# Weakly Supervised Temporal Convolutional Networks for Fine-Grained Surgical Activity Recognition

Sanat Ramesh, Diego Dall'Alba, Cristians Gonzalez, Tong Yu, Pietro Mascagni, Didier Mutter, Jacques Marescaux, Paolo Fiorini, *Life Fellow, IEEE*, and Nicolas Padoy

***Abstract*— Automatic recognition of fine-grained surgical activities, called steps, is a challenging but crucial task for intelligent intra-operative computer assistance. The development of current vision-based activity recognition methods relies heavily on a high volume of manually annotated data. This data is difficult and time-consuming to generate and requires domain-specific knowledge. In this work, we propose to use coarser and easier-to-annotate activity labels, namely phases, as weak supervision to learn step recognition with fewer step annotated videos. We introduce a step-phase dependency loss to exploit the weak supervision signal. We then employ a Single-Stage Temporal Convolutional Network (SS-TCN) with a ResNet-50 backbone, trained in an end-to-end fashion from weakly annotated videos, for temporal activity segmentation and recognition. We extensively evaluate and show the effectiveness of the proposed method on a large video dataset consisting of 40 laparoscopic gastric bypass procedures and the public benchmark CATARACTS containing 50 cataract surgeries.**

Manuscript received 14 September 2022; revised 5 December 2022; accepted 21 March 2023. Date of publication 29 March 2023; date of current version 31 August 2023. This work was supported in part by the European Union's Horizon 2020 Research and Innovation Program under the Marie Sklodowska-Curie Grant through the Project AuTonomous intraLuminAl Surgery (ATLAS) under Grant 813782, in part by the French State Funds managed within the Investissements d'Avenir Program by Banque Publique d'Investissement (BPI) France Project Connected Optimized Network and Data in Operating Rooms (CONDOR), and in part by the ANR under Grant ANR-16-CE33-0009 and Grant ANR-10-IAHU-02. *(Corresponding author: Sanat Ramesh.)*

Sanat Ramesh is with the Altair Robotics Laboratory, Department of Computer Science, University of Verona, 37134 Verona, Italy, and also with the ICube Laboratory, CNRS, IHU Strasbourg, University of Strasbourg, 67000 Strasbourg, France (e-mail: sanat.ramesh@univr.it).

Diego Dall'Alba and Paolo Fiorini are with the Altair Robotics Laboratory, Department of Computer Science, University of Verona, 37134 Verona, Italy (e-mail: diego.dallalba@univr.it; paolo.fiorini@univr.it).

Cristians Gonzalez is with the IHU Strasbourg, University Hospital of Strasbourg, 67000 Strasbourg, France (e-mail: cristians.gonzalez@ihu-strasbourg.eu).

Tong Yu and Nicolas Padoy are with the ICube Laboratory, CNRS, IHU Strasbourg, University of Strasbourg, 67000 Strasbourg, France (e-mail: tyu@unistra.fr; npadoy@unistra.fr).

Pietro Mascagni is with the IRCCS, Fondazione Policlinico Universitario Agostino Gemelli, 00168 Rome, Italy, and also with the ICube Laboratory, CNRS, IHU Strasbourg, University of Strasbourg, 67000 Strasbourg, France (e-mail: p.mascagni@unistra.fr).

Didier Mutter is with the IHU Strasbourg, University Hospital of Strasbourg, 67000 Strasbourg, France, and also with the IRCAD, 67000 Strasbourg, France (e-mail: didier.mutter@chru-strasbourg.fr).

Jacques Marescaux is with the IRCAD, 67000 Strasbourg, France (e-mail: jacques.marescaux@ircad.fr).

Digital Object Identifier 10.1109/TMI.2023.3262847

***Index Terms*— Endoscopic videos, surgical step recognition, temporal convolutional networks, weak supervision, gastric bypass procedures, cataracts procedures.**

## I. Introduction

RESEARCH in developing advanced clinical decision support systems in computer-assisted interventions (CAI) and robot-assisted surgeries (RAS) for the demanding situations of a modern Operating Room (OR) [1], [2], [3] has seen significant progress in the last decade. One of the primary functions of these advanced systems is automatic surgical workflow analysis, i.e., reliable recognition of the current surgical activities. Surgical activity recognition could play a key role in assisting clinical decisions, report generation, and data annotation by providing valuable semantic information.

Depending on the level of granularity, a surgical procedure can be decomposed into activities, such as the whole procedure, phases, stages, steps, and actions [4], [5]. Surgical phases are defined as a set of fundamental surgical aims to accomplish in order to successfully complete the surgical procedure. Similarly, steps are defined as a set of surgical actions to perform in order to accomplish a surgical phase. These definitions help clinicians define an ontology for each procedure, e.g. [6], [7] define ontologies for cataract and gastric bypass procedures. Although the ontologies are well defined, automatically recognizing these activities from available endoscopic videos is a topic of high interest.

Phase recognition has received a lot of attention and is a very active area of research in the medical computer vision community [8], [9], [10], [11], [12]. Alongside phases, there has been substantial research focusing on fine-grained activities such as robotic gestures [13], [14], [15], [16], [17], [18], [19], action triplets [20], and instrument detection and tracking [11], [21], [22]. Recently, there has been a aof research works focusing particularly on step recognition [6], [7], [23].

While steps define a surgical workflow at a more fine-grained level than phases, the time required to annotate a dataset with steps is significantly higher than with phase annotations. For example, in Laparoscopic Roux-en-Y gastric bypass (LRYGB) procedures, the workflow consists of 44 steps and 11 phases (Table II). Precisely defining and annotating all the steps requires a considerably higher time


Fig. 1. Sample images from Bypass40 and CATARACTS datasets. Each column of Bypass40 images present similar steps.

of experts due to the number of steps and more importantly lower inter-class variances between steps. Since recent works in surgical phase/step recognition employ deep learning models, they rely on the availability of large-scale annotated datasets. Curation of these annotated datasets is difficult and time-consuming as these tasks require domain-specific medical knowledge.

To address this issue, a few works [24], [25], [26], [27] have proposed methods based on semi-supervision. These approaches involve either pre-training the model on proxy tasks or training on synthetic labels generated by a teacher model trained on a small subset for phase recognition. Unlike these works, inspired by [22] and [28], we address the annotation scarcity issue by proposing a weakly supervised learning approach utilizing relatively economical annotations.

The main contributions of our work are summarized as follows:

1) We propose a weakly supervised learning method for surgical workflow analysis to tackle the problem of fine-grained surgical activity (step) recognition. We exploit the hierarchical step-phase relationships and utilize easier-to-annotate weak phase annotations on videos with missing step annotations.
2) We introduce a novel dependency loss to enforce the weak supervision and encode the step-phase hierarchical relationship as a matrix. By optimizing for this loss, it encourages the model to learn possible step sequences and transitions from videos with only phase annotations.
3) We present an end-to-end model consisting of ResNet-50 and Single-Stage Temporal Convolutional Network (SS-TCN) to learn both visual and temporal cues jointly.
4) We extend the CATARACTS[1] dataset (containing step annotations) with phase annotations. These annotations will be released upon acceptance of this manuscript.
5) We extensively evaluate our approach on two surgical video datasets, namely Bypass40 [7] and CATARACTS [29], demonstrating the effectiveness and generalizability of our method.

[1]https://cataracts2020.grand-challenge.org/

## II. Related Work

### A. Surgical Activity Recognition

Research on developing deep learning methods for surgical phase recognition has seen significant progress with initial works of EndoNet [8] and DeepPhase [9] on cholecystectomy and cataract surgeries, respectively. EndoNet proposed a Convolutional Neural Network (CNN) followed by a hierarchical Hidden Markov Model (HMM) to perform both phase and tool detection. Similarly, DeepPhase introduced an architecture with ResNet [30] and Recurrent Neural Network (RNN), instead of HMMs, for temporal modeling, for both phase recognition and tool detection. EndoLSTM [31], [32] extended EndoNet by utilizing a Long Short-Term Memory (LSTM) for temporal refinement of spatial features. Similarly, SV-RCNet [10] trained a ResNet and LSTM model end-to-end and proposed a prior knowledge inference scheme for surgical phase recognition. MTRCNet-CL [11] presented a multi-task model to detect tool presence and perform phase recognition along with a novel correlation loss to capture the relationship between tool presence and phase identification. Recently, TeCNO [12] adapted the multi-stage Temporal Convolutional Network (MS-TCN) [33] architecture for online surgical phase prediction by implementing causal convolutions [34].

On the other hand, step recognition has seen a spark in research with the initial work of [23]. A Content-Based Video Retrieval (CBVR) system, for real-time step recognition, was proposed utilizing a novel pupil center and scale tracking method as pre-processing of motion features. In [6], the CBVR system along with surgical tool presence information was used as input to statistical models consisting of Bayesian Network and HMMs for multi-level online recognition of step and phase. Recently, MTMS-TCN [7] adapted TeCNO utilizing TCNs for multi-level online recognition of step and phase. In this work, we build upon the architectures of TeCNO and MTMS-TCN by utilizing a variant of MS-TCN in an end-to-end fashion for online step recognition.

### B. Weak Supervision

Weak supervision has seen a great interest in the medical computer vision community to tackle the need for high-volume annotated datasets that are difficult to generate. Some of the interesting applications of weak supervision are seen in surgical tool localization [22], tool segmentation [28], cancerous tissue segmentation [35], and detection of the region of interest in chest X-rays and mammograms [36]. To reduce the number of labeled videos, most of the recent research works in phase recognition have proposed approaches based on semi-supervised learning. These approaches follow a similar strategy of pre-training the models on different proxy tasks of frame-sorting [24], predicting the temporal distance between multiple frames [25], and predicting the remaining surgery duration [26]. The most closely related work to this paper in terms of objectives is [27], which proposed a teacher/student approach for phase recognition in scenarios of extreme manual annotation scarcity ($\leq$25% of the training set). The teacher model (trained on a small set) generated synthetic phase annotations for a large number of videos on which the student model was then trained.

Weakly supervised coarse-to-fine methods have received considerable interest in the computer vision community [37], [38], [39] for image classification. Reference [37] proposed an image-based weakly supervised end-to-end model for object classification consisting of a CNN followed by two self-expressive layers. One self-expressive layer captures the global structures through coarse labels and the other captures the local structures for fine-grained classification. Reference [38] tackled the problem of learning finer representations from coarser labels without any fine-grained labels. Their proposed method consists of CNN based trunk-target network that learns coarse representations from labels and finer representations with nearest-neighbor classifier objective. Recently, [39] tackled the problem of Coarse-to-Fine Few-Shot (C2FS) and proposed a novel 'angular normalization' module that effectively combines supervised and self-supervised contrastive pre-training for C2FS.

Although these previous works in the vision community propose weakly supervised learning methods exploiting hierarchical structures, the focus solely lies on object recognition in natural images containing a single object in each image. In this work, we focus on weakly supervised learning from videos instead of images. We aim to recognize fine-grained activity, as opposed to object, exploiting the temporal information available in videos. In particular, we target fine-grained surgical activity recognition on videos from endoscopic procedures on two different types of surgeries, i.e., gastric bypass and cataract.

## III. Methodology

The overview of our proposed method is presented in Fig. 2. In this section, we first present our end-to-end Spatio-temporal (ResNet-50 + SS-TCN) model for the task of fine-grained activity, i.e, step, recognition. Then we introduce the phase-step dependency loss for weak supervision of step recognition using phase annotation.

### A. Spatio-Temporal Model

Our weakly supervised step recognition network consists of a ResNet-50 model for visual feature extraction followed by an SS-TCN for modeling the recognition problem temporally. The complete model is trained in an end-to-end fashion. The overview of the model setup is depicted in Fig. 2.

For phase segmentation, ResNet-50 [40] has been successfully employed as the backbone in many previous works [10], [11], [12], [27]. In this work, we utilize the same architecture for visual feature extraction. We use a single-stage TCN (SS-TCN), a single-stage variant of MS-TCN, to learn the spatial coherence across video frames. The choice of SS-TCN was motivated by the work of [7] where MS-TCN did not provide a significant improvement over SS-TCN for both the step and phase recognition. Following the design of MS-TCN, the SS-TCN contains neither pooling layers nor fully connected layers and is constructed with only temporal convolutional layers, specifically dilated residual layers performing dilated

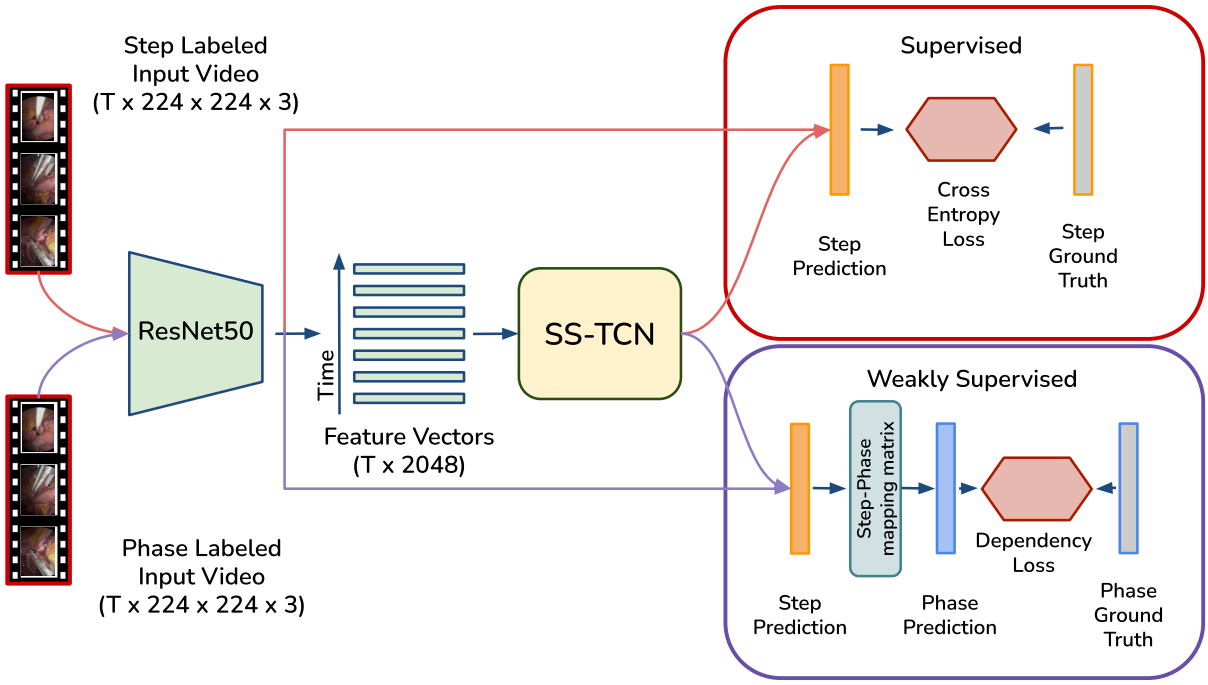



Fig. 2. Overview of our end-to-end Spatio-temporal model setup: ResNet50 + SS-TCN (Single-Stage Temporal Convolutional Networks). When step labels are available, the model is trained through the supervised pathway (red) and weakly supervised pathway (purple) utilizing phase labels. The model is trained end-to-end in a single learning stage.

convolutions. With the aim of online activity segmentation, we perform at each layer causal convolutions [7], [12], [34] that depend only on the current frame and $n$ previous frames.

The complete model takes an input video consisting of $T$ frames $x_{1:T}$. The ResNet-50 maps $224 \times 224 \times 3$ RGB images to a feature space of size $N_f = 2048$. These frame-wise features are collected over time and are inputs to the TCN model that predicts $\hat{y}^s_{1:T}$ where $\hat{y}^s_t$ is the class label for the current timestamp $t$, $t \in [1, T]$. Since step recognition is a multi-class classification problem that exhibits an imbalance in the class distribution, softmax activation and class-weighted cross-entropy loss are utilized. Additionally, the dependency loss used when step labels are not available also relies on softmax activation and weighted cross-entropy loss, utilizing phase labels instead. The class weights for both steps and phases are calculated using the median frequency balancing [41] on the training set. The total loss is given by:

$$\mathcal{L}_{total} = \delta_{step} \cdot \mathcal{L}_{step} + (1 - \delta_{step}) \cdot \mathcal{L}_{dep}, \qquad (1)$$

where $\mathcal{L}_{step}$ represents weighted cross-entropy loss for steps, $\mathcal{L}_{dep}$ is the step-phase dependency loss (subsection III-B), and $\delta_{step}$ is a binary variable that indicates if the video contains step labels.

### B. Weak Supervision: Step-Phase Dependency Loss

Steps and phases are two types of activities describing the surgical workflow that are defined at different levels of granularity and possess an inherent hierarchical relationship [4], [7]. Steps are defined at a higher level of detail compared to phases. This brings about lower inter-class variances between steps, compared to phases, making it a more complex task to clearly define and distinguish between them. The challenges can be seen in the sample images presented in Fig. 1. For instance, in the Bypass40 dataset, similar actions are performed across different steps belonging to different phases. Dissection is performed in at least 7 steps spread across 3 different phases. Similarly, Stapling is performed in 5 steps across 4 different phases. Designing and training a deep learning model to distinguish between these similar steps poses a great challenge. Even the state-of-the-art method, MTMS-TCN [7], trained on a fully annotated dataset achieves an accuracy of ~76% with a precision of ~56%, accentuating the difficulty of the problem. The class imbalance further creates a challenge for training deep learning models that require large datasets with plenty of samples for each class.

In the scenario presented in this paper where the number of annotations is scarce, the recognition difficulties increase drastically. To overcome some of the challenges, this work proposes a weakly supervised approach that utilizes labels of less granular activities, i.e., phases. Phase information alone could help the model in two ways. Firstly, phase information could help the model reduce errors related to recognizing similar looking steps, e.g., 'S6: horizontal stapling' and 'S18: gastrojejunal stapling', belonging to two different phases. Secondly, we can gather a smaller subset of probable steps that could occur in a given phase eliminating the rest. For example, given the phase to be 'Phacoemulsification' of cataract surgery, only 5 out of 19 steps are likely to occur (Table I). Similarly, a phase such as 'P5: anastomosis test' in the Bypass40 dataset, reduces the possible steps to 7 out of 44 (Table II). Here, the phase information provides cues to the model to learn to distinguish between steps belonging to the subset rather than the whole set. Thus we hypothesize that the additional

TABLE I
PHASES AND STEPS FOR THE CATARACT PROCEDURE

| Phases | Idle | Opening | Phacoemulsification | Implantation | Closure |
|---|---|---|---|---|---|
| Steps | Idle | Idle<br>Toric Marking<br>Implant Ejection<br>Incision<br>Viscodilatation<br>Capsulorhexis<br>Hydrodissection | Idle<br>Nucleus Breaking<br>Phacoemulsification<br>Vitrectomy<br>Irrigation/Aspiration | Idle<br>Incision<br>Viscodilatation<br>Preparing Implant<br>Manual Aspiration<br>Implantation<br>Positioning<br>OVD Aspiration | Idle<br>Suturing<br>Sealing Control<br>Wound Hydration |

TABLE II
PHASES AND STEPS FOR THE LAPAROSCOPIC RUE-EN-Y GASTRIC BYPASS PROCEDURE

| Phases | Steps |
|---|---|
| P1: preparation | S0: null step, S1: cavity exploration, S2: trocar placement, S3: retractor placement, S14: adhesiolysis, S22: gastric tube placement |
| P2: gastric pouch creation | S0: null step, S4: crura dissection, S5: his angle dissection, S6: horizontal stapling, S7: retrogastric dissection, S8: vertical stapling, S9: gastric remnant reinforcement, S10: gastric pouch reinforcement, S11: gastric opening, S22: gastric tube placement, S43: calibration |
| P3: omentum division | S0: null step, S12: omental lifting, S13: omental section, S14: adhesiolysis |
| P4: gastrojejunal anastomosis | S0: null step, S15: treitz angle identification, S16: biliary limb measurement, S17: jejunum opening, S18: gastrojejunal stapling, S19: gastrojejunal defect closing, S26: gastrojejunal anastomosis reinforcement, S30: alimentary limb measurement |
| P5: anastomosis test | S0: null step, S22: gastric tube placement, S23: clamping, S24: ink injection, S25: visual assessment, S26: gastrojejunal anastomosis reinforcement, S39: coagulation |
| P6: jejunal separation | S0: null step, S20: mesenteric opening, S21: jejunal section |
| P7: closure petersen space | S0: null step, S27: petersen space exposure, S28: petersen space closing |
| P8: jejunojejunal anastomosis | S0: null step, S29: biliary limb opening, S30: alimentary limb measurement, S31: alimentary limb opening, S32: jejunojejunal stapling, S33: jejunojejunal defect closing, S34: jejunojejunal anastomosis reinforcement, S35: staple line reinforcement |
| P9: closure mesenteric defect | S0: null step, S36: mesenteric defect exposure, S37: mesenteric defect closing, S38: anastomosis fixation, |
| P10: cleaning coagulation | S0: null step, S39: coagulation, S40: irrigation aspiration |
| P11: disassembling | S0: null step, S40: irrigation aspiration, S41: parietal closure, S42: trocar removal |

available weak phase information could be very beneficial for step recognition in the low data regime.

We propose to represent the relationship as a step-phase mapping matrix $M_{s\to p}$, where the elements $m_{ij}$ of the matrix are binary indicator variables which are 1 if step $s_i$ occurs in phase $p_j$. The matrix encodes the weak information about which steps can occur in a particular phase and does not provide details of their occurrence, duration, and/or order. To enforce this weak link between steps and phases, the step predictions $\hat{y}_t^s$ of our Spatio-temporal model (as described earlier) are linearly transformed by $M_{s\to p}$ into the phase space. Then a weighted cross-entropy loss ($\mathcal{L}_{CE}$) captures the similarity between the phase labels ($y_t^p$) and the transformed predictions ($M_{s\to p} \times \hat{y}_t^s$) of the model. The dependency loss ($\mathcal{L}_{dep}$) is given by:

$$\mathcal{L}_{dep} = \mathcal{L}_{CE}(y_t^p, M_{s\to p} \times \hat{y}_t^s). \quad (2)$$

## IV. EXPERIMENTAL SETUP

In this section, we discuss the experimental setup of our method. First, we present the datasets used for evaluation. Next, we discuss the experimental study followed by the training setup and evaluation metrics.

### A. Datasets

*1) Bypass40:* The *Bypass40* dataset [7] consists of 40 videos of LRYGB procedures with resolution 854 × 480 or 1920 × 1080 pixels recorded at 25 fps. Each frame is manually assigned to one of the 11 phases and one of the 44 steps [7]. For example, steps such as *gastric opening, gastric tube placement, horizontal stapling*, and *vertical stapling* occur in *gastric pouch creation* phase. A detailed list of phases and steps along with their hierarchical relationship is presented in Table II. For more information, we ask the readers to refer to [7]. We split the 40 videos into 24, 6, and 10 videos for training, validation, and test sets, respectively, and sub-sampled them at 1 frame-per-second (fps). This amounts to 150k, 40k, and 65k images in each set. The images are resized to ResNet-50's input dimension of 224 × 224, and the training dataset is augmented by applying horizontal flip, saturation, and rotation.

*2) Cataracts:* The CATARACTS dataset, proposed in [29], contains 50 videos of cataract surgery. With the recent CATARACTS2020 challenge, the dataset has been released with step annotations. Similar to [6], we define a phase ontology for available step labels. Cataract surgery consists of 5 phases and 19 steps that are summarized in Table I. The dataset is extended with phase labels that is automatically

generated using the available step annotations and the ontology presented in Table I. For each frame in a video, the phase label is obtained by a simple lookup of the step label in Table I. The only constraint while generating phase labels is when there are steps that can occur in several phases. In this case, the phase of the immediately preceding frame is assigned to the current frame. Since the only steps that occur in more than one phase are Idle, Incision, and Viscodilatation, and they do not occur at the beginning or at the end of a phase, it is therefore always possible to identify the correct phase by checking the phase of the previous step. Since very few steps occur in multiple phases, the automatically generated phase labels by table lookup are accurate and do not require expert knowledge or verification from a clinical expert.

We split the 50 videos (following the challenge[2]) into 25, 5, and 20 videos for training, validation, and test sets, respectively. Each set consists of 66k, 3.5k, and 11.8k frames extracted at 1 fps from the videos. The frames are resized from $1920 \times 1080$ to $224 \times 224$, and the training set is augmented with horizontal flip, saturation, and rotation.

### B. Study

To demonstrate the effectiveness of our approach, we train and evaluate different configurations of the model. Given $n$ videos, of which $k$ are annotated with steps and the rest $(n-k)$ are weakly annotated with phases, the Spatio-temporal model is trained in the proposed weakly supervised setting utilizing the dependency loss, presented as 'DEP'. To analyze the efficacy of 'DEP', we compare it against the Spatio-temporal model trained only on $k$ videos in a fully-supervised approach for the task of step recognition, which we refer to as 'FSA'. Additionally, we add a state-of-the-art semi-supervised learning method proposed by Yu et al. [42] to our results. Yu et al. [42], proposed a teacher/student semi-supervised learning method where both the teacher and student models consisted of spatial and temporal components, CNN-biLSTM-CRF and CNN-LSTM respectively. As noted in Section II-B, [42] is a closely related work in the literature to the work presented in this paper. Hence, we have implemented and adapted the method of Yu et al. [42] for the task of step recognition. We repeat all the experiments for different values of $k \in \{3, 6, 12, 18\}$.

Furthermore, to analyze the influence of the number of additional videos with phase labels on the model performance, we conduct experiments where we fix $k$ videos with step annotations and vary the number of videos with phase annotations from 0 to $n-k$ (i.e., 3, 6, 12, etc.).

### C. Training

The ResNet-50 model is initialized with weights pre-trained on ImageNet. The complete ResNet-50 + SS-TCN model is then trained end-to-end for the task of step recognition. Since SS-TCN models the temporal information in an online setup, features from all the past frames in the video needs to be cached. To achieve this, a feature buffer is maintained to store features from the spatial model of the past frames. The feature buffer is reset at the end of the video. In all the experiments, the model is trained for 50 epochs with a learning rate of 1e-5, weight regularization of 5e-4, and a batch size of 64. The test results presented are from the best performing model on the validation set. The models were implemented in PyTorch and trained on NVIDIA RTX 2080 Ti.

### D. Evaluation Metrics

To effectively analyze our models, we observe the accuracy (ACC), precision (PR), recall (RE), and F1 score (F1) metrics used in related publications [10], [11], [12]. Accuracy quantifies the total correct classification of activity in the whole video. PR, RE, and F1 are computed class-wise, defined as:

$$PR = \frac{|GT \cap P|}{|P|}, \quad RE = \frac{|GT \cap P|}{|GT|}, \quad F1 = \frac{2}{\frac{1}{PR} + \frac{1}{RE}}, \tag{3}$$

where GT and P represent the ground truth and prediction for one class, respectively. These values are averaged across all the classes to obtain PR, RE, and F1 for each video in the test set. All four metrics, computed per video, are averaged across all the videos in the test set. Furthermore, where applicable, standard deviations are also computed across all the videos in the test set.

## V. Results and Discussions

### A. Bypass40

*1) Effect of Weak Supervision:* To quantitatively evaluate our method, the results of step recognition on the test set are presented in Table III. The table contains the results of our model with a varying number of videos in the training set labeled with steps (3, 6, 12, and 18) along with the rest of the training set containing phase annotations. The introduction of dependency loss 'DEP' for weak supervision significantly improves the performance over the model (FSA) trained only on the step labeled subset of the dataset. We notice a 10-13% improvement of the model trained with 'DEP' loss containing only 3 videos annotated with steps. Similarly, we see a 10-13% and 5-7% increase in performance in all the metrics of the 'DEP' model in experiments corresponding to 6 and 12 step annotated videos, respectively. Interestingly, our 'DEP' model, trained on a dataset with 50% of step and 50% of phase annotated videos, achieves performance close to the upper baseline 'FSA' model trained on the whole fully labeled dataset.

Moreover, the results of Yu et al. [42] semi-supervised method are also presented in Table III for different step annotated videos (3, 6, 12, and 18) used to train both teacher and student model. The student model's performance increases by 3-8% over 'FSA' in all the metrics for 6 videos with step annotations. Furthermore, an increase of 6% and 2% is noticed in recall and F1-score above 'FSA' with 12 step annotated videos. However, the method falls short of our proposed 'DEP' method. We notice a 10-15%, 2-6%, and 1-6% increase in performance in all the metrics of the 'DEP' model over Yu et al. with 3, 6 and 12 step annotated videos, respectively.

[2]https://www.synapse.org/#!Synapse:syn21680292/wiki/601563

TABLE III

BYPASS40: EFFECT OF WEAK SUPERVISION ON VARYING AMOUNT OF STEP LABELED VIDEOS. ACCURACY (ACC), PRECISION (PR), RECALL (RE), AND F1-SCORE (F1) (%) ARE REPORTED. 'FSA' DENOTES THE MODEL TRAINED FOR STEP RECOGNITION WITHOUT ANY PHASE ANNOTATIONS. 'DEP' DENOTES THE DEPENDENCY LOSSADDED FOR WEAK SUPERVISION USING PHASE LABELS ON THE REMAINING VIDEOS

| Model | # Videos Step | # Videos Phase | ACC | PR | RE | F1 |
|---|---|---|---|---|---|---|
| FSA | 3 (12%) | - | $45.02 \pm 9.96$ | $26.62 \pm 5.32$ | $21.87 \pm 4.70$ | $19.44 \pm 5.31$ |
| Yu et al. [42] | 3 (12%) | - | $43.27 \pm 11.8$ | $23.63 \pm 4.41$ | $23.91 \pm 5.71$ | $19.77 \pm 4.89$ |
| DEP | 3 (12%) | 21 | **57.20 ± 8.31** | **33.44 ± 6.04** | **33.16 ± 6.37** | **29.38 ± 6.11** |
| FSA | 6 (25%) | - | $59.80 \pm 10.17$ | $37.19 \pm 8.52$ | $35.93 \pm 7.31$ | $32.15 \pm 8.03$ |
| Yu et al. [42] | 6 (25%) | - | $62.55 \pm 10.09$ | $40.63 \pm 7.85$ | $43.71 \pm 8.35$ | $37.68 \pm 8.54$ |
| DEP | 6 (25%) | 18 | **68.03 ± 9.04** | **50.05 ± 6.82** | **45.86 ± 6.46** | **42.05 ± 7.44** |
| FSA | 12 (50%) | - | $68.26 \pm 8.31$ | $47.57 \pm 7.84$ | $44.74 \pm 7.59$ | $41.30 \pm 8.44$ |
| Yu et al. [42] | 12 (50%) | - | $67.89 \pm 11.04$ | $46.26 \pm 9.97$ | $50.11 \pm 8.20$ | $43.41 \pm 10.33$ |
| DEP | 12 (50%) | 12 | **73.43 ± 8.43** | **53.40 ± 7.43** | **51.19 ± 8.20** | **48.34 ± 8.85** |
| FSA | 18 (75%) | - | $72.82 \pm 6.76$ | $50.60 \pm 7.90$ | $48.98 \pm 8.33$ | $46.08 \pm 8.61$ |
| Yu et al. [42] | 18 (75%) | - | $73.33 \pm 10.15$ | **54.78 ± 11.05** | **57.21 ± 8.51** | **51.72 ± 10.59** |
| DEP | 18 (75%) | 6 | **73.88 ± 8.11** | $54.33 \pm 6.38$ | $51.79 \pm 7.10$ | $48.62 \pm 7.49$ |
| FSA | 24 (100%) | - | $76.12 \pm 7.39$ | $54.23 \pm 8.24$ | $50.94 \pm 7.53$ | $48.17 \pm 8.02$ |

TABLE IV

BYPASS40: EFFECT OF THE NUMBER OF PHASE ANNOTATED VIDEOS FOR STEP RECOGNITION USING 'DEP' LOSS FOR WEAK SUPERVISION. ACCURACY (ACC), PRECISION (PR), RECALL (RE), AND F1-SCORE (F1) (%) ARE REPORTED FOR SETUPS WITH 6, 12, AND 24 VIDEOS FULLY ANNOTATED WITH STEPS

| Model | # Videos Step | # Videos Phase | ACC | PR | RE | F1 |
|---|---|---|---|---|---|---|
| FSA | 6 | - | 59.80 | 37.19 | 35.93 | 32.15 |
| DEP | 6 | 3 | 62.15 | 40.48 | 37.15 | 33.48 |
| DEP | 6 | 6 | 67.94 | 46.17 | 42.61 | 39.67 |
| DEP | 6 | 12 | 68.07 | 47.18 | 43.18 | 40.42 |
| DEP | 6 | 18 | 68.03 | 50.05 | 45.86 | 42.05 |
| FSA | 12 | - | 68.26 | 47.57 | 44.74 | 41.30 |
| DEP | 12 | 3 | 72.79 | 50.10 | 48.39 | 45.06 |
| DEP | 12 | 6 | 72.43 | 53.02 | 51.20 | 47.26 |
| DEP | 12 | 12 | 73.43 | 53.40 | 51.19 | 48.34 |
| FSA | 24 | - | 76.12 | 54.23 | 50.94 | 48.17 |

Although both methods use 100% of the training videos for the task of step recognition, Yu et al. aim at exploiting the knowledge learned by an offline teacher model to generate pseudo labels for additional videos without step annotations while 'DEP' aims to use weak supervision through phase annotations. Hence, the method of Yu et al. is limited by the knowledge learned by the teacher model which uses only $k$ step annotated videos although it learns from both current and future frames. On the other hand, the superior performance of the 'DEP' model indicates the additional cues present in phase annotated videos, although weak, is advantageous and that the proposed method effectively utilizes this information in the lower data settings.

*2) Effect of the Amount of Phase Annotated Videos:* In Table IV, we present the results of our model with a varying number of phase annotated videos. Utilizing 6 videos containing step annotations, the addition of phase labeled videos as weak supervision improves all metrics: accuracy, F1, precision, and recall. With 6 videos annotated with phases, the model performance increases by 7-8% in all metrics over the baseline 'FSA' model. The addition of more videos does not affect the accuracy but further improves both precision and recall by 4%. This is due to our weakly-supervised method, which only provides supervision information if a step can occur in the given phase. This information helps to distinguish steps belonging to different phases, as opposed to steps belonging to the same phase. Therefore, the precision and recall of the model improve with more phase annotated videos, and no significant improvement in accuracy is seen. We see a similar trend when using 12 videos annotated with steps and increasing the number of videos annotated with phase labels. Thus, ultimately it is beneficial to train our method utilizing all additional videos in the dataset with phase annotations for weak supervision.

## B. Cataracts

*1) Effect of Weak Supervision:* We quantitatively evaluate our method and present the results of step recognition in Table V. The table contains the results of our model, on a similar set of experiments as with *Bypass40*, by varying the number of videos in the training set labeled with steps (3, 6, 12, and 18) along with the rest of the training set containing phase annotations. We see a similar trend as with bypass where the 'DEP' model outperforms 'FSA'. We notice a 13-22% improvement 'DEP' model considering only 3 step annotated videos. Furthermore, we see a 6-13% and 1-3% increase in performance in all the metrics of the 'DEP' model in experiments corresponding to 6 and 12 step annotated videos, respectively. We see that our method achieves a similar performance improvement on a relatively easier surgical workflow, such as cataracts, consistently surpassing the FSA in all labeled ratios. The semi-supervised method of Yu et al. achieves performance improvement of 16%, 8%, and

TABLE V

CATARACTS: EFFECT OF WEAK SUPERVISION ON VARYING AMOUNT OF STEP LABELED VIDEOS.ACCURACY (ACC), PRECISION (PR), RECALL (RE), AND F1-SCORE (F1) (%) ARE REPORTED. 'FSA' DENOTES THE MODEL TRAINED FOR STEP RECOGNITION WITHOUT ANY PHASE ANNOTATIONS. 'DEP' DENOTES THE DEPENDENCY LOSS ADDED FOR WEAK SUPERVISION USING PHASE LABELS ON THE REMAINING VIDEOS

| Model | # Videos Step | # Videos Phase | ACC | PR | RE | F1 |
|---|---|---|---|---|---|---|
| FSA | 3 (12%) | - | $48.47 \pm 10.62$ | $51.32 \pm 11.91$ | $37.44 \pm 9.85$ | $37.12 \pm 10.15$ |
| Yu et al. [42] | 3 (12%) | - | $59.61 \pm 10.67$ | $56.02 \pm 14.31$ | **61.82 ± 14.45** | $53.26 \pm 13.61$ |
| DEP | 3 (12%) | 22 | **66.78 ± 12.21** | **64.29 ± 12.50** | $59.73 \pm 11.93$ | **58.31 ± 12.73** |
| FSA | 6 (25%) | - | $69.51 \pm 11.16$ | $71.05 \pm 14.13$ | $56.70 \pm 12.67$ | $59.28 \pm 13.50$ |
| Yu et al. [42] | 6 (25%) | - | $74.62 \pm 8.22$ | $67.71 \pm 11.48$ | **75.93 ± 12.48** | $67.67 \pm 12.46$ |
| DEP | 6 (25%) | 19 | **75.28 ± 11.50** | **71.84 ± 14.30** | $69.19 \pm 12.72$ | **68.09 ± 13.97** |
| FSA | 12 (50%) | - | $78.02 \pm 9.05$ | $79.02 \pm 13.20$ | $69.55 \pm 12.04$ | $71.18 \pm 13.04$ |
| Yu et al. [42] | 12 (50%) | - | $77.84 \pm 12.55$ | $71.48 \pm 13.41$ | **79.92 ± 15.28** | $72.96 \pm 14.46$ |
| DEP | 12 (50%) | 13 | **79.94 ± 9.17** | **80.52 ± 12.93** | $72.62 \pm 11.91$ | **73.52 ± 13.29** |
| FSA | 18 (75%) | - | $82.5 \pm 8.07$ | **82.58 ± 11.91** | $76.05 \pm 11.62$ | $77.39 \pm 12.12$ |
| Yu et al. [42] | 18 (75%) | - | $78.59 \pm 10.71$ | $74.55 \pm 14.17$ | **78.16 ± 12.64** | $73.55 \pm 13.67$ |
| DEP | 18 (75%) | 7 | **82.64 ± 9.72** | $82.20 \pm 13.70$ | $77.32 \pm 12.70$ | **77.67 ± 13.56** |
| FSA | 25 (100%) | - | $83.37 \pm 9.50$ | $85.29 \pm 12.05$ | $78.96 \pm 11.93$ | $80.09 \pm 13.34$ |

TABLE VI

CATARACTS: EFFECT OF THE NUMBER OF PHASE ANNOTATED VIDEOS FOR STEP RECOGNITION USING 'DEP' LOSS FOR WEAK SUPERVISION. ACCURACY (ACC), PRECISION (PR), RECALL (RE), AND F1-SCORE (F1) (%) ARE REPORTED FOR SETUPS WITH 6, 12, AND 25 VIDEOS FULLY ANNOTATED WITH STEPS

| Model | # Videos Step | # Videos Phase | ACC | PR | RE | F1 |
|---|---|---|---|---|---|---|
| FSA | 6 | - | 69.51 | 71.05 | 56.70 | 59.28 |
| DEP | 6 | 3 | 71.34 | 67.84 | 62.27 | 62.01 |
| DEP | 6 | 6 | 74.30 | 71.70 | 64.18 | 64.96 |
| DEP | 6 | 12 | 73.57 | 70.88 | 65.68 | 66.03 |
| DEP | 6 | 19 | 75.28 | 71.84 | 69.19 | 68.09 |
| FSA | 12 | - | 78.02 | 79.02 | 69.55 | 71.18 |
| DEP | 12 | 3 | 77.60 | 78.26 | 68.60 | 69.87 |
| DEP | 12 | 6 | 80.11 | 81.60 | 72.46 | 73.98 |
| DEP | 12 | 13 | 79.94 | 80.52 | 72.62 | 73.52 |
| FSA | 25 | - | 83.37 | 85.29 | 78.96 | 80.09 |

1.5% over 'FSA' in F1-score for experiments corresponding to 3, 6, and 12 videos, respectively. However, as seen earlier, it falls short of 'DEP' by 5%, 0.5%, and 0.5% in the F1-score for experiments corresponding to 3, 6, and 12 videos. Interestingly, Yu et al. achieves high recall on both datasets (Table III & V). On CATARACTS, it even outperforms the 'DEP' model in recall in all the experiments but falls short significantly in precision. This could be credited to the student model which learns from imperfect pseudo labels generated by the teacher model. Since our proposed 'DEP' model learns from true phase labels on additional videos its performance increases in both precision and recall. This validates the applicability of our approach to different surgical workflows.

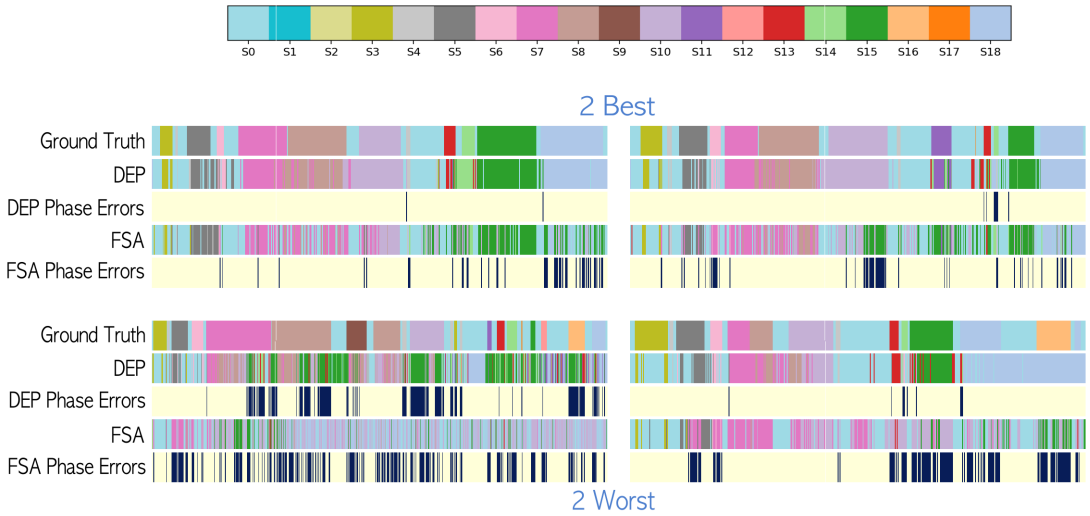
*2) Effect of the Amount of Phase Annotated Videos:* We present the results of our experiments, with a varying number of phase annotated videos, on CATARACTS in Table VI. We notice that utilizing 6 step annotated videos with additional phase labeled videos improves all the metrics by 6-13%. In particular, with 6 videos annotated with phases, we see a performance increase of 5% in accuracy and F1-score and 8% in recall of the 'DEP' model over the baseline 'FSA'. The addition of more videos provides a fractional improvement in accuracy but further improves both recall and F1-score by 1-4%. We see a similar trend when using 12 videos with step annotations reaffirming our hypothesis that it is beneficial to train our method utilizing all additional videos in the dataset with phase annotations for weak supervision.

### C. Weak Supervision on Step Predictions

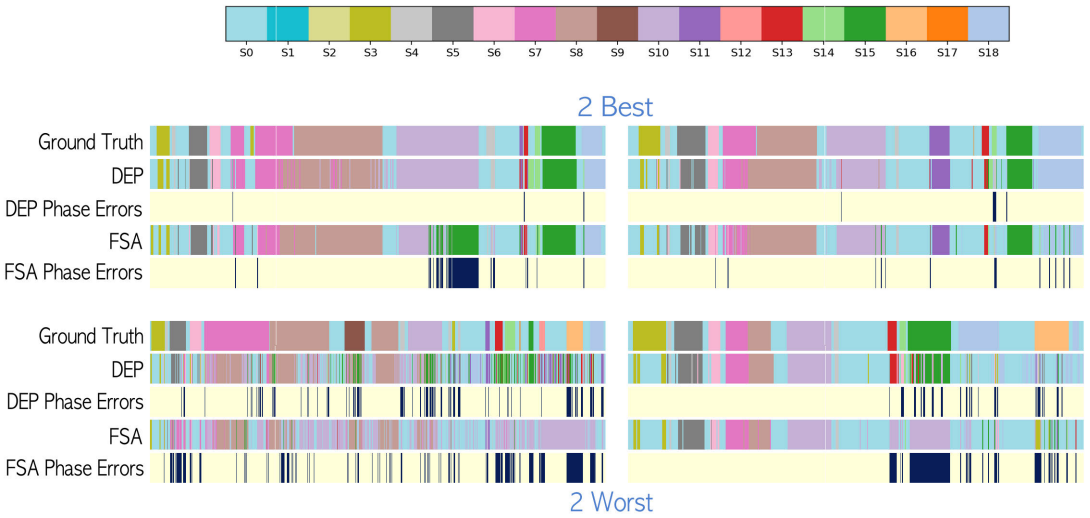
To visualize the effectiveness of our method, we visualize the step predictions of our method on the CATARACTS dataset which contains fewer phases and steps thereby enabling us to render a simple and clearer graphical diagram. We compare the step predictions of our 'DEP' model against 'FSA' for 2 best and 2 worst videos in CATARACTS in Fig. 3 for different labeled ratios (3, 6, and 12 videos with step annotations). Along with the step predictions we present the errors in the phase predictions for both models. The phase prediction error plot is computed as the errors in phase predictions derived from step predictions, using the step-phase mapping matrix, against ground truth phase predictions. Fig. 3 clearly depicts the effectiveness of our method for different labeled ratios. By correcting for the phase labels through dependency loss, our 'DEP' model is able to correct for corresponding step labels without explicit supervision for step recognition (e.g. S10, S15, S18). The top row of Fig. 3a shows this effect where we see a marked improvement in recognition of steps S18 (first video) and S10 (second video) by correcting for phase errors.

### D. Limitations

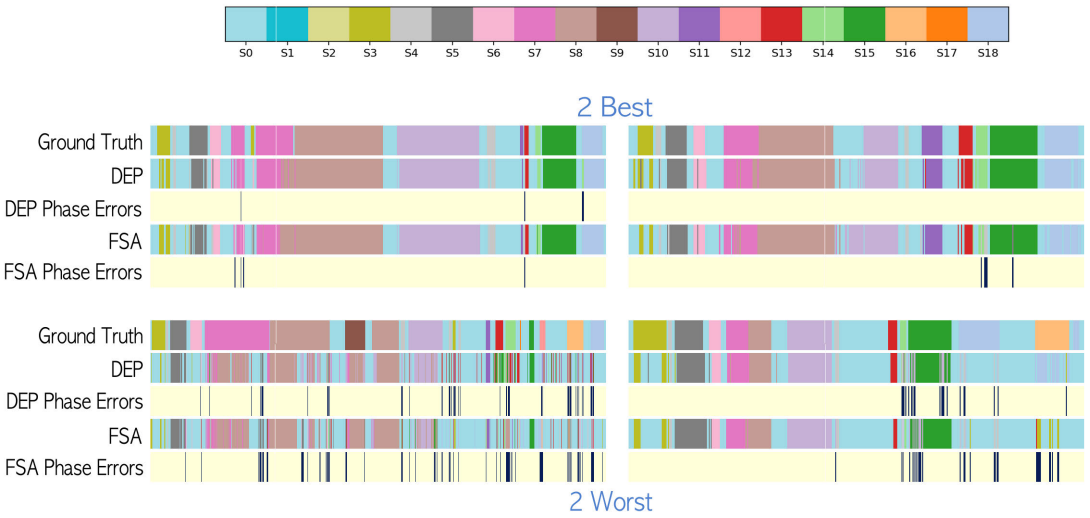
In some cases, for example, S16 (Fig. 3a, 3b, 3c), correcting for phase errors does not improve step recognition. The step is misrecognized with another step that occurs in the same phase.




(a) FSA vs DEP: 3 videos with step annotations.




(b) FSA vs DEP: 6 videos with step annotations.




(c) FSA vs DEP: 12 videos with step annotations.

Fig. 3. Step predictions on two best and two worst videos on the CATARACTS dataset for different labeled ratios. For each video, we visualize the step prediction of ground truth, DEP model predictions, DEP model phase prediction errors, FSA model predictions, and phase prediction errors of FSA model.

This is an expected outcome due to the intrinsic limitations of our weakly supervised method using coarser phase labels. Given the phase to be 'P2: gastric pouch creation' (Table II), it is impossible for a model to differentiate between 'crura dissection' and 'his angle dissection' or between 'horizontal stapling' and 'vertical stapling'. As can be seen in Fig. 1, the steps are quite similar in appearance and perform similar actions on the same anatomy (i.e., stomach or small intestine). This makes it challenging for a model to learn even when all the annotations are available. Furthermore, the phase information is too weak and does not provide any cues to better distinguish between the steps because both are valid steps in the current phase. Another limitation of our method is that adding more videos with phase annotations is not always beneficial. This limitation also stems from weak phase signals. If the fully supervised 'FSA' model learns to separate steps belonging to different phases, i.e., it has no or few phase-step correspondence errors, then additional videos with phase labels add no significant value as the model, during training, makes no/few errors in phase-step correspondence that helps improve feature learning. The significant errors by the model would be the inter-class separation of steps belonging to the same phase. Learning good representations to reduce these errors without supervision is a challenging task that needs to be tackled in future works.

Meanwhile, the effect of utilizing more phase annotated videos as weak supervision for improving the model performance on step recognition is presented in Tables IV & VI. As observed in Sections V-A.2 & V-B.2, it is beneficial to train the 'DEP' model utilizing all the additional phase annotated videos in the dataset for weak supervision. We also observe that in the lower data setting (6 videos with step annotations) model performance improves even when the phase annotated videos are increased from 12 to 18 (19 for cataracts). However, our study doesn't provide insights as to how many phase annotated videos are truly required to achieve the best performance by our proposed 'DEP' model. This is another limitation of our study, irrespective of the complexity of the procedure, that is hindered by the size of the available labeled datasets (24 in Bypass40 & 25 in CATARACTS). Understanding the extent of the 'DEP' model would require extending these datasets which is an important direction that needs to be pursued in future studies.

## VI. Conclusion

In this paper, we introduce a weakly-supervised learning method for surgical step recognition utilizing less demanding phase annotations. To model the weak supervision between steps and phases, we introduce a step-phase dependency loss and train a ResNet-50 + SS-TCN model end-to-end. The proposed method is extensively evaluated on a *Bypass40* dataset consisting of 40 LRYGB procedures and on the CATARACTS dataset containing 50 cataracts surgeries. The proposed 'DEP' model significantly improves the step recognition metrics over the baseline 'FSA' model for all the amounts of step annotations available. We hope that this work will inspire and foster future research in weak supervision for surgical workflow analysis utilizing multi-level descriptions of the workflow.

**Ethical Approval** The surgical videos were recorded and anonymized following the informed consent of patients in compliance with the local Institutional Review Board (IRB) requirements.

**Informed Consent** The patients consented to data recording.

## References


[1] K. Cleary and A. Kinsella, "OR 2020: The operating room of the future—Workshop report," *J. Laparoendoscopic Adv. Surgical Techn., A*, vol. 15, no. 5, pp. 495–573, 2005.

[2] L. Maier-Hein et al., "Surgical data science for next-generation interventions," *Nature Biomed. Eng.*, vol. 1, no. 9, pp. 691–696, 2017.

[3] T. Vercauteren, M. Unberath, N. Padoy, and N. Navab, "CAI4CAI: The rise of contextual artificial intelligence in computer-assisted interventions," *Proc. IEEE*, vol. 108, no. 1, pp. 198–214, Jan. 2020.

[4] D. Katic et al., "LapOntoSPM: An ontology for laparoscopic surgeries and its application to surgical phase recognition," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 10, no. 9, pp. 1427–1434, Jun. 2015.

[5] O. R. Meireles et al., "SAGES consensus recommendations on an annotation framework for surgical video," *Surgical Endoscopy*, vol. 35, no. 9, pp. 4918–4929, Jul. 2021.

[6] K. Charrière et al., "Real-time analysis of cataract surgery videos using statistical models," *Multimedia Tools Appl.*, vol. 76, no. 21, pp. 22473–22491, May 2017.

[7] S. Ramesh et al., "Multi-task temporal convolutional networks for joint recognition of surgical phases and steps in gastric bypass procedures," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 16, pp. 1111–1119, May 2021.

[8] A. P. Twinanda, S. Shehata, D. Mutter, J. Marescaux, M. de Mathelin, and N. Padoy, "EndoNet: A deep architecture for recognition tasks on laparoscopic videos," *IEEE Trans. Med. Imag.*, vol. 36, no. 1, pp. 86–97, Jan. 2017.

[9] O. Zisimopoulos et al., "DeepPhase: Surgical phase recognition in cataracts videos," in *Proc. MICCAI*, 2018, pp. 265–272.

[10] Y. Jin et al., "SV-RCNet: Workflow recognition from surgical videos using recurrent convolutional network," *IEEE Trans. Med. Imag.*, vol. 37, no. 5, pp. 1114–1126, May 2018.

[11] Y. Jin et al., "Multi-task recurrent convolutional network with correlation loss for surgical video analysis," *Med. Image Anal.*, vol. 59, Jan. 2020, Art. no. 101572.

[12] T. Czempiel et al., "TeCNO: Surgical phase recognition with multi-stage temporal convolutional networks," in *Proc. MICCAI*, 2020, pp. 343–352.

[13] L. Zappella, B. B. Haro, G. Hager, and R. Vidal, "Surgical gesture classification from video and kinematic data," *Med. Image Anal.*, vol. 17, no. 7, pp. 732–745, 2013.

[14] C. Lea, G. D. Hager, and R. Vidal, "An improved model for segmentation and recognition of fine-grained activities with application to surgical training tasks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, Jan. 2015, pp. 1123–1129.

[15] C. Lea, R. Vidal, and G. D. Hager, "Learning convolutional action primitives for fine-grained action recognition," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2016, pp. 1642–1649.

[16] N. Ahmidi et al., "A dataset and benchmarks for segmentation and recognition of gestures in robotic surgery," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 9, pp. 2025–2041, Sep. 2017.

[17] D. Liu and T. Jiang, "Deep reinforcement learning for surgical gesture segmentation and classification," in *Medical Image Computing and Computer Assisted Intervention—MICCAI 2018*. Cham, Switzerland: Springer, 2018, pp. 247–255.

[18] I. Funke, S. Bodenstedt, F. Oehme, F. von Bechtolsheim, J. Weitz, and S. Speidel, "Using 3D convolutional neural networks to learn spatiotemporal features for automatic surgical gesture recognition in video," in *Proc. MICCAI*, 2019, pp. 467–475.

[19] X. Gao, Y. Jin, Q. Dou, and P.-A. Heng, "Automatic gesture recognition in robot-assisted surgery with reinforcement learning and tree search," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2020, pp. 8440–8446.

[20] C. I. Nwoye et al., "Recognition of instrument-tissue interactions in endoscopic videos via action triplets," in *Medical Image Computing and Computer Assisted Intervention—MICCAI 2020*. Cham, Switzerland: Springer, 2020, pp. 364–374.

[21] H. Al Hajj, M. Lamard, P.-H. Conze, B. Cochener, and G. Quellec, "Monitoring tool usage in surgery videos using boosted convolutional and recurrent neural networks," *Med. Image Anal.*, vol. 47, pp. 203–218, Jul. 2018.

[22] C. I. Nwoye, D. Mutter, J. Marescaux, and N. Padoy, "Weakly supervised convolutional LSTM approach for tool tracking in laparoscopic videos," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 14, no. 6, pp. 1059–1067, Jun. 2019.

[23] K. Charriere, G. Quellec, M. Lamard, G. Coatrieux, B. Cochener, and G. Cazuguel, "Automated surgical step recognition in normalized cataract surgery videos," in *Proc. 36th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Aug. 2014, pp. 4647–4650.

[24] S. Bodenstedt et al., "Unsupervised temporal context learning using convolutional neural networks for laparoscopic workflow analysis," 2017, *arXiv:1702.03684*.

[25] I. Funke, A. Jenke, S. T. Mees, J. Weitz, S. Speidel, and S. Bodenstedt, "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, Switzerland: Springer, 2018, pp. 85–93.

[26] G. Yengera, D. Mutter, J. Marescaux, and N. Padoy, "Less is more: Surgical phase recognition with less annotations through self-supervised pre-training of CNN-LSTM networks," 2018, *arXiv:1805.08569*.

[27] T. Yu, D. Mutter, J. Marescaux, and N. Padoy, "Learning from a tiny dataset of manual annotations: A teacher/student approach for surgical phase recognition," in *Proc. Int. Conf. Inf. Process. Comput.-Assist. Intervent. (IPCAI)*, 2019, pp. 1–13.

[28] F. Fuentes-Hurtado, A. Kadkhodamohammadi, E. Flouty, S. Barbarisi, I. Luengo, and D. Stoyanov, "EasyLabels: Weak labels for scene segmentation in laparoscopic videos," *Int. J. Comput. Assist. Radiol. Surgery*, vol. 14, no. 7, pp. 1247–1257, Jul. 2019.

[29] H. A. Hajj et al., "CATARACTS: Challenge on automatic tool annotation for cataRACT surgery," *Med. Image Anal.*, vol. 52, pp. 24–41, Feb. 2019.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[31] A. P. Twinanda, D. Mutter, J. Marescaux, M. de Mathelin, and N. Padoy, "Single- and multi-task architectures for surgical workflow challenge at M2CAI 2016," 2016, *arXiv:1610.08844*.

[32] A. P. Twinanda, "Vision-based approaches for surgical activity recognition using laparoscopic and RBGD videos," Ph.D. theses, L'UFR de Mathématique et d'Informatique, Univ. de Strasbourg, Jan. 2017.

[33] Y. A. Farha and J. Gall, "MS-TCN: Multi-stage temporal convolutional network for action segmentation," in *Proc. CVPR*, 2019, pp. 3575–3584.

[34] A. van den Oord et al., "WaveNet: A generative model for raw audio," 2016, *arXiv:1609.03499*.

[35] Z. Jia, X. Huang, E. I-C. Chang, and Y. Xu, "Constrained deep weak supervision for histopathology image segmentation," *IEEE Trans. Med. Imag.*, vol. 36, no. 11, pp. 2376–2388, Nov. 2017.

[36] S. Hwang and H.-E. Kim, "Self-transfer learning for weakly supervised lesion localization," in *Proc. MICCAI*. Cham, Switzerland: Springer, 2016, pp. 239–246.

[37] F. Taherkhani, H. Kazemi, A. Dabouei, J. Dawson, and N. Nasrabadi, "A weakly supervised fine label classifier enhanced by coarse supervision," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6458–6467.

[38] H. Touvron, A. Sablayrolles, M. Douze, M. Cord, and H. Jégou, "Grafit: Learning fine-grained image representations with coarse labels," 2020, *arXiv:2011.12982*.

[39] G. Bukchin et al., "Fine-grained angular contrastive learning with coarse labels," 2020, *arXiv:2012.03515*.

[40] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Computer Vision—ECCV 2016*. Amsterdam, The Netherlands: Springer, 2016.

[41] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2015, pp. 2650–2658.

[42] T. Yu, D. Mutter, J. Marescaux, and N. Padoy, "Learning from a tiny dataset of manual annotations: A teacher/student approach for surgical phase recognition," 2018, *arXiv:1812.00033*.

Open Access funding provided by 'Università degli Studi di Verona' within the CRUI CARE Agreement