Medical Image Analysis 73 (2021) 102158

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Semi-supervised learning with progressive unlabeled data excavation for label-efficient surgical workflow recognition

Xueying Shi[a], Yueming Jin[a,*], Qi Dou[a,b], Pheng-Ann Heng[a,b]

[a] Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong
[b] T Stone Robotics Institute, The Chinese University of Hong Kong, Hong Kong

ARTICLE INFO

Article history:
Received 10 January 2021
Revised 4 June 2021
Accepted 29 June 2021
Available online 8 July 2021

Keywords:
Surgical workflow recognition
Semi-supervised learning
Video representation learning
Visual temporal dynamic consistency
Pseudo label generation

ABSTRACT

Surgical workflow recognition is a fundamental task in computer-assisted surgery and a key component of various applications in operating rooms. Existing deep learning models have achieved promising results for surgical workflow recognition, heavily relying on a large amount of annotated videos. However, obtaining annotation is time-consuming and requires the domain knowledge of surgeons. In this paper, we propose a novel two-stage **S**emi-**S**upervised **L**earning method for label-efficient **Surg**ical workflow recognition, named as *SurgSSL*. Our proposed SurgSSL progressively leverages the inherent knowledge held in the unlabeled data to a larger extent: from implicit unlabeled data excavation via motion knowledge excavation, to explicit unlabeled data excavation via pre-knowledge pseudo labeling. Specifically, we first propose a novel intra-sequence Visual and Temporal Dynamic Consistency (VTDC) scheme for implicit excavation. It enforces prediction consistency of the same data under perturbations in both spatial and temporal spaces, encouraging model to capture rich motion knowledge. We further perform explicit excavation by optimizing the model towards our pre-knowledge pseudo label. It is naturally generated by the VTDC regularized model with prior knowledge of unlabeled data encoded, and demonstrates superior reliability for model supervision compared with the label generated by existing methods. We extensively evaluate our method on two public surgical datasets of Cholec80 and M2CAI challenge dataset. Our method surpasses the state-of-the-art semi-supervised methods by a large margin, e.g., improving 10.5% Accuracy under the severest annotation regime of M2CAI dataset. Using only 50% labeled videos on Cholec80, our approach achieves competitive performance compared with full-data training method.

1. Introduction

Automatic surgical workflow recognition from surgical videos is a crucial prerequisite for various applications in computer-assisted surgery inside the modern operating room (Cleary and Kinsella, 2005; Padoy, 2019). By enhancing cognitive understanding of surgical procedures, surgical workflow recognition enables the context-aware computer-assisted system to have the capability of optimizing surgical process, generating real-time warning and supporting decision making of minimally invasive surgery (Bricon-Souf and Newman, 2007; Padoy et al., 2012; Dergachyova et al., 2016). Additionally, automatically analyzing surgical videos can be utilized beyond intra-operation, benefiting the postoperative review, surgeon training and surgeon skill evaluation (Maier-Hein et al., 2020).

Deep learning model has greatly advanced the surgical workflow recognition (Twinanda et al., 2016; Jin et al., 2017; Yi and Jiang, 2019), however, it inevitably relies on large-scale and frame-wise annotations of surgical videos for model learning. Unfortunately, accurate annotation demands expertise on domain knowledge of surgical operation. Meanwhile, obtaining high-quality labels of surgical video is laborious and expensive, given that the whole procedures last at least for hours. Considering that unlabeled surgical videos from new patients are generally abundant and easily collected in clinical sites, in this paper, we aim to develop a semi-supervised learning method for label-efficient surgical workflow recognition by effectively leveraging the unlabeled surgical videos, which can greatly reduce the labor of large scale data annotation.

Semi-supervised learning has shown potential for improving surgical workflow recognition with the scarce labeled data. DiPietro and Hager (2019) propose to improve the activity recognition by performing auxiliary self-learning tasks of kinematic data reconstruction and future prediction. Yengera et al. (2018) introduce the remaining surgery duration prediction as a pre-training task to enhance the network representation capability.

[*] Corresponding author.
E-mail address: ymjin@cse.cuhk.edu.hk (Y. Jin).

https://doi.org/10.1016/j.media.2021.102158
1361-8415/© 2021 Elsevier B.V. All rights reserved.

Yu et al. (2019) propose to first generate pseudo labels of unlabeled data by using the model only trained on labeled data. The produced pseudo labels are then utilized for model training, which introduces additional learning constraints for model, achieving the state-of-the-art results of semi-supervised learning for surgical workflow recognition. However, all these methods only utilize the labeled data for network training, while the informative knowledge within the abundant unlabeled data fails to be used to enhance the representation learning.

Current semi-supervised methods on natural image recognition stand on the cutting edge and can be broadly divided into three categories. One stream also follow the pseudo label generation (Lee, 2013; Berthelot et al., 2019; Xie et al., 2020), by regarding the predicted pseudo labels as the ground truth to further finetune the model. Other stream focus on consistency regularization strategy (Sajjadi et al., 2016; Laine and Aila, 2017; Tarvainen and Valpola, 2017), by enforcing the model predictions to remain unchanged for the same input under perturbations of input image itself or model parameters. This strategy can introduce the information of unlabeled data into models, increasing the model generality and stability. Very recently, Xie et al. (2020) comprise a combination of these two schemes. However, these existing studies leverage the unlabeled data in a form of single image, which are solely based on the visual space. Therefore, they are suboptimal solutions for sequential surgical video recognition.

While for action recognition of natural videos, which is introduced given its high task-wise similarity with surgical workflow recognition, existing semi-supervised methods are dedicated to model pre-training. Some transfer the knowledge from extra data source, such as the large-scale 2D still-image dataset (Girdhar et al., 2019), the large and easily-accessible web videos with weak labels of hashtags (Ghadiyaram et al., 2019). Others leverage the auxiliary tasks to learn the intrinsic representation of video, including using visual correspondence (Wang et al., 2019), using discriminator to differentiate a real or generated sample (Ahsan et al., 2018). The strategy of consistency regularization on temporal space remains unexplored so far.

In this work, we propose a novel two-stage **S**emi-**S**upervised **L**earning method for **Surg**ical workflow recognition (SurgSSL). We aim to progressively leverage the inherent knowledge from unlabeled data to a larger extent, from implicit unlabeled data excavation via motion knowledge excavation to explicit unlabeled data excavation via pre-knowledge pseudo labeling. In the first stage, we propose a novel intra-sequence Visual and Temporal Dynamic Consistency (VTDC) scheme with a form of self-supervision. By enforcing the consistency of data under different perturbations from both visual and temporal space, VTDC can implicitly dig up motion knowledge from unlabeled data to enhance video representation learning. It is motivated by the fact that features corresponding to video understanding like surgical workflow recognition should pay more attention to the dynamic action, for example, the interaction between tissues and surgical tools. Such features can be better extracted from video clips containing moving motion instead of separate frames with static scene. This naturally suggests that conventional visual consistency on perturbations of a single image is not enough (Fig. 1(a)), while performing both visual and temporal consistency across a video clip by our VTDC can better benefit workflow recognition (Fig. 1(b)).

In the second stage, we further generate pre-knowledge pseudo labels for unlabeled data using the regularized model from stage-1, where motion cues of unlabeled data have been encoded to provide valuable prior knowledge for label generation. Our pre-knowledge pseudo labels are therefore more reliable compared with those generated by existing methods, as they use the model solely trained with labeled data while without unlabeled data excavation. The distribution mismatch of labeled and unlabeled data cause inferior ability of model for recognizing the phase of unlabeled videos, therefore leads to high potentials for producing low-quality and noisy pseudo labels. Instead, we leverage our stage-1 model involving the information of unlabeled data and estimate more trustworthy pseudo labels. We then perform the label excavation from unlabeled data by jointly optimizing the model towards both pseudo labels and ground truth annotations of labeled data.

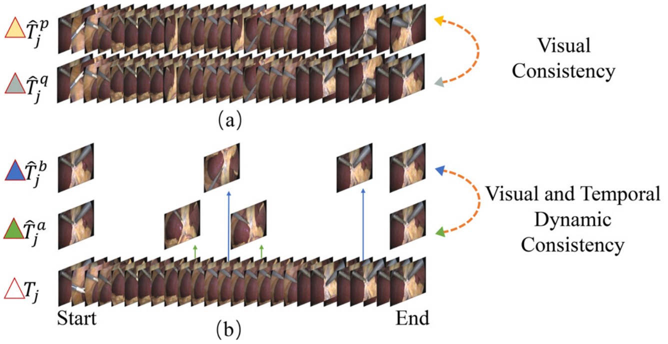

Fig. 1. Conventional visual consistency and our proposed intra-sequence visual and temporal dynamic consistency (VTDC) scheme for semi-supervised surgical workflow recognition. For a video sequence T_j, conventional semi-supervised schemes emphasize the consistency in image-level. While our VTDC explicitly enforces the consistency of data under different perturbations from both visual (random perturbations on every single image) and temporal domains (sampling a clip in time dimension with flexible strides). Besides visual variation, it encourages the model to additionally learn motion variation of video and forces the model to encode discriminative motion features instead of just background or motionless object's semantic features, thus aggregating motion knowledge from unlabeled data.

Our main contributions are summarized as follows:

1. We propose a novel **S**emi-**S**upervised **L**earning method for label-efficient **Surg**ical workflow recognition. We progressively utilize unlabeled data in two learning stages, from implicit excavation to explicit excavation.
2. We present a novel intra-sequence Visual and Temporal Dynamic Consistency (VTDC) scheme for implicit excavation from unlabeled data. By adding regularization from both visual and temporal perspectives, it encourages model to excavate motion cues from unlabeled videos.
3. We continue to perform explicit excavation from unlabeled data, by optimizing the model towards pre-knowledge pseudo labels. They can be naturally generated from stage-1 regularized model with prior knowledge encoded, and demonstrate more precise supervision capability compared with conventional pseudo labels.
4. We extensively validate our proposed SurgSSL on two popular surgical video datasets. Our approach largely outperforms state-of-the-art semi-supervised methods. Using only 50% labeled videos, our method achieves competitive results with full-data training, endorsing great potential in clinical practice. Code will be publicly available.

2. Related works

Before presenting the proposed approach, we first review the literature about our targeted task of surgical workflow recognition. Next, we recall methods about semi-supervised learning on general natural data and self-supervised learning for excavating inherent cues that are closely relevant to our work.

2.1. Surgical workflow recognition

Existing state-of-the-art fully supervised surgical workflow recognition methods are all based on the convolutional neural network and long short term memory (CNN-LSTM) backbone to integrate both visual and temporal information (Twinanda et al., 2016; Jin et al., 2017; Zisimopoulos et al., 2018; Jin et al., 2019b; 2019a; DiPietro et al., 2019; van Amsterdam et al., 2020; Qin et al., 2020). More specifically, they used CNN to first extract features for every single frame then utilized LSTM to embed these continuous image-level features to a sequential feature with the aim of capturing temporal information. Therefore, the CNN-LSTM model generally produced a more precise prediction for a given frame with the assistant of knowledge from previously continuous frames.

Semi-supervised learning has been studied for surgical workflow recognition given the scarce data annotations. DiPietro and Hager (2019) proposed to improve the activity recognition by first learning the representation of surgical kinematic data in an unsupervised learning fashion. Yengera et al. (2018) introduced remaining surgery duration prediction as a self-supervised pre-training task to increase the recognition performance. Yu et al. (2019) proposed to generate the pseudo labels of unlabeled data by using the model only trained on labeled data, in which the produced pseudo labels can introduce additional learning constraint for model learning, achieving the state-of-the-art results of semi-supervised learning for workflow recognition. However, these previous methods ignore the crucial clues inside the unlabeled data when doing network training. Also, they only consider neighbor temporal information while ignoring distant temporal cues. Therefore, for surgical videos with large appearance variance across frames (e.g., smoke noise, motion blur and fast action changing), they cannot precisely recognize frames without cross-frame dependency clues.

2.2. Semi-supervised learning

Semi-supervised learing has been broadly utilized in many medical image applications(Cheplygina et al., 2019; Xia et al., 2020; Ganaye et al., 2019; Xie et al., 2019; Shi et al., 2020b; Zheng et al., 2019; Wang et al., 2020b; Bouget et al., 2017). For natural image recognition, existing semi-supervised methods can be broadly categorized as pseudo label generation, consistency regularization or the combination of them.

For pseudo label based methods, Lee (2013), Berthelot et al. (2019), Berthelot et al. (2020), Xie et al. (2020) generated pseudo labels for unlabeled data and then added them to jointly train with labeled samples for enlarging data quantity. In this regard, the quality of pseudo label is vital and the difference among these methods is how to promote the accuracy of generated pseudo labels: Lee (2013) only utilized the high-confidence pseudo labels by setting a threshold and avoided using the low-confidence label that may introduce the label noise; Berthelot et al. (2019) enhanced the pseudo label by generating more distinguished prediction probabilities. They first generated K image-level augmentations and made predictions on all of them, then the probability was enhanced by ensembling predictions and temperature sharpen of all the K augmented samples; Xie et al. (2020) refined the generated pseudo labels round by round with the introduced noisy student model, which is based on the assumption that the model would learn the essential and distinct features under the noisy input of each round and gradually generate more precise pseudo labels.

However, the model trained only on labeled data tends to generate low-quality and noisy annotations of unlabeled data. Apart from pseudo label generation, Sajjadi et al. (2016), Laine and Aila (2017), Tarvainen and Valpola (2017) utilized consistency regularization to leverage the unlabeled data by enforcing the prediction consistency under different perturbations. Sajjadi et al. (2016) proposed an unsupervised loss function to minimize the difference between the predictions of multiple passes of a training sample through the network, where each pass conducts different perturbations, such as randomized data augmentation, dropout and random max-pooling. Apart from introducing perturbations of the same image in training, Temporal Ensembling (TE) (Laine and Aila, 2017) proposed to further ensemble predictions of the same example among the current model with earlier version models and thus improve the prediction quality. As the model updating is very slow with epoch-level, it is unclear to define when TE's prediction can be used. Therefore, Mean Teacher (MT) framework (Tarvainen and Valpola, 2017) further improved the ensemble scheme, where the ensemble is directly conducted on model's weights instead of predictions.

A recent work (Xie et al., 2020) combined pseudo labeling and consistency regularization, which first generated pseudo labels using the predictions on weakly-augmented unlabeled images, and model was then trained to enforce the predictions of strongly-augmented version to match the pseudo label. However, though these image-level semi-supervised methods have achieved great successes, they process the unlabeled data solely on the spatial space, therefore, are not the best choice for sequential surgical video recognition.

Compared with semi-supervised learning methods for natural image analysis, there is still rare study that can be inferred from natural video recognition task. Some methods transferred the knowledge from extra data (Ghadiyaram et al., 2019). Others leveraged the auxiliary tasks to learn the intrinsic representation of video, such as self-supervised learning of visual correspondence (Wang et al., 2019), using discriminator to differentiate a real or generated sample (Ahsan et al., 2018). However, the consistency regularization scheme on temporal dimension has not been applied on the video data.

2.3. Self-supervised learning

Recently, self-supervised is a promising direction for excavating inherent information of data by enhancing feature representation learning and has been applied on medical image analysis (Zhu et al., 2020). Also, it is beneficial for cost-effective learning by aggregating unlabeled data information jointly trained with ground truth labels. In surgical data science area, Funke et al. (2018) applied self-supervised pre-training of neural networks on unlabeled laparoscopic videos using temporal coherence, while Yengera et al. (2018) utilized an auxiliary self-supervised pre-training task before surgical phase recognition by predicting the remaining surgery duration from laparoscopic videos. da Costa Rocha et al. (2019) proposed a self-supervised optimization method to obtain good labels from an imprecise kinematic model projection on 2D image and utilized those generated labels for model fine-tuning. In addition to surgical video, self-supervised learning is also utilized in natural video area for improving feature learning (Wang et al., 2020a; Han et al., 2020; Kong et al., 2020).

3. Methods

3.1. Overview of SurgSSL framework

An overview of our proposed **S**emi-**S**upervised **L**earning framework for **Surg**ical workflow recognition (SurgSSL) is illustrated in Fig. 2. We optimize the model with progressive unlabeled data excavation in two stages: from implicit excavation via motion knowledge excavation, to explicit excavation via pre-knowledge pseudo labeling. We denote the **L**abeled set as $\mathcal{D}_L = \{(T_i, y_i)\}_{i=1}^{L}$ and **U**nlabeled set as $\mathcal{D}_U = \{(T_j)\}_{j=1}^{U}$, where $T_i = \{x_1, \dots, x_l, \dots, x_n\}$ and $T_j = \{x_1, \dots, x_u, \dots, x_n\}$ represent the n-frame video clips cutting from the labeled video and unlabeled video, respectively; y_i represents the label for the last frame of clip T_i. In stage-1, we propose a

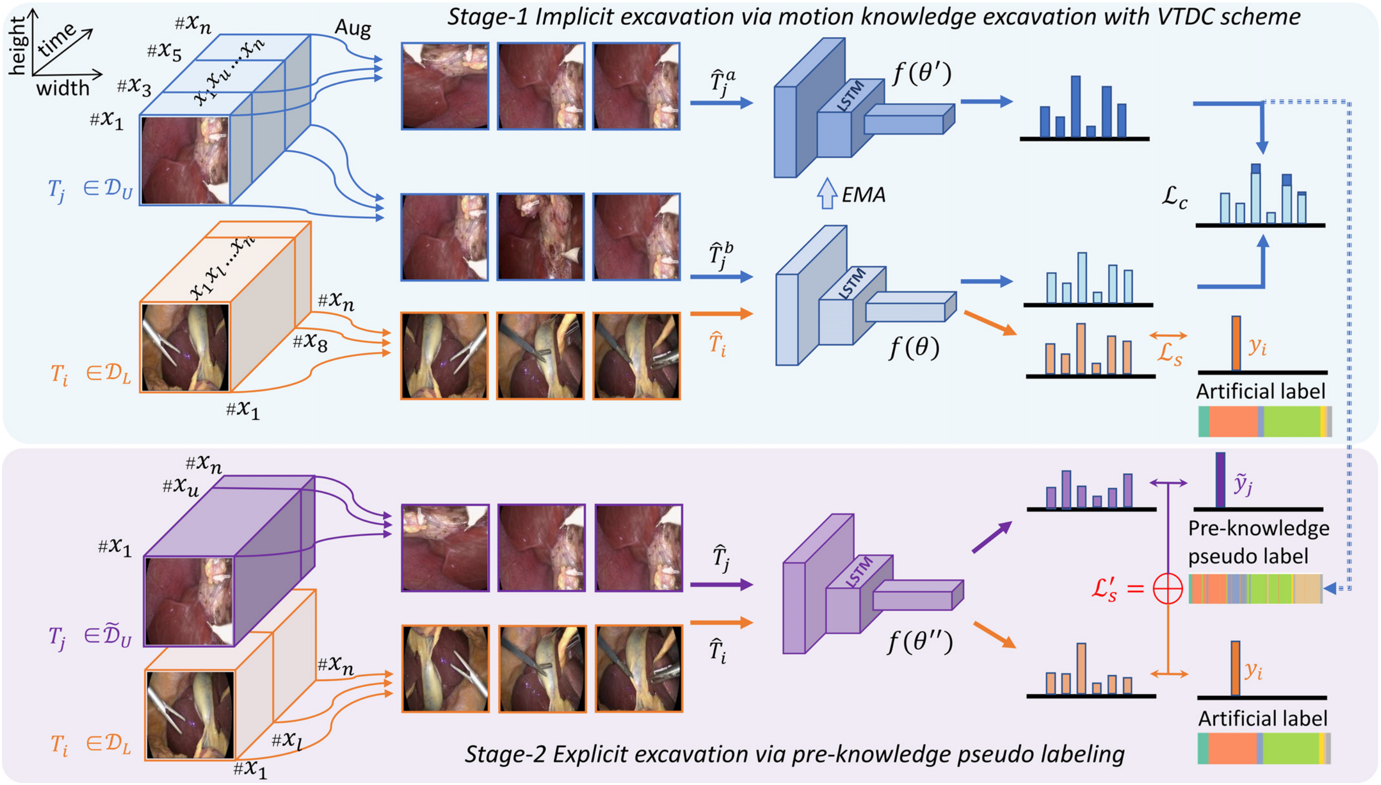

Fig. 2. The overview of our **S**emi-**S**upervised **L**earning (SSL) framework SurgSSL for **Surg**ical workflow recognition which consists of two stages: (1) semi-supervised learning by implicit unlabeled data excavation with proposed self-supervised intra-sequence visual and temporal dynamic consistency (VTDC); (2) explicit unlabeled data excavation with a novel pre-knowledge pseudo labeling. Stage-1 constrains unlabeled data $\mathcal{D}_U$ with self-supervised VTDC (loss $\mathcal{L}_c$) and labeled data with fully supervised manner (loss $\mathcal{L}_s$). Stage-2 generates pre-knowledge pseudo labels $\tilde{y}_j$ from the regularized model $f(\theta')$ and jointly training on entire dataset $\mathcal{D}_L \cup \tilde{\mathcal{D}}_U$ in supervised manner (loss $\mathcal{L}'_s$). The model $f(\theta')$ is updated as the exponential moving average (EMA) of model $f(\theta)'s$ weights. 'Aug' on the arrow means each frame would take frame-wise augmentation (rotation, flip, color transformation). $f(\theta)$, $f(\theta')$ and $f(\theta'')$ are based on the same network backbone.

novel self-supervised intra-sequence Visual and Temporal Dynamic Consistency (VTDC) strategy for encoding richer motion knowledge from unlabeled data. Labeled data are also included to jointly regularize the model in the same VTDC manner to yield its maximum efficacy. In stage-2, in order to fully utilize the unlabeled data, we generate a novel Pre-knowledge Pseudo Label for unlabeled data and update the unlabeled set as $\tilde{\mathcal{D}}_U = \{(T_j, \tilde{y}_j)\}_{j=1}^{U}$, where $\tilde{y}_j$ represents the generated pre-knowledge pseudo label for the last frame of clip T_j. Then, we leverage the generated pseudo labels and mix them with artificial labels annotated by surgeons to further optimize the network on the whole dataset $\mathcal{D}_L \cup \tilde{\mathcal{D}}_U$.

3.2. Implicit excavation via motion knowledge excavation

Features corresponding to video understanding like surgical workflow recognition should focus on the 'action', such as interaction between tissues and surgical tools. High-level discriminative features of surgical action tend to be easier learned from the motion cues of moving objects, instead of semantic information of background or static objects. Training model by excavating motion knowledge from unlabeled data therefore can facilitate capability of model for accurate surgical workflow recognition.

Motion knowledge exists in continuous video frames instead of single image frame. Previous consistency mechanisms of semi-supervised learning explore unlabeled data's information by imposing various perturbations onto the separately single image, while ignoring an essential property inherent in sequential video about temporal dependency. In order to consider temporal dependency which directly reflects motion knowledge in surgical workflow, we propose a self-supervised intra-sequence Visual and Temporal Dynamic Consistency (VTDC) scheme aiming to exploit motion knowledge held in unlabeled surgical videos. It simultaneously takes both visual representations and sequential dynamics into account, encouraging network to excavate more motion knowledge of unlabeled data to boost recognition capability.

Specifically, given a video clip $T = \{x_1, \ldots, x_n\}$, where T has the length of n, we design our VTDC scheme of creating sub-clips to follow three crucial design principles: i) We downsample the clip in time dimension with flexible stride $\tau = \{\tau_1, \tau_2, \ldots\}$, to generate various subsequences $\hat{T} = \{x_1, x_{1+\tau_1}, x_{1+\tau_1+\tau_2} \ldots, x_n\}$. The flexible stride τ introduces the diversity in time-dimension, greatly enhancing the temporal representation of surgical videos with dynamic and variational frame-to-frame intervals. ii) Conventional visual-level data augmentation (Aug) such as flipping, rotation and mirroring, is performed on every single frame of subsequence randomly, to increase the visual diversity with various context appearances. iii) The first and last frames in the subsequence are maintained to the same as the original video clip, in order to keep the original temporal duration.

The subsequence creation of our VTDC scheme is then exploited on unlabeled dataset $\mathcal{D}_U = \{(T_j)\}_{j=1}^{U}$. When inputting two subsequences from the same video clip, we argue that two network predictions should be consistent. Based on this assumption, in practice, we perform our VTDC scheme twice on the same video clip T_j, to derive two subsequences $\hat{T}_j^a$ and $\hat{T}_j^b$ distinguished in spatio-temporal feature. The generated two subsequences are then separately fed into the student model and teacher model, outputting the two prediction probabilities. Notably, the student and teacher models share the same network backbone of NL-RCNet (Shi et al., 2020a), with different network weights as $f(\theta)$ and $f(\theta')$, respectively. We devise a new consistency loss with VTDC scheme on unlabeled data as:

$$\mathcal{L}_c = -\frac{1}{U}\sum_{j=1}^{U} \| f(\theta', \hat{T}_j^a) - f(\theta, \hat{T}_j^b) \|_2^2, \\ \text{where } (\hat{T}_j^a, \hat{T}_j^b) \in \hat{T},\ \hat{T}_j^a \neq \hat{T}_j^b. \tag{1}$$

By minimizing $\mathcal{L}_c$ during the training process, our scheme regularizes the network by enforcing the prediction consistency between the teacher and student models with inputted samples under different perturbations in both visual and temporal spaces. The network would be enhanced the intra-subsequence similarity through compacting those subsequences from a same clip. On the other hand, the definition of $\mathcal{L}_c$ with VTDC scheme can be interpreted as enforcing model to learn the motion variation in surgical videos. By introducing VTDC, a given unlabeled clip can generate a pair of sub-clips which hold same start and end frame of the entire clip but with dynamic temporal motion evolution, so learning consistency of those sub-clip pairs can encourage the model to learn the motion variation of videos instead of background information or motionless object's semantic information. In this regard, unlabeled data implicitly promote surgical workflow recognition capability by enhancing motion representation of model.

In order to encode the inherent motion knowledge from unlabeled data and ground truth phases from labeled data in a same model, as shown in Fig. 2, we jointly train labeled and unlabeled data in semi-supervised manner in stage-1. The proposed VTDC scheme is exploited on the unlabeled data for aggregating the inherent motion information. While for labeled data, instead of inputting the complete video clip T_i to the network, we also create subsequence for the labeled data as we do for the unlabeled data, and feed the generated subsequence $\hat{T}_i$ into the network. Notably, such strategy randomly augments the sequence in both visual and temporal space, therefore also benefits the network learning from labeled data by remarkably enlarges the labeled training samples. We train our network using supervised loss function (i.e., cross-entropy loss $\mathcal{L}_{ce}$) for labeled subsequence data, which is defined as: $\mathcal{L}_s = -\frac{1}{L}\sum_{i=1}^{L} \mathcal{L}_{ce}(y_i, f(\theta, \hat{T}_i))$, where $f(.)$ is the recognition student model, with parameter weights of θ; $y_i \in \{0, C\}$ (C is class number) represents the label of video clip $\hat{T}_i$.

The parameter update in each training batch is driven by following overall loss function:

$$\mathcal{L} = \mathcal{L}_c + \lambda \mathcal{L}_s, \tag{2}$$

where λ is the trade-off weight to balance the unsupervised loss and supervised loss. Here, we calculate the typical cross-entropy loss as supervised loss for labeled data given the ground-truth labels available. While for unlabeled data, L2 loss is employed as it is bounded and less sensitive to incorrect predictions compared with cross-entropy loss. Additionally, recent literature on semi-supervised learning points out that the teacher model tend to be more reliable to predict consistent targets by ensembling weights of student model at different training steps. Following the same spirit, we optimize the student model θ during the network training, and update the weights $f(\theta')$ of the teacher model using exponential moving average (EMA) mechanism: $\theta'_t = \alpha\theta'_{t-1} + (1-\alpha)\theta_t$, where t represents each mini-batch step and α is a smoothing coefficient hyper-parameter that controls the weights updating rate.

3.3. Explicit excavation via pre-knowledge pseudo labeling

In order to take full utilization of unlabeled data, we continue to perform explicit excavation from unlabeled data by offline generating pseudo labels. The mean-teacher stream methods often focus on how to implicitly utilize unlabeled data by consistency constraint, which is a semi-supervised way of utilizing unlabeled data. While using the generated pseudo labels can provide more explicit supervision for model learning, by minimizing the predictions towards the labels. These two types of supervisions are complementary with each other. Existing pseudo-labeling paradigms follow the original pseudo-labeling framework (Lee, 2013) where one generates labels of unlabeled data by making predictions from a model trained only on labeled data. The core insight of these methods is that they create a mechanism to make selection of those generated pseudo label candidates, by either selecting pseudo label candidates with high probabilities, or generating more distinguished softmax prediction for those candidates and then making selection. However, the main drawback of those conventional pseudo-labeling paradigms is the Train-on-Labeled-Test-on-Unlabeled mode. The distribution mismatch of labeled and unlabeled data would cause performance degradation when generating labels for unlabeled data.

Instead of applying previous mode of pseudo label generation, we aim to produce pre-knowledge pseudo labels by considering the distribution among the global set (including unlabeled data). Our VTDC regularized model from stage-1 can naturally serve as a good base for progressively pseudo labeling, given that unlabeled data's motion knowledge has been already encoded in that model. Therefore, it can surpass existing pseudo label generation methods which only involve labeled data in network training, and contribute to more precise pseudo label estimation. Formally, we employ the regularized model $f(\theta')$ from stage-1 to produce the prediction probabilities for unlabeled data:

$$z_j = f(\theta', T_j). \tag{3}$$

By referring the highest probability, we generate the one-hot pseudo label $\tilde{y}_j = \text{argmax}(z_j)$ for unlabeled clip T_j. The unlabeled dataset is updated as $\tilde{\mathcal{D}}_U = \{(T_j, \tilde{y}_j)\}_{j=1}^{U}$.

We use confusion matrix to intuitively compare our pre-knowledge pseudo label with the one generated by conventional methods, in which models are trained only by labeled data. The confusion matrices are illustrated in Fig. 3 with X and Y-axis indicating pseudo label and ground truth. The diagonal number represents the true positive (TP) of a given class (7 classes in Cholec80 dataset). We can observe that TP is significantly larger in almost every single class of pre-knowledge pseudo label compared with conventional pseudo label, which means that our model regularized by VTDC and encoding unlabeled data's motion knowledge has better label estimation ability of unlabeled data and can produce more reliable pseudo labels.

After attaining the trustworthy pseudo labels, we then mix them with the artificial labels of labeled data as the final annotations. Stage-2 supervised training is then conducted on the entire dataset $\mathcal{D}_L \cup \tilde{\mathcal{D}}_U$ using the standard cross-entropy loss:

$$\mathcal{L}_s' = -\frac{1}{L}\sum_{i=1}^{L} \mathcal{L}_{ce}(y_i, f(\theta'', \hat{T}_i)) \\ -\frac{1}{U}\sum_{j=1}^{U} \mathcal{L}_{ce}(\tilde{y}_j, f(\theta'', \hat{T}_j)), \tag{4}$$

where L and U correspond to the numbers of labeled and unlabeled video clips; $f(.)$ represents the stage-2 recognition network, with the parameter weight denoted as θ''; y_i and $\tilde{y}_j$ respectively represent the ground truth for labeled clip $\hat{T}_i$, and one-hot pseudo label for unlabeled clip $\hat{T}_j$; $y_i, \tilde{y}_j \in \{0, C\}$.

As shown in Eq. (3), we only retain the class whose softmax result shows the largest probability and estimate one-hot hard pseudo labels. Such usage makes pseudo-labeling highly close to entropy minimization (Grandvalet and Bengio, 2005), where the

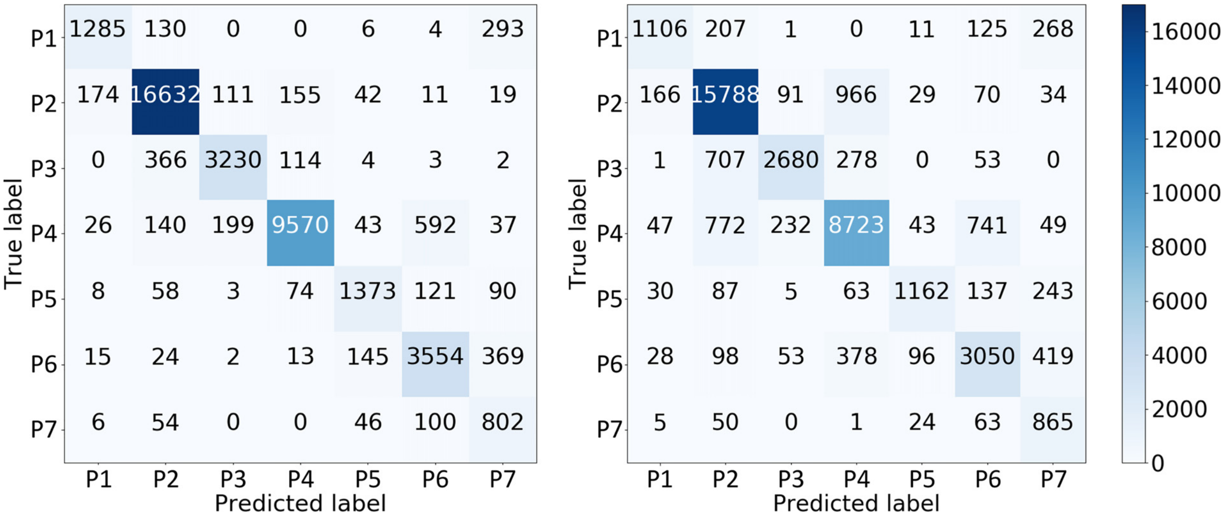

Fig. 3. Confusion matrix of pre-knowledge pseudo label vs ground truth (left) and conventional pseudo label vs ground truth (right) on Cholec80 dataset. Our pre-knowledge pseudo labels are much more reliable than conventional pseudo labels with more true positive (diagonal line). The reason is that pre-knowledge pseudo generation utilizes VTDC regularized model which encodes unlabeled data's pre-knowledge, but existing conventional pseudo labels are generated from model trained without unlabeled data knowledge encoding.

model's predictions are encouraged to be low-entropy (i.e., high-confidence) on unlabeled data. Through introducing our more accurate pre-knowledge pseudo label, the recognition capability of network can be further promoted in stage-2.

3.4. Network configuration and training details

3.4.1. Network configuration

We exploit an advanced architecture for surgical workflow recognition task, i.e., the non-local recurrent convolutional network (NL-RCNet) (Shi et al., 2020a) as our backbone. The framework takes advantage of the recurrent convolutional network (Jin et al., 2017) as well as non-local block (Wang et al., 2018), to model the temporal dependency within surgical workflow. A deep 50-layer residual network (ResNet50) (He et al., 2016) is used to extract high-level visual features from each frame and a LSTM network is harnessed to model the temporal information of sequential frames. We then seamlessly integrate these two components to form an end-to-end recurrent convolutional network, so that the complementary information of visual and temporal features can be sufficiently encoded for more accurate recognition. Based on this high qualitative feature, a non-local block is employed to capture long-range temporal dependency of frames within each clip. Different from the progressive behavior of convolutional and recurrent operations, non-local operations can directly compute interactions between any two positions in each clip, regardless of their positional distance. Therefore, it can enhance the feature distinctiveness for better recognition, with the capability of deducing the cross-frame dependency of arbitrary intervals. Experimental results have validated that this network configuration can provide a strong performance support to encourage our semi-supervised strategy yielding promising efficacy.

3.4.2. Technical details

Our SurgSSL framework conducts three types of perturbations during the network training: 1) Image-level data augmentation: we apply automatic augmentation with random 224×224 cropping, horizontal flips by a factor of 0.5, random rotations of $[-10, 10]$ degrees, brightness, saturation and contrasts by a random factor of 0.2, and hue by a random factor of 0.05. 2) Temporal regularization: our spatio-temporal regularization scheme randomly downsamples each input video clip from 10-frame clip to 4-frame clip in stage-1 and 40-frame clip to 10-frame clip in stage-2. 3) Dropout layer: we add dropout layer before the last fully convolutional layer in NL-RCNet model, with dropout rate as 0.2. The overall loss in stage-1 (Eq. (2)) is the weighted sum of supervised classification loss and unsupervised consistency loss, where the weight λ of classification loss is updated per minibatch, subject to the ramp-ups described below. We apply a ramp-up period at the beginning of training. The consistency loss coefficient $\lambda = 10 \times \gamma$ is ramped up from 0 to maximum value 10, using a sigmoid-shaped function $\gamma(t) = e^{-5(1-t)^2}$, where $t \in [0, 1]$.

3.4.3. Implementation details

Our framework is implemented based on the PyTorch library and we use 4 GeForce RTX 2080 Ti GPUs for acceleration. The network is trained using Adam optimizer with total 100 epochs (50 epochs in each training stage). The learning rates are initialized by 5×10^{-4} and then divided by a factor of 10 every 5 epochs in stage-1 and every 3 epochs in stage-2. We resize the frames from the original resolution of 1920×1080 and 854×480 into 250×250 to dramatically save memory and reduce network parameters.

4. Experiments

We validate our semi-supervised framework SurgSSL on two popular public surgical datasets, Cholec80 (Twinanda et al., 2016) and 2016 M2CAI Workflow Challenge (M2CAI16)[1], with extensive analytical ablation studies and comparison with state-of-the-art methods.

4.1. Dataset and evaluation metrics

Cholec80 dataset consists of 80 videos recording the cholecystectomy procedures performed by 13 surgeons. The videos are captured at 25 fps and each frame has the resolution of 854×480 or 1920×1080. We downsample the video to 1fps for training in all the experiments. All the frames are labeled with 7 defined phases by experts (see Table 1). Tool annotations also consisting of 7 categories, are conducted in this dataset. For fair comparison, we follow the same evaluation procedure reported in (Twinanda et al.,

[1] http://camma.u-strasbg.fr/m2cai2016/

Table 1
Surgical phases of cholecystectomy procedures in Cholec80 and M2CAI16 dataset, where M2CAI16 consists of phase P0-P7 while Cholec80 consists of phase P1-P7 with first two phases merged into one phase.

ID	Phase Name
P0	Trocar Placement
P1	Preparation
P2	Calot Triangle Dissection
P3	Clipping and Cutting
P4	Gallbladder Dissection
P5	Gallbladder Packaging
P6	Cleaning and Coagulation
P7	Gallbladder Retraction

2016), splitting the dataset into two subsets with equal size, with 40 videos for training and the rest 40 videos for testing.

M2CAI16 dataset consists of 41 videos recording the cholecystectomy procedures. These videos are acquired at 25fps and each frame has a resolution of 1920 × 1080. These videos are segmented into 3-8 phases by experienced surgeons (see Table 1). The dataset is divided into training set (27 videos) and testing set (14 videos) following the challenge evaluation.

Regarding the utilization of annotations, basically, we split the labeled data and unlabeled data in video-level instead of clip-wise in previous method (Shi et al., 2020a). Such strategy is more practically significant for surgeons. On the one hand, by seeing the whole and long-range contextual information, surgeons can perform annotations with higher precise and confidence. On the other hand, surgeons are no longer required to understand the newly coming videos for labelling, once adequate videos are annotated. For the network input in training and inference, we create video clips by sequentially sliding the window over the videos, with each time shifting one frame forward. The phase label of each video clip corresponds to the label of the last frame in each clip. We conduct all the experiments in the online mode, by only using the preceding frames for recognition.

To quantitatively analyze the performance of our method, we employ five metrics to evaluate our methods, including Accuracy(ACC), Precision (PR), Recall (RE), Jaccard (JA) and F1 Score (F1). PR, RE, JA and F1 Score are computed in phase-wise, defined as:

$$\begin{aligned} &PR = \frac{|GT \cap P|}{|P|}, \quad RE = \frac{|GT \cap P|}{|GT|}, \\ &JA = \frac{|GT \cap P|}{|GT \cup P|}, \quad F1 = \frac{2}{\frac{1}{PR} + \frac{1}{RE}}, \end{aligned} \tag{5}$$

where GT and P represent the ground truth set and prediction set of one phase, respectively. After PR, RE, JA and F1 of each phase are calculated, we average these values over all the phases and obtain them of the entire video. The ACC is calculated at video-level, defined as the percentage of frames correctly classified into the ground truths in the entire video. Next, we compare our proposed semi-supervised method with full data training methods as well as semi-supervised methods in order to show its effectiveness.

4.2. Comparison with state-of-the-art methods

4.2.1. Full-supervised results

In order to validate the effectiveness of subsequence creation scheme, also the capability of the NL-RCNet to serve as a powerful backbone support, we compare our NL-RCNet++ (using proposed scheme of creating visual and temporal dynamic subsequences on full labeled data to train NL-RCNet) with the state-of-the-art methods: 1) EndoNet (Twinanda et al., 2016) is a 9-layer multi-task network which leverages both tool and phase annotations, followed by hierarchical hidden Markov model to refine the results. As this method requires the tool presence annotations, only results on Cholec80 are available. 2) SV-RCNet (Jin et al., 2017) seamlessly integrates CNN and LSTM to jointly learn visual and temporal features. 3) NL-RCNet (Shi et al., 2020a) introduces non-local block to capture the long-range temporal dependency among continuous frames. 4) OHFM (Yi and Jiang, 2019) proposes a three-step strategy to alleviate the negative effect of hard frames for both network training and testing. Results on Cholec80 and M2CAI16 dataset are shown in Tables 2 and 3, respectively.

We see that our NL-RCNet++ outperforms others in all the five evaluation metrics on Cholec80 dataset, increasing JA with 2.7% and F1 with 2.3%. Our model also achieves superior results on M2CAI16 dataset. This demonstrates that by introducing non-local block to capture the long temporal cross-frame dependency, the NL-RCNet can be regarded as a strong backbone to assist our method yielding its best capacity. Our method also attains better performance than pure NL-RCNet. It verifies the efficacy of our scheme about subsequence creation for excavating video representation on full data training setting, which can alleviate overfitting issue by remarkably augmenting data in both visual and temporal spaces. Our scheme presents more substantial contribution on semi-supervised setting (shown in next section).

4.2.2. Semi-supervised results

We evaluate our semi-supervised framework SurgSSL on two datasets and results are shown in the semi-supervised part of Tables 2 and 3. In each table, we train the backbone network NL-RCNet (Shi et al., 2020a) with only the labeled data as the baseline performance. We also implement the state-of-the-art semi-supervised method CNN-biLSTM-CRF (Yu et al., 2019) on surgical workflow recognition and the advanced semi-supervised framework mean teacher (MT) (Tarvainen and Valpola, 2017) for comparison. In order to ensure the comparison fairness, we use the same network backbone in MT. To investigate the impact of using different numbers of labeled videos in these methods and provide more comprehensive comparison, we conduct the experiments with various labeled videos, i.e., 5/10/15/20 on Cholec80 (total 40 videos for training) and 3/6/9/12 on M2CAI16 (total 27 videos for training).

We observe that our SurgSSL framework consistently improves the fully supervised-only method NL-RCNet in all the labeled ratios. Also, our method outperforms the semi-supervised frameworks CNN-biLSTM-CRF and MT by a large margin. For Cholec80 dataset (Table 2), when leveraging extremely less data quantity (5 videos), our SurgSSL remarkably outperforms MT, with improvement of 7.0% in Accuracy, 8.9% in Jaccard and 7.6% in F1 score. In order to show the clearer comparison, we further visualize the recognition results under this extremely-scarce-labels regime. Results on three complete surgical procedures (patients) are shown in Fig. 4. Each row illustrates the phase recognition results of different methods. We use different colors to represent the phases during surgery, whose names are shown in legend. We observe that our SurgSSL can achieve more continuous predictions in each phase compared with supervised-only NL-RCNet and MT framework, especially in P2 and P4 (two phases with long duration). In addition, NL-RCNet++ with fully-supervised training on 40 videos can be regarded as the upper bound of our experiments. On Cholec80 dataset (Table 2), when annotations are extremely scarce with 5-labeled videos available, our SurgSSL framework narrows the performance gap with the upper bound, from 14.2% (NL-RCNet backbone v.s. upper bound) to 4.5% (ours v.s. upper bound). Notably, our SurgSSL achieves competitive results when trained using only up to 50% of samples compared with fully annotation training, with the performance difference of only 0.7%.

In Table 3, we see that our method achieves the same outstanding performance on M2CAI16 dataset, consistently surpassing MT framework with training on different numbers of labeled videos.

Table 2
Recognition performance comparison with different methods on Cholec80 dataset(mean±std., %).

Methods		Video used		Accuracy	Precision	Recall	Jaccard	F1 Score
		labeled	unlabeled					
Fully supervised	EndoNet (Twinanda et al., 2016)	40	0	81.7 ± 4.2	73.7 ± 16.1	79.6 ± 7.9	-	-
	SV-RCNet (Jin et al., 2017)	40	0	86.4 ± 7.3	82.9 ± 5.9	84.5 ± 8.0	-	-
	NL-RCNet (Shi et al., 2020a)	40	0	85.7 ± 7.0	82.9 ± 6.2	85.0 ± 5.2	70.0 ± 8.8	82.1 ± 6.5
	OHFM (Yi and Jiang, 2019)	40	0	87.3 ± 5.7	-	-	67.0 ± 13.13	-
	NL-RCNet+ (ours)	40	0	**87.7 ± 7.7**	**86.1 ± 6.2**	**85.9 ± 9.4**	**72.7 ± 9.0**	**84.4 ± 6.0**
Semi-supervised	NL-RCNet (Shi et al., 2020a)	5	0	73.5 ± 10.9	72.8 ± 9.8	71.4 ± 11.4	52.0 ± 9.3	67.3 ± 7.9
	CNN-biLSTM-CRF (Yu et al., 2019)	5	35	76.5 ± 8.7	75.3 ± 12.2	74.3 ± 9.5	55.9 ± 10.4	70.9 ± 9.2
	MT (Tarvainen and Valpola, 2017)	5	35	76.2 ± 11.0	73.4 ± 12.0	77.1 ± 12.1	56.7 ± 13.3	71.0 ± 11.1
	SurgSSL (ours)	5	35	**83.2 ± 7.7**	**81.8 ± 9.9**	**81.6 ± 12.3**	**65.6 ± 14.8**	**78.6 ± 11.7**
	NL-RCNet (Shi et al., 2020a)	10	0	80.0 ± 9.0	76.6 ± 10.2	77.2 ± 9.7	59.4 ± 11.3	73.5 ± 9.5
	CNN-biLSTM-CRF (Yu et al., 2019)	10	30	80.4 ± 8.6	78.7 ± 10.9	78.7 ± 9.3	61.3 ± 11.1	75.3 ± 9.2
	MT (Tarvainen and Valpola, 2017)	10	30	82.8 ± 7.1	80.4 ± 9.9	79.9 ± 10.8	63.9 ± 13.8	77.3 ± 10.7
	SurgSSL (ours)	10	30	**85.0 ± 7.7**	**83.3 ± 8.3**	**83.1 ± 12.3**	**68.0 ± 13.5**	**80.6 ± 10.1**
	NL-RCNet (Shi et al., 2020a)	15	0	81.5 ± 8.2	77.5 ± 9.5	79.4 ± 9.7	61.7 ± 11.7	75.4 ± 9.4
	CNN-biLSTM-CRF (Yu et al., 2019)	15	25	83.4 ± 7.4	80.3 ± 10.2	79.7 ± 12.8	63.8 ± 13.4	77.1 ± 10.8
	MT (Tarvainen and Valpola, 2017)	15	25	83.8 ± 7.1	81.0 ± 9.9	81.0 ± 9.3	65.3 ± 13.1	78.3 ± 10.3
	SurgSSL (ours)	15	25	**86.1 ± 6.6**	**83.2 ± 9.2**	**84.1 ± 8.0**	**69.1 ± 12.3**	**81.2 ± 9.2**
	NL-RCNet (Shi et al., 2020a)	20	0	82.5 ± 8.4	79.7 ± 9.0	80.9 ± 8.1	64.2 ± 10.2	77.6 ± 7.9
	CNN-biLSTM-CRF (Yu et al., 2019)	20	20	83.5 ± 8.2	80.8 ± 9.6	82.0 ± 6.4	66.0 ± 10.8	78.9 ± 8.1
	MT (Tarvainen and Valpola, 2017)	20	20	84.0 ± 8.2	80.3 ± 10.0	82.6 ± 7.5	66.4 ± 11.2	79.0 ± 8.3
	SurgSSL (ours)	20	20	**87.0 ± 7.4**	**84.2 ± 8.9**	**85.2 ± 11.1**	**70.5 ± 12.6**	**82.6 ± 8.9**

Table 3
Recognition performance comparison with different methods on M2CAI16 dataset(mean±std., %).

Methods		Video used		Accuracy	Precision	Recall	Jaccard	F1 Score
		labeled	unlabeled					
Fully supervised	SV-RCNet (Jin et al., 2017)	27	0	81.7 ± 8.1	81.0 ± 8.3	81.6 ± 7.2	65.4 ± 8.9	-
	NL-RCNet (Shi et al., 2020a)	27	0	81.7 ± 10.6	81.3 ± 9.7	82.0 ± 8.5	65.6 ± 10.1	79.4 ± 7.1
	OHFM (Yi and Jiang, 2019)	27	0	**84.8 ± 8.0**	-	-	68.5 ± 11.1	-
	NL-RCNet+ (ours)	27	0	83.2 ± 9.7	**84.5 ± 8.2**	**85.6 ± 7.9**	**69.7 ± 7.1**	**82.8 ± 4.4**
Semi-supervised	NL-RCNet (Shi et al., 2020a)	3	0	49.9 ± 20.1	54.5 ± 16.2	56.7 ± 25.5	32.1 ± 13.2	46.0 ± 14.2
	CNN-biLSTM-CRF (Yu et al., 2019)	3	24	52.0 ± 24.4	54.5 ± 20.0	55.3 ± 24.3	32.9 ± 15.3	47.2 ± 14.8
	MT (Tarvainen and Valpola, 2017)	3	24	50.0 ± 13.7	62.4 ± 16.7	59.1 ± 20.4	36.1 ± 9.3	50.9 ± 10.4
	SurgSSL (ours)	3	24	**60.5 ± 11.1**	**71.8 ± 17.3**	**64.8 ± 24.2**	**43.7 ± 13.7**	**60.9 ± 13.2**
	NL-RCNet (Shi et al., 2020a)	6	0	67.0 ± 13.6	67.6 ± 12.0	69.2 ± 13.5	49.2 ± 10.7	64.1 ± 10.0
	CNN-biLSTM-CRF (Yu et al., 2019)	6	21	65.6 ± 15.0	69.7 ± 15.1	68.8 ± 13.3	47.6 ± 10.2	63.7 ± 8.6
	MT (Tarvainen and Valpola, 2017)	6	21	69.2 ± 12.9	74.5 ± 11.1	69.1 ± 17.7	49.6 ± 10.2	64.6 ± 11.1
	SurgSSL (ours)	6	21	**74.5 ± 11.4**	**78.1 ± 11.9**	**72.1 ± 19.1**	**53.4 ± 12.1**	**69.4 ± 10.2**
	NL-RCNet (Shi et al., 2020a)	9	0	69.4 ± 14.3	70.5 ± 13.5	73.3 ± 13.0	51.6 ± 9.7	67.7 ± 7.5
	CNN-biLSTM-CRF (Yu et al., 2019)	9	18	70.6 ± 12.0	73.9 ± 12.7	70.9 ± 12.6	51.5 ± 9.4	67.6 ± 7.5
	MT (Tarvainen and Valpola, 2017)	9	18	70.9 ± 13.7	76.3 ± 15.2	73.4 ± 16.4	53.5 ± 10.6	68.3 ± 9.7
	SurgSSL (ours)	9	18	**77.6 ± 10.1**	**80.3 ± 12.8**	**78.7 ± 12.2**	**59.5 ± 9.6**	**75.0 ± 6.9**
	NL-RCNet (Shi et al., 2020a)	12	0	72.1 ± 13.7	74.1 ± 14.9	74.0 ± 10.4	54.4 ± 12.9	69.8 ± 11.3
	CNN-biLSTM-CRF (Yu et al., 2019)	12	15	72.6 ± 10.7	74.5 ± 14.7	69.9 ± 12.3	51.7 ± 11.3	67.8 ± 9.9
	MT (Tarvainen and Valpola, 2017)	12	15	78.1 ± 10.2	78.7 ± 11.9	78.5 ± 11.7	60.3 ± 10.6	75.2 ± 8.6
	SurgSSL (ours)	12	15	**79.6 ± 9.4**	**80.2 ± 11.3**	**79.6 ± 11.5**	**62.0 ± 11.1**	**76.6 ± 9.3**

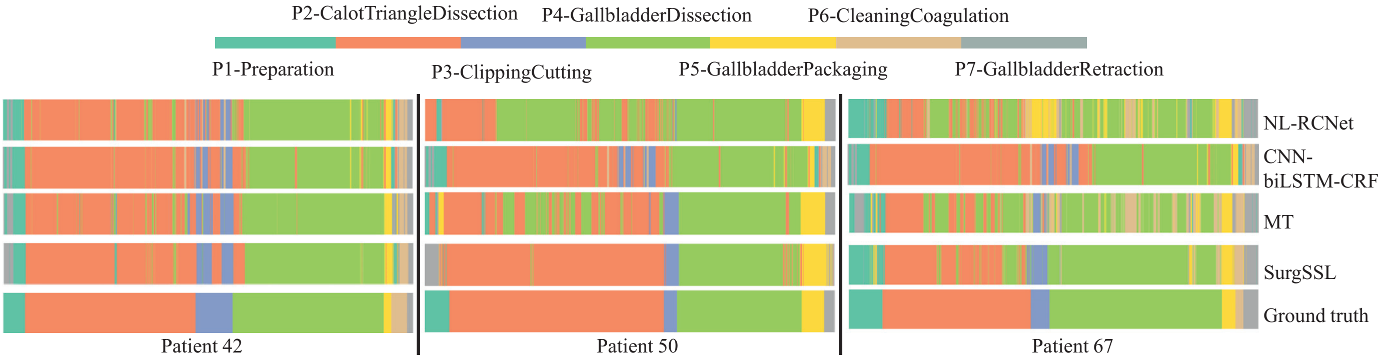

Fig. 4. Color-coded ribbon illustration for recognition comparison of different methods under the severest annotation case (only 5 labeled videos) on Cholec80 dataset. Each subfigure represents results of one surgical procedure for one single patient. The phase names are shown in legend.

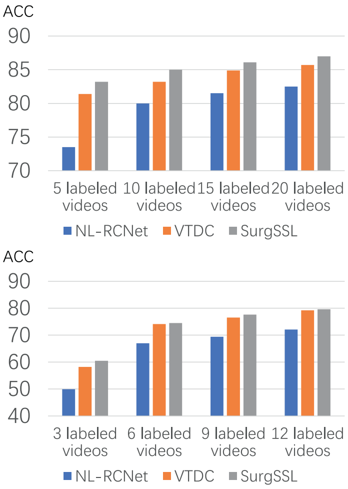

Fig. 5. Ablation study for two-stage excavation from unlabeled data under three settings, i.e., NL-RCNet as baseline, VTDC using only stage-1 excavation, and SurgSSL by adding stage-2 excavation. Experiments with different label ratios are conducted on both Cholec80 (top) and M2CAI16 (bottom) datasets.

Specifically, we outperform MT with 10.5% in Accuracy, 7.6% in Jaccard and 10% in F1 score using extremely less labeled data (3 videos). This verifies the generalization ability of SurgSSL framework for semi-supervised surgical workflow recognition.

4.3. Analytical ablation studies

4.3.1. Effectiveness of two-stage excavation

We conduct the ablation experiments to validate the effectiveness of different unlabeled data excavations in two stages. We draw the bar charts to illustrate performances in different annotated ratios with three configurations: NL-RCNet: baseline with solely labeled data for training; VTDC: using motion representation excavation with our VTDC scheme in stage-1; SurgSSL: our full model by adding the pre-knowledge pseudo label excavation for jointly network training in stage-2. The results on two datasets are shown in Fig. 5.

We can see that VTDC consistently attains better results over NL-RCNet on both Cholec80 (top) and M2CAI (bottom) datasets when using different quantities of labeled data, corroborating the efficacy of introducing temporal regularization among unlabeled sequence to semi-supervised learning. Particularly, in extremely less label quantity (5 videos) of Cholec80 dataset, VTDC surpasses NL-RCNet by a large margin.

We further analyze the importance of the stage-2 (pre-knowledge pseudo label excavation) in our semi-supervised method. Compared with VTDC, our entire SurgSSL framework achieves superior recognition results in all the annotation ratios after incorporating the generated pseudo labels for network training. This demonstrates the efficacy of combining the two strategies towards semi-supervised learning. In other words, training on our generated pseudo labels can further augment the consistency regularized model, by encouraging the high confidence towards correct ground truth for the unlabeled data. For intuitive interpretation, we illustrate our generated pre-knowledge pseudo labels on unlabeled training data in Fig. 6. It is observed that the pseudo labels predicted by VTDC is not only much more precise than NL-RCNet, but also very close to ground truth labels annotated by surgeons. This verifies that our VTDC can serve as a strong base model to invoke the effective joint training on the whole dataset with its generated pseudo labels and artificial labels.

In order to make more comprehensive comparison of our two-stage framework, we further compare our method with several strong semi-supervised baselines, including Π model (Laine and Aila, 2017), temporal embedding model (TE) (Laine and Aila, 2017) and directly using pseudo labels for training. We do the comparison on the larger dataset of Cholec80 with the most severe condition of only 5 labeled data. Among all the methods we compare, Π, TE and MT follow the pattern of our stage-1 training: semi-supervised method with consistency learning, while CNN-biLSTM-CRF and NL-RCNet+Pseudo follow the pattern of our stage-2 training: pseudo labeling. The results are listed in Table 4. We can see that our SurgSSL outperforms them with a large margin, which verifies the effectiveness of two-stage training. We also investigate whether updating the pseudo label after each epoch using moving average can benefit the model learning. We therefore derive SurgSSL+EMAPseudo with EMA-updated (exponential moving average (EMA) mechanism introduced in Sec 3.2) pseudo label generation like stage-1. We found that the performances of SurgSSL and SurgSSL+EMAPseudo are comparable, which indicates that the pseudo label gets more precise when model converges to the best. Therefore using EMA gradually updating pseudo label brings negligible benefits compared with directly using the pseudo label generated in the best learning epoch.

4.3.2. In-depth analysis of phase-level results

In order to comprehensively analyze the performance, we draw the phase-level bar chart to illustrate the results of four evaluation metrics in Fig. 7. We take the severest annotation case with only five annotated videos from Cholec80 as an example. It can be seen that our SurgSSL dominates other methods in six phases (seven phases in total) in terms of Precision, Jaccard and F1 score. For

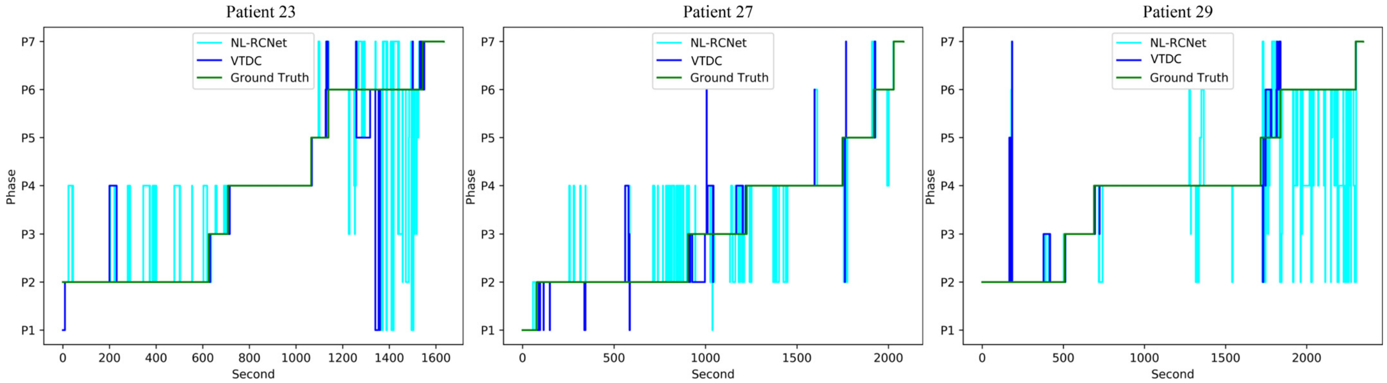

Fig. 6. Ablation study on pseudo label generalization quality. The figures visualize our pre-knowledge pseudo labels generated from VTDC, conventional pseudo labels generated from NL-RCNet, and ground truth annotations for three entire surgical videos from Cholec80 dataset. Models for generation are optimized with 20 labeled videos and 20 unlabeled videos.

Table 4
Recognition performance comparison with different frameworks for stage-1 and stage-2 on Cholec80 dataset of limited 5 labeled data (mean±std., %).

Methods	Accuracy	Precision	Recall	Jaccard	F1 Score
Π (Laine and Aila, 2017)	73.6 ± 10.3	73.5 ± 10.0	71.6 ± 9.0	53.1 ± 9.1	68.0 ± 8.3
TE (Laine and Aila, 2017)	74.0 ± 9.3	73.3 ± 10.3	71.6 ± 10.6	52.8 ± 10.0	67.9 ± 9.0
MT (Tarvainen and Valpola, 2017)	76.2 ± 11.0	73.4 ± 12.0	77.1 ± 12.1	56.7 ± 13.3	71.0 ± 11.1
CNN-biLSTM-CRF (Yu et al., 2019)	76.5 ± 8.7	75.3 ± 12.2	74.3 ± 9.5	55.9 ± 10.4	70.9 ± 9.2
NL-RCNet+Pseudo	78.4 ± 8.9	77.6 ± 8.8	75.5 ± 12.6	58.2 ± 11.2	72.6 ± 9.4
SurgSSL	83.2 ± 7.7	**81.8 ± 9.9**	81.6 ± 12.3	65.6 ± 14.8	78.6 ± 11.7
SurgSSL+EMAPseudo	**83.4 ± 8.2**	80.6 ± 12.5	**82.2 ± 10.1**	**65.7 ± 14.0**	**78.7 ± 11.0**

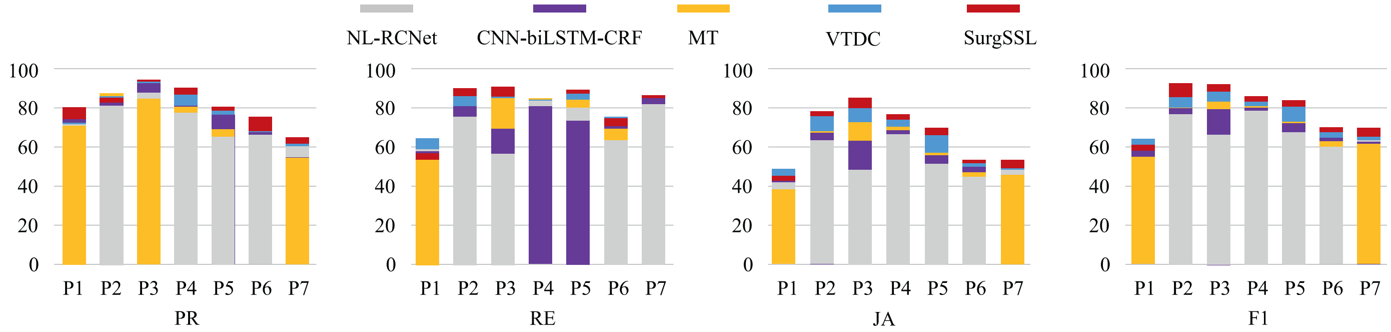

Fig. 7. Phase-level performance comparison of our SurgSSL with others on Cholec80 dataset in various evaluation metrics. From left to right are the results of Precision, Recall, Jaccard and F1 score, under 5 available labeled data.

Recall, SurgSSL surpasses others in four phases of P2, P3, P5 and P7. Such consistent improvements among diverse metrics proof the stable perfection of SurgSSL for recognizing each single phase of surgical procedure. Notably, the increase from other schemes to SurgSSL is especially significant for P2 in F1 score. The underlying reason is that P2 holds the longest duration in Cholec80, where our method can yield its maximum efficacy for model regularization, by providing more training samples compared with other phases ().

5. Discussion

Automatically recognizing surgical workflow from videos plays a key role in the development of intelligent context-aware modern operating rooms. However, existing satisfied performance for surgical workflow recognition on deep learning model always companies with high annotation cost, which is a huge burden for surgeons. In this paper, we propose a novel two-stage **S**emi-**S**upervised **L**earning method for label-efficient **Surg**ical workflow recognition, named as *SurgSSL*. It is designed with the aim of extending unlabeled data to the larger usage progressively: from implicit excavation via learning motion knowledge to explicit excavation via pseudo labeling.

In order to better fit practical usage, the annotation pattern design for surgeons is fundamental and quite significant (Bodenstedt et al., 2019; Shi et al., 2020a). Bodenstedt et al. (2019) utilize clip-wise or patient-wise annotation while (Shi et al., 2020a) utilize clip-wise annotation. From the view of surgeons, semi-supervised surgical workflow recognition is more practical under the patient-wise annotation reducing setting (DiPietro and Hager, 2018; 2019; Yu et al., 2019). In other words, surgeons give annotations video-by-video, with each video corresponding to the surgery procedure of one patient. There are at least two advantages. First, surgeons can provide more precise phase labels after observing the complete surgery procedure compared with only seeing the short term information. It is particularly crucial given the severe challenges of recognizing each surgical frame, including limited inter-phase variance, high intra-phase variance, motion blur and extra noise and artifacts. Second, as long as adequate videos are labeled, surgeons are no longer required to understand the newly coming videos for labelling. In this regard, our work exploits the patient-wise annotation setting and split labeled dataset and unlabeled dataset in video-level.

The temporal duration of input clip is a vital parameter for our visual and temporal dynamic consistency (VTDC) scheme, because information from previous continuous frames can provide significant guidance for the prediction of current frame. However, the longer duration not always generates the better performance. In stage-1, we empirically form a relatively short clip with duration of 10 s and downsample the 10-frame clip to 4-frame clip. The reason is to avoid introducing much ambiguous data among unlabeled samples when generating spatial-temporal perturbation examples. Otherwise, it shall perform a negative impact on network training when the quantity of unlabeled data dominates that of labeled data in stage-1. In stage-2, after obtaining the reliable pseudo labels, we enlarge the number of frames in each clip to 10, to maximize the usage of feasible computational cost. For the clip duration, we conducted variant clip-length trials and find that it is hard for model to remember quite long sequences. Therefore, we empirically form the clip with the duration of 40 s and downsample the 40-frame clip to 10-frame clip.

The proposed semi-supervised framework SurgSSL can be easily extended to many types of surgeries for recognizing phases, thanks to the low annotation cost. Only extremely small quantity of annotated videos are required from surgeons (5 or 10 videos), our SurgSSL framework then can be used for excavating knowledge from unlabeled surgical videos in semi-supervised manner. Workflow recognition task of many surgeries therefore can be conducted even when small quantity of annotations is available, broadening the applicability of our method to diverse surgeries. In addition, we can utilize VTDC scheme on robotic-assisted surgery for gesture recognition task. To better strengthen sequential representation capability, we can adopt VTDC on both non-visual kinematic data as well as visual surgical videos. A good representation of kinematics along time dimension can better model the motion transformation of robotic arms, thus help promote the prediction of gestures. Besides, VTDC with self-supervised manner can be used to integrate

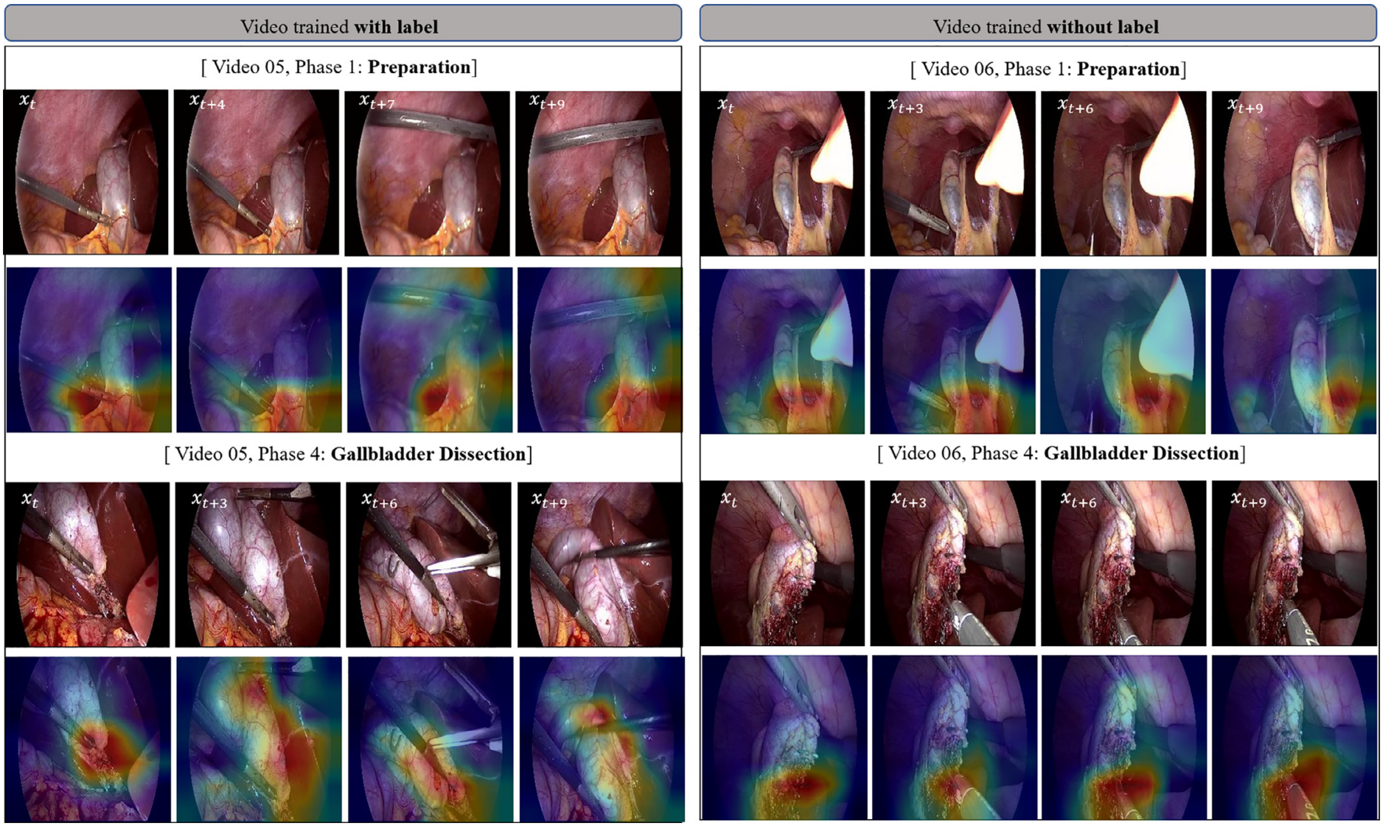
Video trained with label
[Video 05, Phase 1: Preparation]
x_t
x_{t+4}
x_{t+7}
x_{t+9}
[Video 05, Phase 4: Gallbladder Dissection]
x_t
x_{t+3}
x_{t+6}
x_{t+9}
Video trained without label
[Video 06, Phase 1: Preparation]
x_t
x_{t+3}
x_{t+6}
x_{t+9}
[Video 06, Phase 4: Gallbladder Dissection]
x_t
x_{t+3}
x_{t+6}
x_{t+9}

Fig. 8. Attention map visualization results separately on videos trained with and without label captured by our VTDC model on different surgical phases. The VTDC model can locate the interactions of tools and tissues even without supervision of phase label, which corresponds to what model has learned in labeled videos.

diverse information of both kinematic and visual features in the same timestamp, instead of only using one of them.

6. Conclusion

In this work, we propose a novel two-stage semi-supervised method SurgSSL for surgical workflow recognition. We aim to progressively leverage the inherent knowledge from unlabeled data to a larger extent, from implicit unlabeled data excavation via motion knowledge excavation to explicit unlabeled data excavation via pre-knowledge pseudo labeling.

We continue to perform explicit excavation from unlabeled data, by optimizing the model towards pre-knowledge pseudo labels. They can be naturally generated from stage-1 regularized model with prior knowledge encoded, and demonstrate more precise supervision capability compared with conventional pseudo labels. We extensively validate our proposed SurgSSL on two public surgical video datasets. Our approach substantially outperforms other state-of-the-art semi-supervised methods, and achieves competitive results compared with the full-data training method.

Declaration of Competing Interest

We confirm that there are no known conflicts of interest associated with this work. There has been no significant financial support for this work that could have influenced its outcome.

CRediT authorship contribution statement

Xueying Shi: Conceptualization, Methodology, Validation, Writing – original draft, Software. **Yueming Jin:** Conceptualization, Methodology, Formal analysis, Writing – review & editing. **Qi Dou:** Conceptualization, Writing – review & editing, Project administration. **Pheng-Ann Heng:** Writing – review & editing, Project administration, Funding acquisition.

Acknowledgments

This work is partially supported by Hong Kong RGC TRS project T42-409/18-R, and a grant from the National Natural Science Foundation of China (Project No. U1813204) and CUHK T Stone Robotics Institute.

References

Ahsan, U., Sun, C., Essa, I., 2018. DiscrimNet: semi-supervised action recognition from videos using generative adversarial networks. arXiv preprint arXiv:1801.07230.

Berthelot, D., Carlini, N., Cubuk, E.D., Kurakin, A., Sohn, K., Zhang, H., Raffel, C., 2020. ReMixMatch: semi-supervised learning with distribution alignment and augmentation anchoring. In: International Conference on Learning Representations.

Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., Raffel, C.A., 2019. MixMatch: a holistic approach to semi-supervised learning. In: Advances in Neural Information Processing Systems, pp. 5050–5060.

Bodenstedt, S., Rivoir, D., Jenke, A., Wagner, M., Breucha, M., Müller-Stich, B., Mees, S.T., Weitz, J., Speidel, S., 2019. Active learning using deep Bayesian networks for surgical workflow analysis. Int. J. Comput. Assisted Radiol. Surg. 14 (6), 1079–1087.

Bouget, D., Allan, M., Stoyanov, D., Jannin, P., 2017. Vision-based and marker-less surgical tool detection and tracking: a review of the literature. Med. Image Anal. 35, 633–654.

Bricon-Souf, N., Newman, C.R., 2007. Context awareness in health care: a review. Int. J. Med. Inf. 76 (1), 2–12.

Cheplygina, V., de Bruijne, M., Pluim, J.P., 2019. Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. Med. Image Anal. 54, 280–296.

Cleary, K., Kinsella, A., 2005. OR 2020: the operating room of the future.. J. Laparosc. Adv. Surg.Tech. Part A 15 (5), 495–497.

da Costa Rocha, C., Padoy, N., Rosa, B., 2019. Self-supervised surgical tool segmentation using kinematic information. In: 2019 International Conference on Robotics and Automation (ICRA). IEEE, pp. 8720–8726.

Dergachyova, O., Bouget, D., Huaulmé, A., Morandi, X., Jannin, P., 2016. Automatic data-driven real-time segmentation and recognition of surgical workflow. Int. J. Comput. Assist.Radiol. Surg. 11 (6), 1081–1089.

DiPietro, R., Ahmidi, N., Malpani, A., Waldram, M., Lee, G.I., Lee, M.R., Vedula, S.S., Hager, G.D., 2019. Segmenting and classifying activities in robot-assisted surgery with recurrent neural networks. Int. J. Comput. Assist.Radiol. Surg. 14 (11), 2005–2020.

DiPietro, R., Hager, G.D., 2018. Unsupervised learning for surgical motion by learning to predict the future. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 281–288.

DiPietro, R., Hager, G.D., 2019. Automated surgical activity recognition with one labeled sequence. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 458–466.

Funke, I., Jenke, A., Mees, S.T., Weitz, J., Speidel, S., Bodenstedt, S., 2018. 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. Springer, pp. 85–93.

Ganaye, P.-A., Sdika, M., Triggs, B., Benoit-Cattin, H., 2019. Removing segmentation inconsistencies with semi-supervised non-adjacency constraint. Med. Image Anal. 58, 101551.

Ghadiyaram, D., Tran, D., Mahajan, D., 2019. Large-scale weakly-supervised pre-training for video action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 12046–12055.

Girdhar, R., Tran, D., Torresani, L., Ramanan, D., 2019. Distinit: learning video representations without a single labeled video. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 852–861.

Grandvalet, Y., Bengio, Y., 2005. Semi-supervised learning by entropy minimization. In: Advances in Neural Information Processing Systems, pp. 529–536.

Han, T., Xie, W., Zisserman, A., 2020. Self-supervised co-training for video representation learning. Adv. Neural Inf. Process. Syst. 33.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778.

Jin, Y., Cheng, K., Dou, Q., Heng, P.-A., 2019. Incorporating temporal prior from motion flow for instrument segmentation in minimally invasive surgery video. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 440–448.

Jin, Y., Dou, Q., Chen, H., Yu, L., Qin, J., Fu, C.-W., Heng, P.-A., 2017. SV-RCNet: workflow recognition from surgical videos using recurrent convolutional network. IEEE Trans. Med. Imaging 37 (5), 1114–1126.

Jin, Y., Li, H., Dou, Q., Chen, H., Qin, J., Fu, C.-W., Heng, P.-A., 2019. Multi-task recurrent convolutional network with correlation loss for surgical video analysis. Med. Image Anal. 101572.

Kong, Q., Wei, W., Deng, Z., Yoshinaga, T., Murakami, T., 2020. Cycle-contrast for self-supervised video representation learning. Adv. Neural Inf. Process. Syst. 33.

Laine, S., Aila, T., 2017. Temporal ensembling for semi-supervised learning. In: International Conference on Learning Representations.

Lee, D.-H., 2013. Pseudo-label: the simple and efficient semi-supervised learning method for deep neural networks. In: Workshop on Challenges in Representation Learning, ICML, Vol. 3, p. 2.

Maier-Hein, L., Eisenmann, M., Sarikaya, D., März, K., Collins, T., Malpani, A., Fallert, J., Feussner, H., Giannarou, S., Mascagni, P., et al., 2020. Surgical data science–from concepts to clinical translation. arXiv preprint arXiv:2011.02284.

Padoy, N., 2019. Machine and deep learning for workflow recognition during surgery. Minim. Invasive Ther. Allied Technol. 28 (2), 82–90.

Padoy, N., Blum, T., Ahmadi, S.-A., Feussner, H., Berger, M.-O., Navab, N., 2012. Statistical modeling and recognition of surgical workflow. Med. Image Anal. 16 (3), 632–641.

Qin, Y., Pedram, S.A., Feyzabadi, S., Allan, M., McLeod, A.J., Burdick, J.W., Azizian, M., 2020. Temporal segmentation of surgical sub-tasks through deep learning with multiple data sources. In: IEEE International Conference on Robotics and Automation.

Sajjadi, M., Javanmardi, M., Tasdizen, T., 2016. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. In: Advances in Neural Information Processing Systems, pp. 1163–1171.

Shi, X., Jin, Y., Dou, Q., Heng, P.-A., 2020. LRTD: long-range temporal dependency based active learning for surgical workflow recognition. In: International Conference on Information Processing in Computer-Assisted Interventions.

Shi, X., Su, H., Xing, F., Liang, Y., Qu, G., Yang, L., 2020. Graph temporal ensembling based semi-supervised convolutional neural network with noisy labels for histopathology image analysis. Med. Image Anal. 60, 101624.

Tarvainen, A., Valpola, H., 2017. Mean teachers are better role models: weight-averaged consistency targets improve semi-supervised deep learning results. In: Advances in Neural Information Processing Systems, pp. 1195–1204.

Twinanda, A.P., Shehata, S., Mutter, D., Marescaux, J., De Mathelin, M., Padoy, N., 2016. EndoNet: a deep architecture for recognition tasks on laparoscopic videos. IEEE Trans. Med. Imaging 36 (1), 86–97.

van Amsterdam, B., Clarkson, M.J., Stoyanov, D., 2020. Multi-task recurrent neural network for surgical gesture recognition and progress prediction. In: IEEE International Conference on Robotics and Automation.

Wang, J., Jiao, J., Liu, Y.-H., 2020. Self-supervised video representation learning by pace prediction. In: European Conference on Computer Vision.

Wang, X., Girshick, R., Gupta, A., He, K., 2018. Non-local neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7794–7803.

Wang, X., Jabri, A., Efros, A.A., 2019. Learning correspondence from the cycle-consistency of time. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2566–2576.

Wang, Z., Lin, Y., Cheng, K.-T.T., Yang, X., 2020. Semi-supervised mp-MRI data synthesis with StitchLayer and auxiliary distance maximization. Med. Image Anal. 59, 101565.

Xia, Y., Yang, D., Yu, Z., Liu, F., Cai, J., Yu, L., Zhu, Z., Xu, D., Yuille, A., Roth, H., 2020. Uncertainty-aware multi-view co-training for semi-supervised medical image segmentation and domain adaptation. Med. Image Anal. 65, 101766.

Xie, Q., Luong, M.-T., Hovy, E., Le, Q.V., 2020. Self-training with noisy student improves imagenet classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10687–10698.

Xie, Y., Zhang, J., Xia, Y., 2019. Semi-supervised adversarial model for benign–malignant lung nodule classification on chest CT. Med. Image Anal. 57, 237–248.

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

Yi, F., Jiang, T., 2019. Hard frame detection and online mapping for surgical phase recognition. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 449–457.

Yu, T., Mutter, D., Marescaux, J., Padoy, N., 2019. Learning from a tiny dataset of manual annotations: a teacher/student approach for surgical phase recognition. In: International Conference on Information Processing in Computer-Assisted Interventions.

Zheng, Q., Delingette, H., Ayache, N., 2019. Explainable cardiac pathology classification on cine MRI with motion characterization by semi-supervised learning of apparent flow. Med. Image Anal. 56, 80–95.

Zhu, J., Li, Y., Hu, Y., Ma, K., Zhou, S.K., Zheng, Y., 2020. Rubik's cube+: a self-supervised feature learning framework for 3D medical image analysis. Med. Image Anal. 101746.

Zisimopoulos, O., Flouty, E., Luengo, I., Giataganas, P., Nehme, J., Chow, A., Stoyanov, D., 2018. DeepPhase: surgical phase recognition in cataracts videos. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 265–272.