Processing Systems 33 (2020): 12980-12992.

[21] Ye, Fei, and Adrian G. Bors. "Dynamic Self-Supervised Teacher-Student Network Learning." IEEE Transactions on Pattern Analysis and Machine Intelligence 45.5 (2022): 5731-5748.

[22] Ke, Zhanghan, et al. "Dual student: Breaking the limits of the teacher in semi-supervised learning." Proceedings of the IEEE/CVF international conference on computer vision. 2019.

[23] Huo, Xinyue, et al. "ATSO: Asynchronous teacher-student optimization for semi-supervised

image segmentation." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021.

[24] Yang, Lei, et al. "Mix-teaching: A simple, unified and effective semi-supervised learning framework for monocular 3d object detection." IEEE Transactions on Circuits and Systems for Video Technology (2023).

[25] Yu, Tong, et al. "Learning from a tiny dataset of manual annotations: a teacher/student approach for surgical phase recognition." arXiv preprint arXiv:1812.00033 (2018).

[26] Teevno, Mansoor Ali, Gilberto Ochoa-Ruiz, and Sharib Ali. "A semi-supervised Teacher-Student framework for surgical tool detection and localization." Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization 11.4 (2023): 1033-1041.

[27] Yamlahi, Amine, et al. "Self-distillation for surgical action recognition." International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer Nature Switzerland, 2023.

[28] Chang, Ya-Liang, et al. "Free-form video inpainting with 3d gated convolution and temporal patchgan." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019.

[29] Zeng, Yanhong, Jianlong Fu, and Hongyang Chao. "Learning joint spatial-temporal transformations for video inpainting." Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16. Springer International Publishing, 2020.

[30] Li, Zhen, et al. "Towards an end-to-end framework for flow-guided video inpainting." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.

[31] Sitzmann, Vincent, et al. "Implicit neural representations with periodic activation functions." Advances in neural information processing systems 33 (2020): 7462-7473.

[32] Chen, Zeyuan, et al. "Videoinr: Learning video implicit neural representation for continuous space-time super-resolution." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

[33] Kim, Subin, et al. "Scalable neural video representations with learnable positional

features." Advances in Neural Information Processing Systems 35 (2022): 12718-12731.

[34] Attal, Benjamin, et al. "HyperReel: High-fidelity 6-DoF video with ray-conditioned sampling." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

[35] Park, Jeong Joon, et al. "Deepsdf: Learning continuous signed distance functions for shape representation." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.

[36] Mildenhall, Ben, et al. "Nerf: Representing scenes as neural radiance fields for view synthesis." Communications of the ACM 65.1 (2021): 99-106.

[37] Deng, Jiankang, et al. "Arcface: Additive angular margin loss for deep face recognition." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.

[38] Meng, Qiang, et al. "Magface: A universal representation for face recognition and quality assessment." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021.

[39] Kim, Minchul, Anil K. Jain, and Xiaoming Liu. "Adaface: Quality adaptive margin for face recognition." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.

[40] Choi, Joonmyeong, et al. "Video recognition of simple mastoidectomy using convolutional neural networks: Detection and segmentation of surgical tools and anatomical regions." Computer Methods and Programs in Biomedicine 208 (2021): 106251.

[41] Demir, Kubilay Can, et al. "Deep Learning in Surgical Workflow Analysis: A Review of Phase and Step Recognition." (2023).

[42] Buda, Mateusz, Atsuto Maki, and Maciej A. Mazurowski. "A systematic study of the class imbalance problem in convolutional neural networks." Neural networks 106 (2018): 249-259.

[43] Johnson, Justin M., and Taghi M. Khoshgoftaar. "Survey on deep learning with class imbalance." Journal of Big Data 6.1 (2019): 1-54.

[44] Loshchilov, Ilya, and Frank Hutter. "Decoupled weight decay regularization." arXiv preprint arXiv:1711.05101 (2017).

[45] Loshchilov, Ilya, and Frank Hutter. "Sgdr: Stochastic gradient descent with warm

restarts." arXiv preprint arXiv:1608.03983 (2016).

[46] Buslaev, Alexander, et al. "Albumentations: fast and flexible image augmentations." Information 11.2 (2020): 125.

[47] Ying, Xue. "An overview of overfitting and its solutions." Journal of physics: Conference series. Vol. 1168. IOP Publishing, 2019.

[48] Ouyang, Hao, Tengfei Wang, and Qifeng Chen. "Internal video inpainting by implicit long-range propagation." Proceedings of the IEEE/CVF international conference on computer vision. 2021.

[49] Zhang, Haotian, et al. "An internal learning approach to video inpainting." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019.

[50] Wang, Yuehao, et al. "Neural rendering for stereo 3d reconstruction of deformable tissues in robotic surgery." International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer Nature Switzerland, 2022.

[51] Rex, Douglas K., et al. "Quality indicators for colonoscopy." Official journal of the American College of Gastroenterology| ACG 110.1 (2015): 72-90.

[52] Simmons, Dia T., et al. "Impact of endoscopist withdrawal speed on polyp yield: implications for optimal colonoscopy withdrawal time." Alimentary pharmacology & therapeutics 24.6 (2006): 965-971.

[53] Moritz, V., et al. "Withdrawal time as a quality indicator for colonoscopy–a nationwide analysis." Endoscopy 44.05 (2012): 476-481.

[54] Shaukat, Aasma, et al. "Longer withdrawal time is associated with a reduced incidence of interval cancer after screening colonoscopy." Gastroenterology 149.4 (2015): 952-957.

[55] Jin, Bo, et al. "Diagnosing Parkinson disease through facial expression recognition: video analysis." Journal of medical Internet research 22.7 (2020): e18697.

[56] Liu, Shengfeng, et al. "Deep learning in medical ultrasound analysis: a review." Engineering 5.2 (2019): 261-275.

[57] Quellec, Gwenolé, et al. "Multiple-instance learning for medical image and video analysis." IEEE reviews in biomedical engineering 10 (2017): 213-234.

[58] Van der Maaten, Laurens, and Geoffrey Hinton. "Visualizing data using t-SNE." Journal of machine learning research 9.11 (2008).

Abstract (with Korean)

이 연구는 의료 비디오 분석을 위한 딥 러닝 기술을 탐구합니다. 현재 딥러닝 기술들을 의료 분야에 적용하는 것에는 데이터 수집의 어려움과 자연 비디오와는 다른 의료 비디오만의 독특한 특성으로 인해 몇 가지 도전과제를 안고 있습니다. 본 논문은 지식 증류(Knowledge Distillation), 암시적 신경 표현(Implicit Neural Representation), 그리고 딥 거리 학습(Deep Metric Learning)과 같은 고급 딥 러닝 기술을 의료 비디오 데이터 분석에 적용하여 이러한 도전과제들을 해결합니다. 특히, 이 연구는 유양돌기 절제 수술 단계 인식 및 비디오 인페인팅 기법과 대장내시경 비디오 데이터의 정량화 방법에 중점을 둡니다.

이 연구에서는 의료 비디오를 분석하기 위해 세 가지 실험이 수행되었습니다. 첫 번째 연구는 교사-학생 학습을 사용하여 유양돌기 절제 수술 단계 인식 학습 방법을 제안했으며, 클래스 불균형 및 데이터 부족 상황에서 성능이 향상된 것을 보여주었습니다. 두 번째 연구는 암묵적 신경 표현 방법을 활용하여 대규모 비디오 데이터 수집이 없이 고해상도 비디오 인페인팅을 하는 방법을 소개했으며, 기존 최신 모델과 비교하여 시각적으로 성능이 향상된 것을 확인합니다. 세 번째 연구는 품질 인식 거리 학습과 노이즈에 강한 단계 인식 모델을 결합한 비디오 인식 방법을 제안하여 대장의 특성을 고려함으로써 대장내시경에서 보다 정확한 정량화를 가능하게 했습니다.

우리는 이 연구에서 개발된 방법론들이 유양돌기 절제 수술 및 대장내시경 비디오뿐만 아니라 더 넓은 범위의 의료 비디오 분석 분야에도 적용될 수 있을 것이라고 믿습니다.