SV-RCNet: Workflow Recognition From Surgical Videos Using Recurrent Convolutional Network

Yueming Jin, *Student Member, IEEE*, Qi Dou, *Student Member, IEEE*, Hao Chen, *Student Member, IEEE*, Lequan Yu, *Student Member, IEEE*, Jing Qin, *Member, IEEE*, Chi-Wing Fu, *Member, IEEE*, and Pheng-Ann Heng, *Senior Member, IEEE*

***Abstract*— We propose an analysis of surgical videos that is based on a novel recurrent convolutional network (SV-RCNet), specifically for automatic workflow recognition from surgical videos online, which is a key component for developing the context-aware computer-assisted intervention systems. Different from previous methods which harness visual and temporal information separately, the proposed SV-RCNet seamlessly integrates a convolutional neural network (CNN) and a recurrent neural network (RNN) to form a novel recurrent convolutional architecture in order to take full advantages of the complementary information of visual and temporal features learned from surgical videos. We effectively train the SV-RCNet in an end-to-end manner so that the visual representations and sequential dynamics can be jointly optimized in the learning process. In order to produce more discriminative spatio-temporal features, we exploit a deep residual network (ResNet) and a long short term memory (LSTM) network, to extract visual features and temporal dependencies, respectively, and integrate them into the SV-RCNet. Moreover, based on the phase transition-sensitive predictions from the SV-RCNet, we propose a simple yet effective inference scheme, namely the prior knowledge inference (PKI), by leveraging the natural characteristic of surgical video. Such a strategy further improves the consistency of results and largely boosts the recognition performance. Extensive experiments have been conducted with the *MICCAI 2016 Modeling and Monitoring of Computer Assisted Interventions Workflow Challenge* dataset and *Cholec80* dataset to validate SV-RCNet. Our approach not only achieves superior performance on these two datasets but also outperforms the state-of-the-art methods by a significant margin.**

***Index Terms*— Recurrent convolutional network, surgical workflow recognition, joint learning of spatio-temporal features, very deep residual network, long short-term memory.**

Manuscript received October 16, 2017; revised December 16, 2017; accepted December 18, 2017. Date of publication December 27, 2017; date of current version May 1, 2018. This work was supported in part by the Research Grants Council of Hong Kong Special Administrative Region under Project CUHK 14202514 and Project CUHK 14203115, in part by the Shenzhen Science and Technology Program under Project JCYJ20170413162617606, and in part by the Hong Kong Polytechnic University under Project 1-ZE8J. *Yueming. Jin and Qi. Dou contributed equally to this work. (Corresponding author: Qi Dou.)*

Y. Jin, Q. Dou, H. Chen, L. Yu, C.-W. Fu, and P.-A. Heng are with the Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong (e-mail: ymjin@cse.cuhk.edu.hk; qdou@cse.cuhk.edu.hk).

J. Qin is with the Centre for Smart Health, School of Nursing, The Hong Kong Polytechnic University, Hong Kong.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMI.2017.2787657

I. Introduction

AIMING to improve the quality of patient treatment, modern operating rooms are in requirement of context-aware systems to monitor surgical processes [1], [2], schedule surgeons [3], [4] and enhance coordination among surgical teams [5]. Particularly, automatic workflow recognition has become a key component when developing the context-aware systems. Furthermore, if the workflow recognition can be intra-operatively performed online, the real-time recognition can interpret specific activity currently performing, which helps to alert surgeons when approaching possible complications [6], to reduce their operative mistakes and to support decision making [2], especially for less experienced surgeons.

Diverse attempts have been made to recognize the surgical workflow or phase, by using various information, including binary instrument usage signals [7], RFID tags [8], data acquired via sensors on tool tracking devices [9], and surgical robots [10]. However, collecting these signals mostly requires tedious manual annotation or extra equipment installation, which would introduce extra workload in the surgery process [2]. Therefore, recent researches have explored to identify the workflow purely based on video data routinely collected during the surgical process [2], [6], [11]. Apart from the merit of avoiding auxiliary devices, automated workflow recognition from surgical videos is also useful for surgeon skill evaluation [12] and documentation of the surgical video databases, given that the current practice of doing manual indexing is tedious and time-consuming [11].

However, purely using video scenes to automatically recognize surgical phase is quite challenging. First, there is limited inter-class variance between different phases while significant intra-class variance within the same phase (see Fig. 1 (a) and (b)). Second, severe scene blur occurs due to the camera motion and the gas produced during the surgery, which increases the recognition difficulty (see Fig. 1 (c)). Third, in the complex surgical procedures, the camera may not always focus on the surgical scenes, introducing additional noise and artifacts into the recorded videos (see Fig. 1 (d)).

To meet these challenges, lots of studies have been dedicated to extracting discriminative visual features from video frames and modeling the temporal dependencies among frames to improve the recognition accuracy. In terms of visual feature extraction, early studies utilized hand-crafted features, such as intensity and gradient [12], shape, color and texture-based

0278-0062 © 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

Authorized licensed use limited to: Johns Hopkins University. Downloaded on February 18,2022 at 04:18:34 UTC from IEEE Xplore. Restrictions apply.

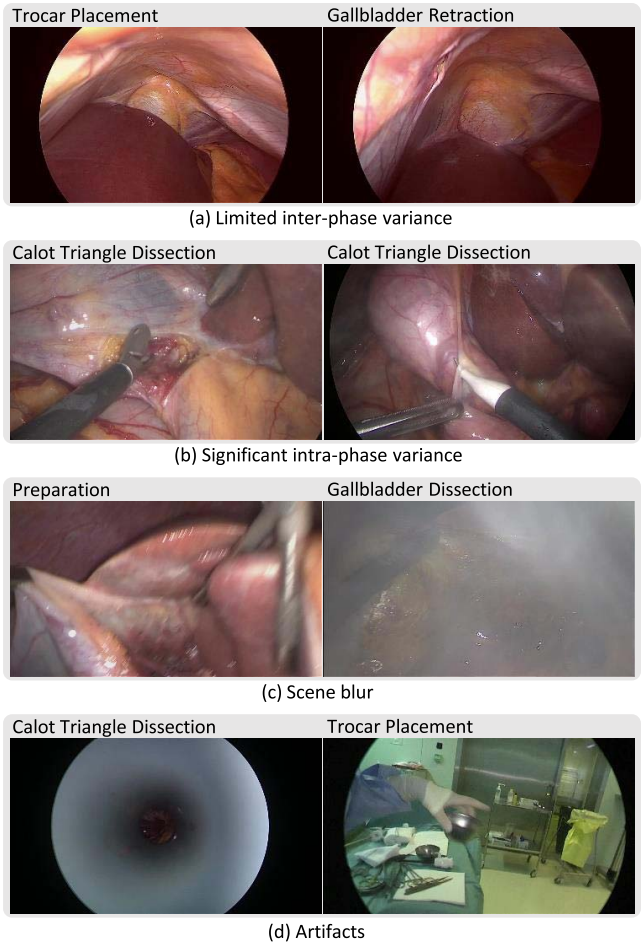

Fig. 1. Illustration of various challenges for automated workflow recognition from surgical videos. The text on the top-left corner of each image indicates which phase it belongs to. With each row, from top to bottom, we present the challenges of (a) limited inter-phase variance, (b) significant intra-phase variance, (c) scene blur due to camera motion and gas, and (d) artifacts.

descriptors [13]. However, it would be insufficient for these low-level features to represent the complicated surgical visual appearance [14]. With the revolution of deep learning and its successful applications on medical imaging [15]–[17], recent methods proposed to enhance the feature discrimination capability by employing convolutional neural networks (CNNs) [11]. Meanwhile, given that the surgical video is actually a form of sequential data, leveraging the temporal information and effectively capturing the sequential dynamics are crucial for accurate workflow recognition. A number of approaches have also been proposed in this direction by utilizing dynamic time warping [7], [12], conditional random field [18], and derivations of hidden markov model (HMM) [19], [20]. Specifically, the state-of-the-art performance of surgical workflow recognition was achieved by Twinanda *et al.* [11], who constructed a 9-layer CNN for visual features and designed a two-level hierarchical HMM for modeling temporal information.

However, it is still challenging for existing methods to fully solve this problem and there are great potentials to improve the automatic recognition performance for the following reasons. First, the previously used visual features, either hand-crafted or shallow CNN based, are still far from sufficient to represent the complicated visual characteristics of the frames in surgical videos. In addition, when exploiting the temporal information, most traditional methods rely on linear statistical models with pre-defined dependences, which are incapable of precisely representing the crucial yet subtle motions in the surgical videos, especially for frame series with strong non-linear dynamics. Second, and more importantly, most existing methods harness visual and temporal information separately, i.e. first using visual features with classifiers to predict each frame, and then using temporal dependencies to refine the results. In this way, visual features are unable to play a role in the temporal model and therefore such a scheme hardly benefits from the spatio-temporal information. Third, due to the above-mentioned two reasons, we analyze and find that it would be difficult for previous methods to sensitively identify and locate the transition frames (i.e., when jumping from one phase to another), while recognizing which is very important to achieve accurate and consistent workflow recognition results.

In this paper, we propose to process surgical videos with a novel recurrent convolutional network, termed as SV-RCNet, to comprehensively address the above challenges for accurate surgical workflow recognition. Our SV-RCNet conducts the workflow recognition in online mode, and employs state-of-the-art deep learning networks to extract visual features and model the temporal dependencies. Specifically, we exploit the very deep residual network, the ResNet [21], to extract highly discriminative visual features from the video frames. The importance of network depth for extracting discriminative features has been manifested by both computational theories [22], [23] and practical applications [24]–[26]. We further propose to learn the temporal dependencies by utilizing the long short term memory (LSTM) network. It is powerful in handling sequential data by non-linearly modeling long-range temporal dependencies [27], and has been successfully applied to many challenging tasks [28]–[30]. More importantly, SV-RCNet seamlessly integrates the ResNet and the LSTM network, so that we can jointly train them in an end-to-end manner to generate high-level features that encode both spatio (visual) and temporal information. Particularly, the spatio-temporal features learned by SV-RCNet are sensitive to motions in surgical videos and can precisely identify the phase transition frames. Considering that the results produced from SV-RCNet are transition-sensitive and the surgical videos are well-structured, we design a simple yet effective scheme called prior knowledge inference (PKI) to refine the SV-RCNet output. Our PKI strategy is tailored to make use of the natural characteristics of surgical videos and can greatly improve the recognition accuracy.

Our main contributions are summarized as follows:

1) We present a novel framework, i.e., SV-RCNet, to accurately recognize the workflow from surgical videos. Compared with previous methods that utilize visual and temporal information independently, the SV-RCNet can learn high-level representations that encode both visual features and temporal dependencies in an end-to-end architecture for improving the recognition accuracy.

Authorized licensed use limited to: Johns Hopkins University. Downloaded on February 18,2022 at 04:18:34 UTC from IEEE Xplore. Restrictions apply.

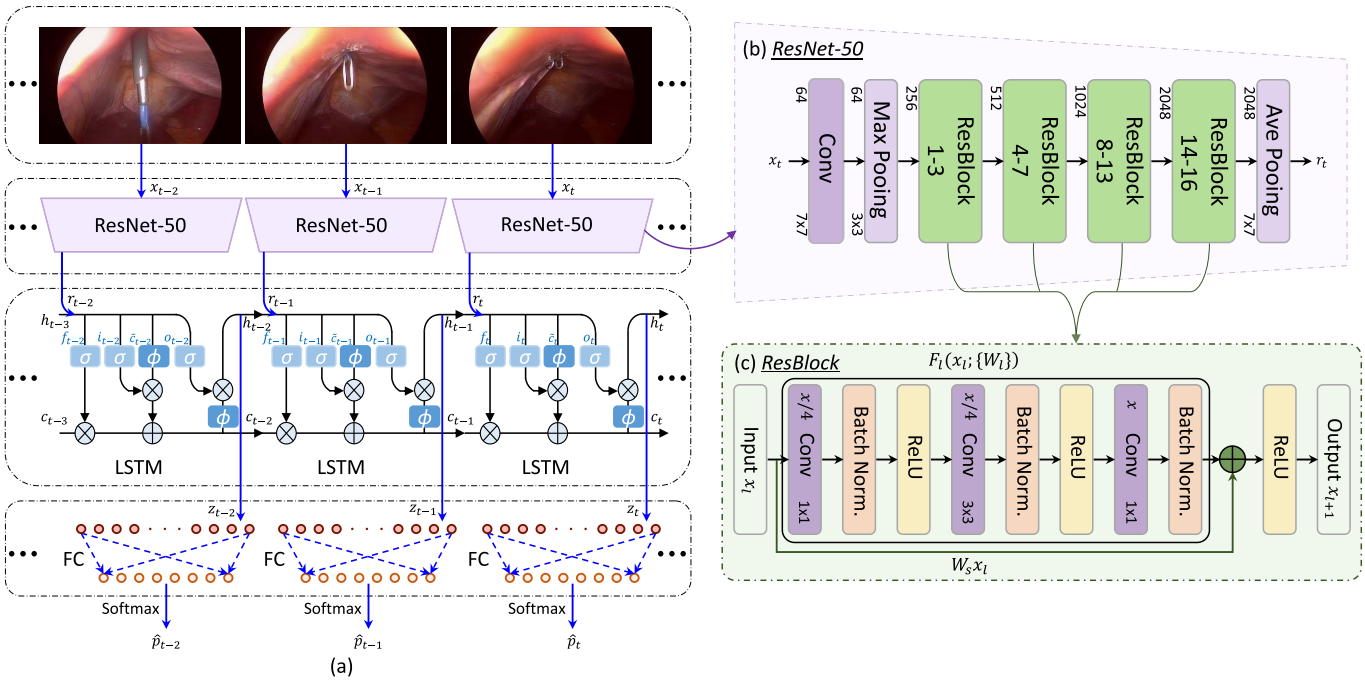

Fig. 2. (a) An overview of the proposed SV-RCNet for workflow recognition from surgical videos. The LSTM networks are instantiated by diagrams to indicate how temporal information is modeled. (b) The architecture of the ResNet to extract visual features from video frames. (c) Illustration of the structure within the residual block.

2) To enhance the discrimination capability of SV-RCNet, we integrate a very deep ResNet and a LSTM network to learn visual and temporal features, respectively, which can produce more representative features compared with traditional methods for surgical video analysis.
3) Based on the high-quality output from the SV-RCNet and the well-ordered structure of surgical videos, we develop the PKI scheme to enhance the consistency of phase predictions and largely improve the recognition accuracy.
4) We extensively evaluate our proposed method on *MICCAI 2016 Modeling and Monitoring of Computer Assisted Interventions Workflow Challenge*. Our achieved results ranked the first in the challenge, outperforming other approaches by a significant margin. In addition, we validate our method on a larger surgical video dataset, i.e., *Cholec80* dataset. Our approach achieved superior performance over the state-of-the-art approaches.

The remainder of this paper is organized as follows. We elaborate our methods in Section II, and report the experimental results in Section III. We further discuss and analyze our method in Section IV. Section V finally draws the conclusions. The source codes and relevant supporting documents can be found on our project website.[1]

II. Methodology

The overview of our proposed SV-RCNet is illustrated in Fig. 2. We exploit a very deep ResNet to extract discriminative visual features from each frame and harness a LSTM network to model the temporal information of sequential frames. More importantly, we seamlessly integrate these two components to form an end-to-end recurrent convolutional network so that the complementary information of the visual and temporal features can be sufficiently encoded for more accurate recognition.

A. Highly Discriminative Visual Descriptor Extraction

Extracting highly discriminative visual features from each frame of the input video is crucial for accurate recognition and forms the basis of our SV-RCNet. It is quite challenging to obtain features with powerful discrimination capability considering the complex surgical environments. Different from previous solutions which utilized either hand-crafted features [12], [13] or shallow CNNs [11], we propose to exploit a very deep ResNet [21] to tackle this crucial while challenging task.

As demonstrated in Fig. 2 (b), our deep residual network is composed of a set of residual blocks. For the l-th residual block B_l, we use x_l and x_{l+1} to respectively denote its input and output representations. Rather than expecting the stacked layers to fit a complicated underlying transformation $x_{l+1} = \mathcal{H}_l(x_l)$, the residual learning aims to ease the optimization difficulty by explicitly making these layers approximate a residual mapping:

$$x_{l+1} = \mathcal{W}_s x_l + \mathcal{F}_l(x_l; \{W_l\}), \tag{1}$$

where $\mathcal{F}_l$ is the residual mapping function; $\{W_l\}$ denotes the set of weights associated with the residual block B_l; the $\mathcal{W}_s$ is an identity mapping matrix to linearly match the input/output dimensions. The detailed construction of each residual block is shown in Fig. 2 (c). In our implementation, each residual block contains three convolutional layers, each followed by a batch normalization (BN) layer and a ReLU

[1] https://github.com/YuemingJin/SV-RCNet

Authorized licensed use limited to: Johns Hopkins University. Downloaded on February 18,2022 at 04:18:34 UTC from IEEE Xplore. Restrictions apply.

non-linearity layer. The shortcut identity mapping and element-wise addition are performed between the last BN layer and ReLU layer.

After constructing the residual blocks, we can hierarchically stack the blocks to substantially increase the network depth. Finally, we construct a 50-layer ResNet with one convolutional layer and one max pooling layer added at the beginning of the network as pre-layers to perform downsampling. The ResNet ends with a 7×7 average pooling layer to extract the global features from each frame and finally outputs a 2048-dimensional feature vector. Interested readers are suggested to refer to [21] for basic principles of residual learning. Note that the visual descriptors obtained from the ResNet are directly connected to the LSTM units in our SV-RCNet.

B. Effective Temporal Information Modeling

Due to the sequential nature of video data, temporal information provides valuable contextual clues for recognizing phases in a surgical procedure. For example, single frames from different phases may take very similar appearance and hence are difficult to be distinguished purely based on visual appearance. In contrast, if we can jointly consider its dependencies with adjacent past frames, recognizing the phase of current frame would be greatly eased.

Instead of employing traditional models, e.g. HMM, we propose to tap into the temporal dimension of the surgical video data using the LSTM [31], [32], which has been demonstrated as a very powerful tool to model temporal concepts. In our SV-RCNet, we sequentially input the visual descriptors obtained from the ResNet into the LSTM network and harness its memory cells to maintain the temporal information of past frames and then employ the temporal dependencies for better recognition.

Fig. 2 (a) illustrates the LSTM units used in our SV-RCNet [33]. The LSTM unit employs three gates, i.e., an input gate i_t, a forget gate f_t and an output gate o_t, to modulate the interactions between the memory cell c_t and its environment. The input gate i_t controls how much of new information $\tilde{c}_t$ to be stored to the memory cell. The forget gate f_t enables the memory cell to throw away previously stored information. In this regard, the memory cell c_t is a summation of the incoming information modulated by the input gate i_t and previous memory modulated by the forget gate f_t. The output gate o_t allows the memory cell to have an effect on the current hidden state and output or block its influence. At timestep t, given input r_t (ResNet visual descriptors in our task), hidden state h_{t-1}, and memory cell c_{t-1}, the LSTM unit updates with following equations:

$$\begin{aligned} i_t &= \sigma(W_{ri} r_t + W_{hi} h_{t-1} + b_i), \\ f_t &= \sigma(W_{rf} r_t + W_{hf} h_{t-1} + b_f), \\ o_t &= \sigma(W_{ro} r_t + W_{ho} h_{t-1} + b_o), \\ \tilde{c}_t &= \phi(W_{rc} r_t + W_{hc} h_{t-1} + b_c), \\ c_t &= f_t \odot c_{t-1} + i_t \odot \tilde{c}_t, \\ h_t &= o_t \odot \phi(c_t), \end{aligned} \tag{2}$$

where the hyperbolic tangent function $\phi(a) = \frac{e^a - e^{-a}}{e^a + e^{-a}}$ squashes the activations into $[-1, 1]$, and the sigmoid nonlinear function $\sigma(a) = \frac{1}{1+e^{-a}}$ squashes the activations into $[0, 1]$ for generating the three gates. The set of $\{W\}$ and $\{b\}$ respectively denote the weights and bias terms. The $\odot$ is element-wise multiplication involving computations with gates. The memory cell and all the gates have the same vector size and we initialize h_0 as $\boldsymbol{0}$.

C. End-to-End Learning of Recurrent Convolutional Network

In order to take full advantages of the complementary information of visual and temporal features, superior to existing methods in which the visual and temporal features are learned and utilized independently, we propose a novel recurrent convolutional network, i.e. the SV-RCNet, by seamlessly integrating the deep ResNet for visual descriptors extraction and the LSTM network for temporal dynamics modeling. Note that the inputs of our SV-RCNet are video clips rather than single frames so that both visual and temporal information can be sufficiently utilized and therefore cooperatively enhance the discrimination capability of our SV-RCNet. We train the SV-RCNet in an end-to-end manner, where the parameters of ResNet and LSTM network are jointly optimized towards accurate surgical workflow recognition.

Most of surgical videos contain quite long sequences since each video records the entire surgical operation, and workflow recognition task requires the model to classify the phase for each frame. With these considerations, instead of inputting the complete video into the network [32], [34], we propose to cut the surgical video into short video clips and conduct truncated backpropagation which alleviates the training difficulty and the limitation of computational memory. Specifically, to recognize the surgical phase at timestep t under online mode, we extract a video clip containing the current frame and a set of its former frames. The frame sequence in the video clip is denoted by $\boldsymbol{x} = \{x_{t'}, \ldots, x_{t-1}, x_t\}$ with the length of the sequence as $t - t'$. We denote the ResNet by U_β with weights β. The ResNet produces a representative fixed-length visual descriptor for each single frame x_j, represented as $r_j = U_\beta(x_j)$. The visual features $\boldsymbol{r} = \{r_{t'}, \ldots, r_{t-1}, r_t\}$ of the video clip are sequentially fed into the LSTM network, which is denoted by V_θ with parameters θ. With input r_t and previous hidden state h_{t-1}, the LSTM calculates the output z_t and the updated hidden state h_t as $z_t = h_t = V_\theta(r_t, h_{t-1})$. Note that the parameters θ are shared among every timestep. In this regard, we can learn the generic temporal dynamics from the video clip and simultaneously prevent the parameter scale from growing in proportion to the length of the video clip. Finally, the prediction probability of frame x_t is yielded by forwarding the output z_t to a softmax layer:

$$\hat{p}_t = Softmax(W_z z_t + b_z), \tag{3}$$

where W_z and b_z respectively denote the projection matrix and bias term, $\hat{p}_t \in \mathbb{R}^C$ is the prediction vector with C denoting the number of classes (the number of phases in our task).

Let $\hat{p}_t^c$ be the c-th element of $\hat{p}_t$, which represents the predicted probability of frame x_t belonging to the class c, and let l_t be the ground truth label of frame x_t, the negative

Authorized licensed use limited to: Johns Hopkins University. Downloaded on February 18,2022 at 04:18:34 UTC from IEEE Xplore. Restrictions apply.

log-likelihood loss of the frame at time t can be calculated as:

$$\ell(x_t) = -\log \hat{p}_t^{c=l_t}(V_\theta(U_\beta(\boldsymbol{x}))). \tag{4}$$

During the training, the losses for each single frame in the video clip $\boldsymbol{x}$ are computed and summed. Let $\mathcal{X}$ represent the training database containing N clip samples, with $\boldsymbol{x} \in \mathcal{X}$ being one video clip in the database, the overall joint loss function can be formulated as:

$$\begin{aligned}\mathcal{L}(\mathcal{X};\beta,\theta) &= \frac{1}{N}\sum_{\boldsymbol{x}\in\mathcal{X}}\ell(\boldsymbol{x}) \\ &= -\frac{1}{N}\sum_{\boldsymbol{x}\in\mathcal{X}}\sum_{\tau=t'}^{\tau=t}\log \hat{p}_\tau^{c=l_\tau}(x_{t':\tau}, h_{t':\tau-1};\beta,\theta).\end{aligned} \tag{5}$$

We now look into the training procedure and scrutinize how our end-to-end trainable SV-RCNet leverages both visual and temporal features, as well as their interactions, to enhance the discrimination capability of the network. In the feedforward procedure, the SV-RCNet sequentially inputs the visual descriptors of video frames obtained from the ResNet into the LSTM network and then the LSTM network can model the temporal dependencies of these frames based on these visual features. On the other hand, during the backpropagation procedure, we jointly optimize the ResNet parameters β and the LSTM parameters θ. In this procedure, the temporal information can be considered as a guidance when updating the parameters of the ResNet.

By employing the stochastic gradient descent, the parameters are updated by computing their gradients $\nabla\mathcal{L}(\mathcal{X};\beta,\theta)$ towards the loss. Specifically, denoting the learning rate of ResNet as λ and the learning rate of LSTM as η, we can update the weights $\{\beta,\theta\}$ according to following equations:

$$\theta \leftarrow \theta - \eta\frac{\partial\mathcal{L}}{\partial\theta},\ \beta \leftarrow \beta - \lambda\frac{\partial\mathcal{L}}{\partial\beta}. \tag{6}$$

In the backpropagation procedure, the gradients first flow into the LSTM network V_θ, where the $\frac{\partial\mathcal{L}}{\partial\theta}$ is calculated as follows:

$$\frac{\partial\mathcal{L}}{\partial\theta} = \frac{\partial\mathcal{L}}{\partial z_t}\frac{\partial z_t}{\partial\theta} = \frac{\partial\mathcal{L}}{\partial z_t}\frac{\partial V_\theta(r_t, h_{t-1})}{\partial\theta}. \tag{7}$$

With the gradients propagating backwards, the ResNet parameters β are optimized involving the parameters θ:

$$\frac{\partial\mathcal{L}}{\partial\beta} = \frac{\partial\mathcal{L}}{\partial r_t}\frac{\partial r_t}{\partial\beta} = \frac{\partial\mathcal{L}}{\partial z_t}\frac{\partial z_t}{\partial r_t}\frac{\partial r_t}{\partial\beta} = \frac{\partial\mathcal{L}}{\partial z_t}\frac{\partial V_\theta(r_t, h_{t-1})}{\partial r_t}\frac{\partial U_\beta(x_t)}{\partial\beta}. \tag{8}$$

From the end-to-end training process, we can find that the proposed SV-RCNet makes parameters β and θ learn both visual and temporal information while preserving their respective advantages. That is, the learning of the visual features is influenced by the captured temporal dynamics, and vice versa.

D. Prior Knowledge Inference for Consistency Enhancement

By seamlessly integrating temporal information, the features learned by SV-RCNet enable the predicted results of the whole surgical video to be smoother. However, a surgical video usually contains a number of resting frames, frames with slight motions, and frames with various artifacts in the middle of each phase, which are difficult to be accurately recognized. Fortunately, we find that, different from the natural videos, most surgical video contents are better structured and well ordered, because the surgeons are requested to perform surgeries according to specified workflows as well as instructions. For example, Fig. 3 (b) summarizes the phase transition conditions of *MICCAI 2016 Modeling and Monitoring of Computer Assisted Interventions Workflow Challenge* dataset, referred to as *M2CAI Workflow Challenge* dataset. Specifically, from P0 to P4, these phases are defined sequentially. From P4 to P7, there exists no linearly sequential alignment, yet we can still observe the order information to some extent, for example, P7 cannot happen before P5 triggers.

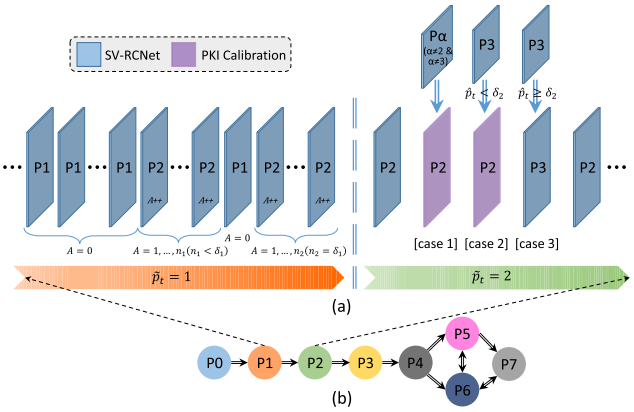

Fig. 3. (a) Illustration of the PKI algorithm, presenting the determination process of the phase prior $\tilde{p}_t$ (left) and calibration of the SV-RCNet predictions according to $\hat{p}_t$ (right). The outputs from SV-RCNet and the calibrated results are denoted in blue and purple colors, respectively. (b) The order information of the phases defined in the M2CAI Workflow Challenge dataset.

With above considerations, one idea should be investigated: by tracing the workflow and instantaneously infer the phase of current frame based on predictions of previous frames, whether we can acquire useful prior knowledge that would greatly help to calibrate those wrong predictions of phase internal frames (yellow arrows in Fig. 6). Actually, during the phase transition period (the beginning of each phase), thanks to the changing of key actions which bring in richer temporal information, SV-RCNet can accurately recognize transition sequences in-between the phases. Examples are shown by the pink arrows in the Fig. 6. In other words, SV-RCNet is motion sensitive and can precisely locate the phase transition points. To this end, we propose a simple yet effective inference scheme following our SV-RCNet, namely prior knowledge inference, aiming to enhance the prediction consistency. With the PKI, we successfully leverage both the well-ordered characteristics of surgical video and the transition-sensitive outputs from our SV-RCNet to boost the workflow recognition performance.

The logic of the PKI algorithm is illustrated in Fig. 3, where we elaborate how PKI works in the transition point from P1 to P2 as an example and other transition points share the same principles. Fig. 3 (a) mainly explains how to determine phase prior $\tilde{p}_t$. We denote the network's phase prediction for a video frame x_t by $\gamma_t \in \{0, 1, ..., C\}$, where $C = 7$ in the M2CAI Workflow Challenge dataset. To provide prior knowledge for

Authorized licensed use limited to: Johns Hopkins University. Downloaded on February 18,2022 at 04:18:34 UTC from IEEE Xplore. Restrictions apply.

the current frame x_t, a prior state collector (denoted by $\mathcal{S}$) is employed to record the phase predictions of all its past frames:

$$\mathcal{S} = (\gamma_0, \gamma_1, ..., \gamma_{t-2}, \gamma_{t-1}). \quad (9)$$

With the prior knowledge collected by $\mathcal{S}$, we infer the phase prior $\tilde{p}_t$ that the current frame x_t is most likely to be. More specifically, we set accumulators A for each possible phase to respectively count the number of frames classified into that phase. The possible phase is P2 in Fig. 3 (a) according to the defined phase transition in Fig. 3 (b). To ensure the accuracy and robustness of $\tilde{p}_t$, the accumulator A for each possible phase only increases when sequential frames are continuously predicted into that phase. Otherwise, A is reset to zero and a new round of accumulation for that phase is invoked.

Finally, the phase prior $\tilde{p}_t$ is determined when the accumulation of this phase reaches a threshold δ_1. In Fig. 3 (a), $\tilde{p}_t$ changes from P1 to P2 only when the A in succession increases to threshold δ_1. Note that at each new timestep, the PKI just updates $\mathcal{S}$, and the phase prior derived from the $\mathcal{S}$ is obtained through the same strategy.

The obtained phase prior $\tilde{p}_t$ is then employed to calibrate the phase prediction of current frame, as illustrated in Fig. 3 (a). If the x_t is classified as neither $\tilde{p}_t$ nor any one of its potential next phases defined in the workflow, which means there is a high possibility that x_t is misclassified into another phase by SV-RCNet due to its indistinguishable current appearance, the PKI calibrates its prediction into $\tilde{p}_t$ to maintain the consistency of prediction results. As shown in the first case of Fig. 3 (a), if the prediction from SV-RCNet is neither phase 2 nor 3, PKI will modify it into $\tilde{p}_t = 2$. In case that the x_t is classified into one of potential next phases, the PKI will check the confidence of this prediction to decide whether it should be maintained. If the prediction probability is lower than a threshold δ_2, the PKI will amend the prediction as $\tilde{p}_t$; otherwise, it will keep the prediction, which are shown in the second and third cases of Fig. 3 (a). The hyper-parameters in the PKI were determined using grid search on a validation subset of the datasets.

E. Training Details of SV-RCNet

In order to effectively train the SV-RCNet, considering the parameter scale of the ResNet is much larger than that of the LSTM network, we first pre-train the ResNet to learn reliable parameters for the following initialization in the overall network. Leveraging the effective generalization capability of transfer learning, we initialize our ResNet with weights trained on the ImageNet dataset [21]. In this stage, we re-sample the original videos to balance training samples of different phases and then resize the frames from the original resolution of 1920×1080 into 250×250 to dramatically save memory and reduce network parameters. The images are further augmented with 224×224 cropping, mirroring and rotation to expand the training database.

After obtaining the pre-trained ResNet model, SV-RCNet which integrates visual and temporal information is trained in an end-to-end manner to convergence. Note that while we use the pre-trained parameters of ResNet as its initialization, the parameters of LSTM network are randomly initialized from Gaussian distribution ($\mu = 0$, $\sigma = 0.01$). Hence, the learning rate of the LSTM is set ten times as that of the ResNet. For training data preparation, we downsample the original videos from 25fps to 5fps to enrich temporal information in the video clips. The resolution of the frames is also resized as 250×250 with the same augmentation strategies. The length of the clip is set to around 2 seconds and the sampling stride is set to 3.

Our framework is implemented with C++ and Python based on the *Caffe* [35] deep learning library, using a TITAN X GPU for acceleration. The hyper-parameters in the network are as follows: momentum=0.9, weight decay=0.005, LSTM dropout rate=0.5. The learning rates are initially set as 0.0005 for ResNet and 0.005 for LSTM, and are divided by a factor of 10 every $20k$ iterations. It took around one day to train the entire framework into convergence. During inference, our framework processes one frame within 0.1 second, which demonstrates its potential to be used for online surgical workflow recognition.

TABLE I
PHASES AND THEIR TIME STATISTICS IN CHOLECYSTECTOMY VIDEOS

ID	Phase	Duration(min)
P0	Trocar Placement	2.2 ± 1.0
P1	Preparation	1.0 ± 0.9
P2	Calot Triangle Dissection	10.4 ± 5.6
P3	Clipping and Cutting	2.8 ± 1.9
P4	Gallbladder Dissection	7.5 ± 8.1
P5	Gallbladder Packaging	1.8 ± 3.2
P6	Cleaning and Coagulation	1.9 ± 1.6
P7	Gallbladder Retraction	4.5 ± 5.0

III. EXPERIMENTS

A. Dataset and Evaluation Metrics

We have extensively validated the proposed SV-RCNet on the public dataset of MICCAI 2016 Challenge on *Modeling and Monitoring of Computer Assisted Interventions*, referred to as *M2CAI Workflow Challenge*.[2] The 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. The names and time statistics of phases are listed in Table I. The dataset is divided into training set (27 videos) and testing set (14 videos). All our surgical workflow recognition experiments were performed in online mode. That is, when estimating the frame at time t, we access no future frame (i.e. frames at time larger than t).

To quantitatively analyze the performance of our method, we employed four metrics including the jaccard index (JA), precision (PR), recall (RE) and accuracy (AC). Among them, the JA and AC were used to evaluate the submissions of M2CAI Workflow Challenge while PR and RE are also commonly used metrics to evaluate video-based workflow recognition methods. The JA, PR and RE are calculated in

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

Authorized licensed use limited to: Johns Hopkins University. Downloaded on February 18,2022 at 04:18:34 UTC from IEEE Xplore. Restrictions apply.

phase-wise, defined as follows:

$$\text{JA} = \frac{|\text{GT} \cap \text{P}|}{|\text{GT} \cup \text{P}|}, \quad \text{PR} = \frac{|\text{GT} \cap \text{P}|}{|\text{P}|}, \quad \text{RE} = \frac{|\text{GT} \cap \text{P}|}{|\text{GT}|}, \tag{10}$$

where GT and P respectively denote the ground truth set and prediction set of one phase. After JA, PR, RE of each phase are calculated, we average these values over all the phases and obtain the corresponding measure of the entire video. The AC is directly calculated at video-level, defined as the percentage of frames correctly classified into its ground truth phase in the entire video.

B. Experiments on Depth of Convolutional Network

Extracting discriminative visual features is crucial for our task. The depth is a key factor in relevance to the performance of the CNNs. While a deeper network may produce more representative features, it will consume more computational resources and increase time complexity both for training and testing. In this regard, we performed extensive experiments to evaluate the impact of network depth on the performance of a convolutional network so that we can find a suitable network architecture to generate highly discriminative visual features while maintaining reasonable computation and time cost.

We implemented four convolutional networks with different depth, i.e. 22-layer GoogLeNet [24], 35-layer, 50-layer, and 101-layer residual networks. The difference among the three residual networks is the number of residual blocks, with 11, 16 and 33 ResBlocks (see Fig. 2 (c)), respectively. The training data of these networks were identical and all the networks were pre-trained based on ImageNet. Note that in these experiments, we obtained the prediction results directly from the outputs of these convolutional networks, which were purely based on the visual information of each frame.

Table II presents the performance of these networks. It is observed that all the residual networks achieved higher performance compared with the 22-layer GoogLeNet, demonstrating that increasing network depth and exploiting residual learning can effectively promote the model performance by extracting highly representative features. The impact of network depth can also be witnessed by the gradually improved results obtained from the 35-layer, 50-layer and 101-layer residual networks. In particular, the major metric JA has achieved 5% increase from the 35-layer ResNet to the 50-layer ResNet.

Nevertheless, we also notice that the performance growth rate tends to be slower as the increase of residual network depth. The accuracy improvement from 50-layer to 101-layer is far less significant than that from 35-layer to 50-layer. One of the underlying reasons might be that when a network is going into the deeper, it will encounter tougher optimization difficulties. For instance, the parameter scale of 101-layer ResNet is around twice as large as that of the 50-layer ResNet, resulting in more risks of overfitting and great increase in computing resource for training. More importantly, we found that when integrating the 101-layer ResNet with the LSTM network, the required computing resource is not affordable even by an advanced GPU and both training and testing time is relatively long, leading to the difficulty in practical application. Therefore, we chose the 50-layer ResNet for our SV-RCNet.

TABLE II
COMPARISON OF PHASE RECOGNITION PERFORMANCE USING NETWORKS WITH DIFFERENT DEPTH

Networks	Jaccard (%)	Precision (%)	Recall (%)	Accuracy (%)
GoogLeNet-22	49.0 ± 7.8	71.4 ± 13.7	68.0 ± 14.0	64.3 ± 18.0
ResNet-35	51.4 ± 10.6	69.5 ± 13.7	73.4 ± 11.5	72.8 ± 10.2
ResNet-50	56.4 ± 10.4	73.5 ± 13.0	76.8 ± 13.0	76.3 ± 8.9
ResNet-101	$\mathbf{57.7 \pm 10.5}$	$\mathbf{74.2 \pm 12.9}$	$\mathbf{77.4 \pm 10.8}$	$\mathbf{76.6 \pm 9.2}$

TABLE III
COMPARISON OF PHASE RECOGNITION RESULTS USING DIFFERENT TEMPORAL MODELING SCHEMES

Experimental Settings	Jaccard (%)	Precision (%)	Recall (%)	Accuracy (%)
ResNet-50	56.4 ± 10.4	73.5 ± 13.0	76.8 ± 13.0	76.3 ± 8.9
ResNet-50+HMM	59.9 ± 10.4	76.5 ± 13.1	78.9 ± 12.6	78.9 ± 9.3
ResNet-50+LSTM	60.8 ± 8.2	77.8 ± 8.6	77.8 ± 7.7	78.4 ± 9.4
SV-RCNet	$\mathbf{65.4 \pm 8.9}$	$\mathbf{81.0 \pm 8.3}$	$\mathbf{81.6 \pm 7.2}$	$\mathbf{81.7 \pm 8.1}$

C. Experiments on Different Temporal Modeling Schemes

How to effectively and sufficiently combine the visual and temporal features lies at the heart of the video-based surgical workflow recognition task. In the meanwhile, how to leverage the ordering information of surgical videos to further enhance the consistency of the results is also important in our task. To this end, we conducted extensive experiments on the dataset of M2CAI Workflow Challenge to validate our proposed method from various perspectives.

1) Effectiveness of the End-to-End Learning: In order to demonstrate the importance of end-to-end learning to extract discriminative features for this task, we first conducted a series of experiments by integrating the 50-layer ResNet with different temporal modeling methods, i.e., (1) pure 50-layer ResNet as a baseline, (2) 50-layer ResNet followed by HMM [36], (3) 50-layer ResNet followed by LSTM (separately trained), and (4) our SV-RCNet (end-to-end training). Note that the video clips input into the last two schemes have the same length for guaranteeing the experiment fairness.

The experimental results are listed in Table III. All schemes integrating temporal information achieve much better results than the pure 50-layer ResNet, demonstrating the importance of temporal cues for more accurate recognition. Furthermore, focusing on the temporal models, both HMM and separately trained LSTM learn temporal dependencies in an independent manner. In contrast, our proposed SV-RCNet is capable of producing features encoding both visual and temporal information via end-to-end learning. It is observed that our SV-RCNet achieves higher results than the above two methods, demonstrating the effectiveness of the spatio-temporal joint learning. Specifically, compared with separately trained ResNet-50+LSTM model, our end-to-end trainable SV-RCNet improves the JA from 60.8% to 65.4%, and similar improvements of PR, RE and AC are also observed. Overall, these results corroborate that through the joint optimization process, there is implicit interaction between the visual and temporal features and such an interaction produces beneficial effects on each other. By taking advantage of complementary information of visual and temporal information, our SV-RCNet can learn more discriminative and high-level spatio-temporal

Authorized licensed use limited to: Johns Hopkins University. Downloaded on February 18,2022 at 04:18:34 UTC from IEEE Xplore. Restrictions apply.

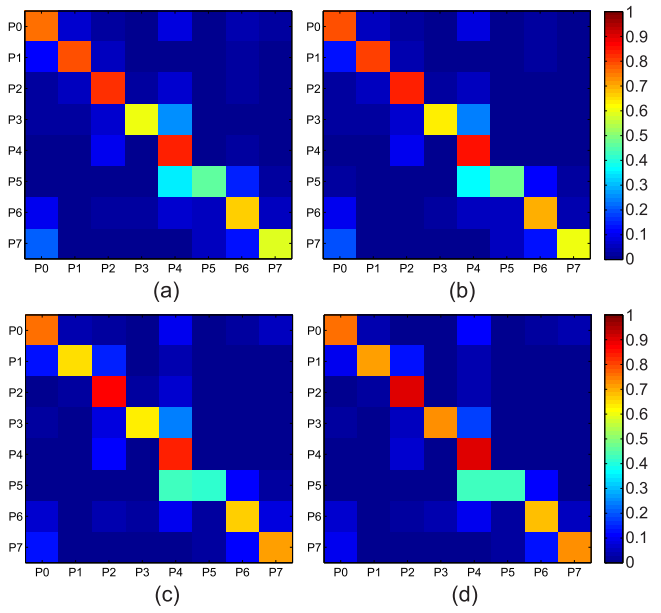

Fig. 4. Confusion matrices visualized by the color brightness for (a) ResNet-50, (b) ResNet-50+HMM, (c) ResNet-50+LSTM, and (d) SV-RCNet. The X and Y-axis represent predicted label and ground truth, respectively. Element (a, b) of each confusion matrix represents the empirical probability of predicting class a given that the ground truth is class b. The probability on diagonal indicates the recall for each phase.

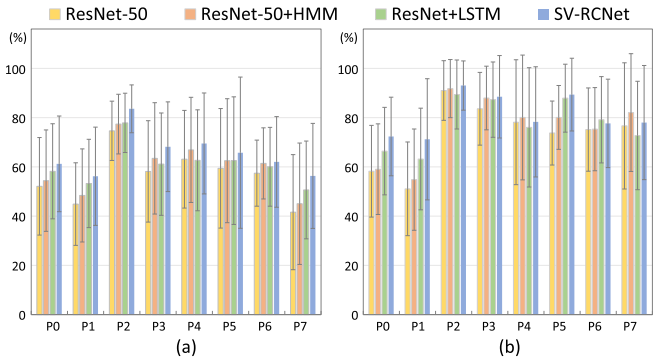

Fig. 5. The phase-level bar chart results of (a) Jaccard and (b) Precision for ResNet-50, ResNet-50+HMM, ResNet-50+LSTM, and SV-RCNet.

features which can increase the performance of workflow recognition in surgical videos.

To more comprehensively analyze effectiveness of the end-to-end learning mechanism, we further visualize the confusion matrices of the methods in Fig. 4. It is observed that, from (a) to (d), the performance rises with increasing on recall and decreasing on misclassification, especially for P7, indicating that while combining temporal information can boost the recognition accuracy, leveraging spatio-temporal features jointly learned from the proposed SV-RCNet can further improve the performance. In addition to using confusion matrix to indicate RE, we further draw a phase-level bar chart to illustrate the results of JA and PR in Fig. 5. We find that across all the 8 phases, the SV-RCNet dominates other schemes on the JA. For PR, the improvement from other schemes to SV-RCNet is especially significant (over 5%) for P0 and P1.

2) Importance of Joint Learning to PKI: Most surgical video contents are structured and ordered. PKI strategy is in particular designed to utilize such a natural characteristic of surgical video to improve the recognition performance. However, the efficacy of PKI heavily depends on the predictions provided by former networks according to the rationale behind the design. To investigate how large the influence can be, we performed the experiments integrating all the above methods with PKI. Besides showing the evaluation metrics in Table IV, the differences of JA (D-Jaccard) between each method and its counterpart in Table III are also shown to more clearly explain the impact degree.

TABLE IV
COMPARISON OF PHASE RECOGNITION RESULTS WITH PKI CALIBRATION CONNECTED

Experimental Settings	D-Jaccard (%)	Jaccard (%)	Precision (%)	Recall (%)	Accuracy (%)
ResNet-50+PKI	7.4	63.8 ± 15.1	79.2 ± 13.5	80.0 ± 12.0	82.9 ± 11.7
ResNet-50+HMM+PKI	7.6	67.5 ± 10.2	83.2 ± 9.8	80.9 ± 9.1	82.3 ± 17.2
ResNet-50+LSTM+PKI	9.3	70.1 ± 9.9	83.4 ± 7.9	81.7 ± 8.8	83.4 ± 19.9
SV-RCNet+PKI	**12.8**	$\mathbf{78.2 \pm 11.0}$	$\mathbf{88.1 \pm 7.8}$	$\mathbf{88.9 \pm 7.4}$	$\mathbf{90.7 \pm 6.9}$

First, compared with the counterparts in Table III, all the methods connected with the PKI strategy can greatly improve the recognition performance, demonstrating the effectiveness of the PKI in refining the results. More importantly, among those four methods, the improvements of JA (D-Jaccard) gradually increase, and SV-RCNet+PKI exceeds others by a large margin. Such phenomenon verifies that although PKI performs well in the calibration, the basis model is also crucial. The considerable improvement of PKI is greatly attributed to the good basis provided by SV-RCNet. In addition, see the results of ResNet-50+LSTM+PKI and SV-RCNet+PKI, where the only difference is w/o end-to-end learning of networks, the latter greatly improves the JA from 70.1% to 78.2%.

Next, we further experimentally investigated the underlying reason why the combination of SV-RCNet and PKI can achieve better results. According to the rationale of PKI, two points decide its efficacy, i.e. transition phase detection sensitivity to decide phase prior $\tilde{p}$ and probability confidence to calibrate wrong predictions. In Fig. 6, we illustrate prediction results of one complete surgical video from ResNet-50 and SV-RCNet. It can be clearly observed from the red arrows that SV-RCNet offers smoother and more accurate estimations at transition frames and provides a better basis for PKI. On the other hand, frequently jumped and garbled results generated by ResNet-50 may impede the performance gains of PKI by providing it incorrect transfer frames. Surgical videos usually contain quite long sequences, hence predictions at each phase transition cannot be shown in details by Fig. 6. In this regard, we present Fig. 7 to have a closer look at the performance during phase transitions ($\pm$500 frames), which more clearly exemplifies the difference between the two networks. We can find that the SV-RCNet can produce much better results during the transition period. Moreover, the high-quality prediction probability from the SV-RCNet, as the other determinant to PKI effectiveness, plays an important role in the PKI. We present some typical prediction results in Fig. 8, where the prediction probabilities for frame x_t using SV-RCNet and ResNet-50 are indicated in top-right corner. It is observed that integrating the LSTM into the end-to-end learning framework can greatly enhance the model's confidence towards correct predictions.

Authorized licensed use limited to: Johns Hopkins University. Downloaded on February 18,2022 at 04:18:34 UTC from IEEE Xplore. Restrictions apply.

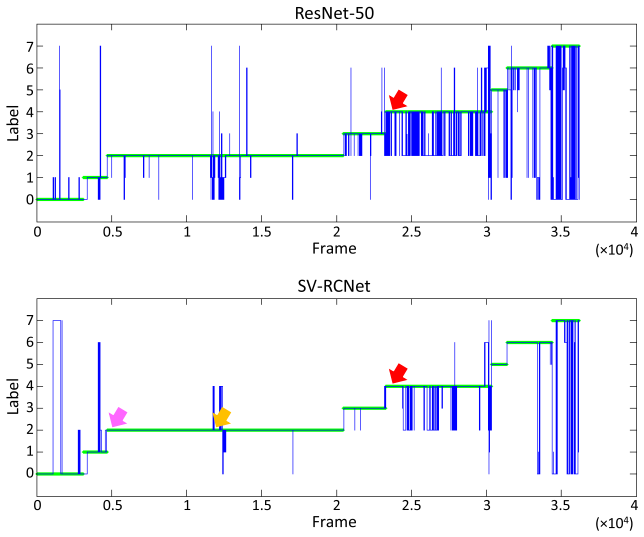

Fig. 6. Illustration of prediction results of one complete typical surgical video from ResNet-50 and SV-RCNet. The phase annotations are shown in green and predictions are shown in blue. Both of them are connected by line to show the performance more clearly. More result overlapping indicates higher prediction accuracy. The pink and yellow arrows represent continuous accurate predictions during transition period and the wrong predictions of internal frames from SV-RCNet, respectively. And red arrows are shown to compare the prediction sensitivities of transition frames between ResNet-50 and SV-RCNet.

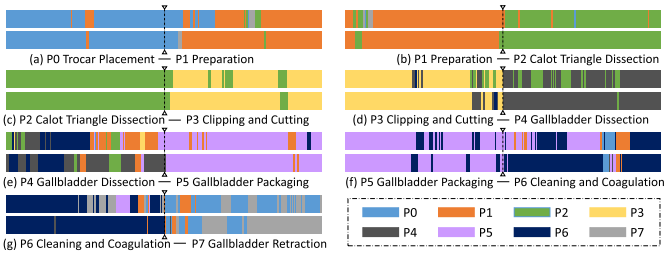

Fig. 7. Color-coded ribbon illustration for phase recognition results during phase transition time (± 500 frames) generated by ResNet-50 (above) and SV-RCNet (bottom), respectively. The texts under each pair of ribbons successively represent transited phase names of one random testing video. The transition points are indicated by triangles and dash lines.

3) Network Interpretation by Visualizing Attention Maps: In order to provide the insight of what the networks learn and investigate why joint optimization can improve the performance, we further visualize the attention maps (class activation maps) to interpret the model. Attention maps can exactly highlight the regions of an image that are important for discrimination by weighted summing the feature maps of last convolutional layer, which expose the implicit attention of networks on an image and intercept the learned information of networks [37], [38]. Specifically, Fig. 9 visualizes attention maps of three typical video clips from three different networks, i.e. (1) ResNet-50, (2) separately trained ResNet-50+LSTM model, (3) our SV-RCNet. We observe that the attention maps are quite reasonable and matched with their probability predictions and predicted classes, which hence can interpret the results in terms of what networks learn.

For the first video clip, the phase label of this video clip is P1. Attention maps from ResNet-50 are shown in the second row, which pay more attention to the fat region distributing in the bottom of the image. Hence, the predictions for the ground truth are rather low and these frames are mistakenly predicted as P0. The third row shows the attention maps from separately trained ResNet-50+LSTM model, where attentions of the network gradually transfer from fat region to the gall bladder and surgical instrument, which are important cues for workflow recognition. Accordingly, the predictions constantly increase and the third frame is eventually classified correctly. The last row displays the attention maps from our SV-RCNet, which successfully focuses the attentions on the key cues for all the three frames, demonstrating that, with the joint spatio-temporal training, the proposed SV-RCNet is capable of robustly generating more discriminative representations encoded both visual and temporal information.

Similarly, the attention maps in the second video clip with P3 phase label also demonstrate the significance of temporal clue and joint optimization of visual and temporal feature. See the third frame in this video clip, the important clue for classification (surgical tool: scissors) nearly disappears from the frame. Due to solely relying on the single frame information, ResNet-50 ignores the tool information and therefore wrongly classifies it as P2. In contrast, separately trained ResNet-50+LSTM model can be aware of such information. However, attention region is still scattered, hence this model cannot achieve rather high prediction confidence. Attention region of our SV-RCNet is on the contrary more compact and relevant to the right phase category, therefore our method produces higher prediction probability.

The attention maps in the third video clip with P7 phase label also clearly demonstrate the superiority of the spatio-temporal feature learning. It can be observed from the first two rows that ResNet-50 and separately trained ResNet-50+LSTM model pay attention to the surgical tool in the first two frames and the background when the tool nearly disappears. Although their attention regions contain some important visual information, since there exists low variance of frame scenes between P0 and P7 (see Fig. 8), these two models fail to recognize the frames based on the low-level features. In contrast, our SV-RCNet not only focuses the surgical tool, but also is aware of its motion path based on the spatio-temporal features. In other words, it can distinguish the motion direction of the tool, i.e. input to or retracted from the patient. Therefore, these frames with limited inter-phase variance can be correctly recognized. We also observed similar meaningful attention maps for other surgical phases during experiments.

From the attention maps, we can intuitively observe what models learn and focus on, therefore have a better understanding on why our method has better behaviors and performance than other networks. Our SV-RCNet is able to localize and concentrate on the discriminative image regions, which helps to correctly identify the phase with high prediction confidence.

D. Results of the M2CAI Workflow Challenge

Five teams took part in the M2CAI Challenge for workflow recognition held in conjunction with MICCAI 2016. The results are reported in Table V. Among the five teams, only our team exploited high level features jointly encoded both visual and temporal information produced by the SV-RCNet, while all other teams used visual features and temporal

Authorized licensed use limited to: Johns Hopkins University. Downloaded on February 18,2022 at 04:18:34 UTC from IEEE Xplore. Restrictions apply.

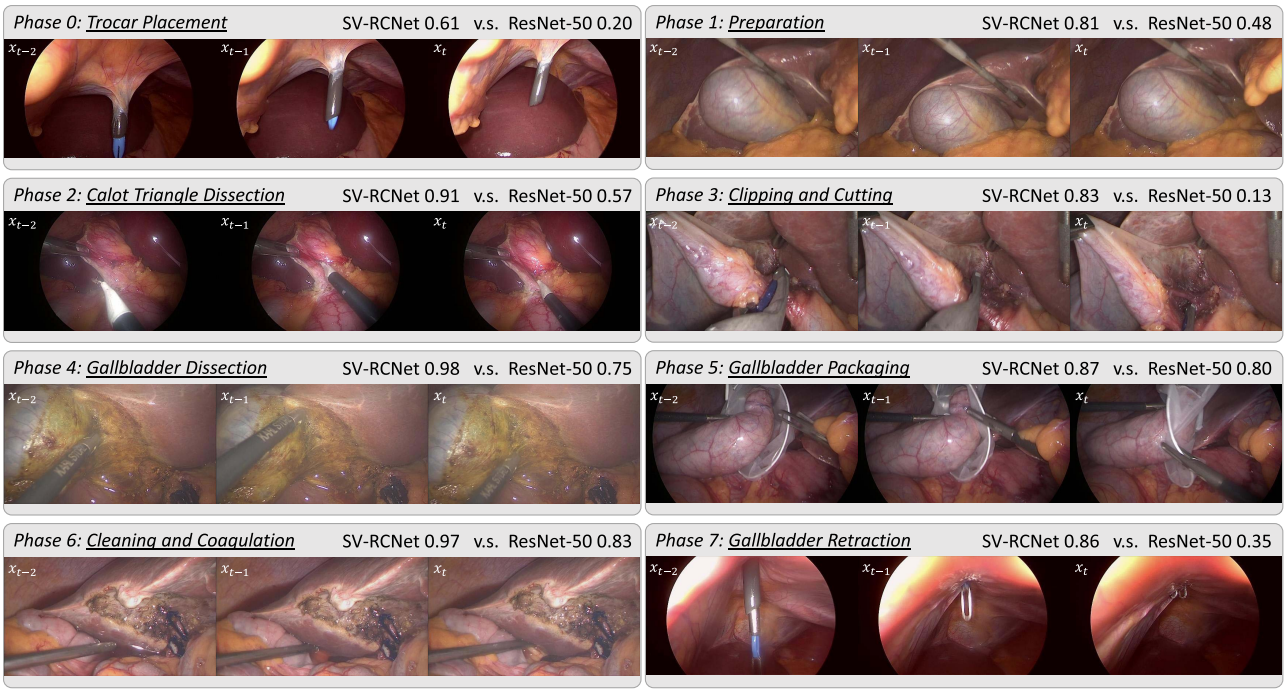

Fig. 8. Illustration of typical actions in different phases during the cholecystectomy procedure. For each phase, we present three continuous frames with the stride of 3 (same as our SV-RCNet setting) to capture the temporal information. In each of the eight section, the phase name is indicated in the top-left corner; the probability predictions towards the ground truth phase of frame x_t using SV-RCNet and ResNet-50 are presented in the top-right corner.

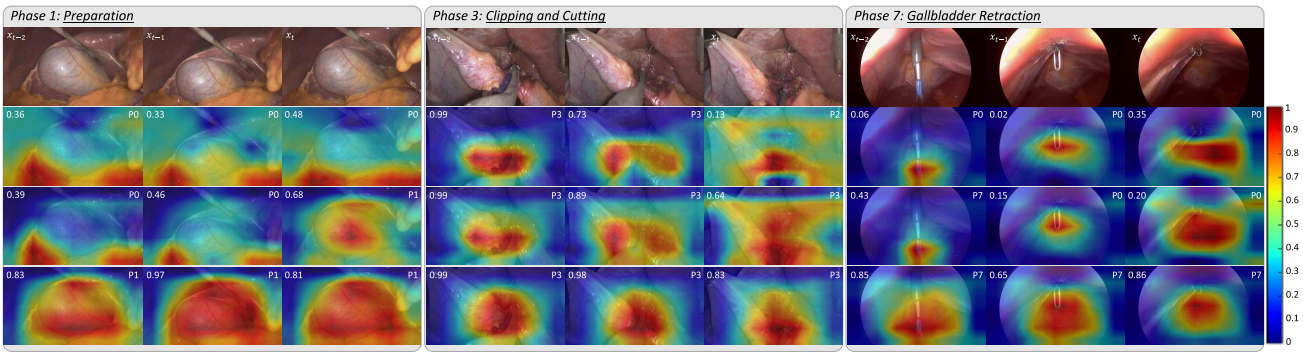

Fig. 9. Visualization of the attention maps of typical video clips indicating the discriminative regions captured by the networks. Three typical video clips are presented, from top to bottom: the input frames, attention maps from pure ResNet-50 model, attention maps from separately trained ResNet-50+LSTM model, and attention maps from our SV-RCNet model. In each of the attention maps, the probability prediction towards the ground truth phase is presented in the top-left corner while the predicted class is indicated in the top-right corner.

information sequentially and separately. Sahu *et. al.* [39] utilized an 8-layer AlexNet to extract visual features and modeled the temporal occurrences of surgical phases by fitting Gaussian distributions to further refine the prediction results. Dergachyova *et. al.* [40] employed hand-crafted visual features based on color, shape and texture, followed by a hidden semi-markov model (HsMM) to refine the results. Twinanda *et. al.* [41] also made use of an 8-layer AlexNet to extract visual features, followed by a hierarchical HMM for refinement, which is same as one method proposed in [11]. Cadène *et. al.* [36] leveraged a 200-layer ResNet, aiming to extract more representative visual descriptors. They further utilized HMM to refine the results. Some of them exploited different types of post-processing strategy for improving performance in fact. For example, Cadène *et. al.* [36] utilized average smoothing on the prediction.

The challenge evaluated the performance of different teams using the metrics of JA and AC. In addition, the challenge evaluation regulation relaxed the boundaries of the phases with a 10-second window, i.e. 250 frames, which means that it tolerates tiny early or late transitions. In Table V, we can find that our method achieved the best performance with both JA of 78.2% and AC of 90.7%. Our results outperformed all other approaches by a significant margin, exceeding the second ranking team by 6.3% on JA and 3.8% on AC, demonstrating the effectiveness of the proposed methods. Note that although the second ranking team employed a 200-layer ResNet, we achieved much better recognition results than them, corroborating the features encoded both visual and temporal information generated from our SV-RCNet are more discriminative and hence powerful than features only containing visual information in workflow recognition from

Authorized licensed use limited to: Johns Hopkins University. Downloaded on February 18,2022 at 04:18:34 UTC from IEEE Xplore. Restrictions apply.

TABLE V
PHASE RECOGNITION RESULTS OF DIFFERENT METHODS IN 2016 MICCAI M2CAI WORKFLOW CHALLENGE

Methods	Jaccard (%)	Accuracy (%)
Ours (SV-RCNet+PKI)	**78.2 ± 11.0**	**90.7 ± 6.9**
Cadene *et. al.* [36]	71.9 ± 12.7	86.9 ± 11.0
Twinanda *et. al.* [41]	64.1 ± 10.3	79.5 ± 12.1
Dergachyova *et. al.* [40]	51.5 ± 14.1	70.7 ± 6.1
Sahu *et. al.* [39]	45.0 ± 14.4	52.7 ± 13.8

surgical videos. Nevertheless, the better results of [36] than other methods further demonstrate that a deeper network may generate more discriminative visual features.

E. Results of the Cholec80 Dataset

To validate the extensibility of our framework, we further evaluated our method on the Cholec80 dataset,[3] which is publicly released by the same organizers. Compared with the dataset of M2CAI Workflow Challenge, the Cholec80 has the same resolution 1920 × 1080 and is also captured at 25fps, while it contains more cholecystectomy procedures (80 videos) and is annotated with 7 defined phases. Specifically, *P0:Trocar Placement* and *P1:Preparation* in Table I are merged into one phase and others remain. This dataset also contains tool annotations indicating the presence of 7 tools in an image.

We compared the recognition performance of our proposed method with the state-of-the-art approach, EndoNet, which utilized 9-layer CNN with two-level hierarchical HMM proposed by Twinanda *et al.* in [11]. EndoNet was designed for multiple tasks, i.e. the phase recognition task and the tool presentation detection task. Therefore, the network recognizes the workflow phase by leveraging both the phase annotations and tool annotations. In addition, we compared the performance with the method of PhaseNet, which was also proposed in [11] while solely utilizing the phase annotations. To guarantee the fairness of comparison, following the same process to dataset in [11], we split it into two subsets of equal size, with 40 videos as training set and the rest as testing set. Moreover, we only utilize phase annotations to train our network.

Experimental results are shown in Table VI. Here, we only list the results on three criteria, i.e. PR, RE and AC, since the other criterion, JA, is not reported in [11]. It is clearly observed that our method SV-RCNet outperforms PhaseNet by a large margin in terms of AC, i.e., 78.8% v.s. 85.3%. Moreover, although EndoNet includes the additional tool annotation information to do phase recognition, our SV-RCNet still achieves better performance than such a state-of-the-art method, improving the AC from 81.7% to 85.3%. Furthermore, our SV-RCNet with PKI strategy can further peak the AC to 92.4%, PR to 90.6% and RE to 86.2%.

The exceeding performances on both M2CAI Workflow Challenge dataset and Cholec80 dataset verify the extensibility of our SV-RCNet architecture. More importantly, compared with the method in [11], the accuracy improving scopes of SV-RCNet also increase as the dataset being larger (from 79.5% to 81.7% on M2CAI Workflow Challenge and from 81.7% to 85.3% on Cholec80). The underlying reason of this phenomenon is that, on the one hand, larger dataset would enrich the training database to gain more powerful spatio-temporal features; on the other hand, larger dataset brings in more complex conditions and challenging issues, where effectively utilizing the complementary information of visual and temporal features is quite important to accurately recognize such challenging cases. The experimental results on the Cholec80 dataset have demonstrated that our proposed framework can effectively address phase recognition task in surgical videos and is general enough to be adapted to various surgical videos.

[3]http://camma.u-strasbg.fr/datasets

TABLE VI
PHASE RECOGNITION RESULTS OF DIFFERENT METHODS ON CHOLEC80 DATASET

Methods	Precision (%)	Recall (%)	Accuracy (%)
SV-RCNet+PKI	**90.6 ± 8.1**	**86.2 ± 15.3**	**92.4 ± 5.2**
SV-RCNet	80.7 ± 7.0	83.5 ± 7.5	85.3 ± 7.3
EndoNet [11]	73.7 ± 16.1	79.6 ± 7.9	81.7 ± 4.2
PhaseNet [11]	71.3 ± 15.6	76.6 ± 16.6	78.8 ± 4.7

IV. DISCUSSION

Automatically recognizing surgical workflow from videos plays a key role in the development of intelligent context-aware operating rooms. While it is a platitude that both visual and temporal information should be harnessed to carry out this task, how to obtain high quality visual features and temporal dependencies and take full advantage of their complementary information do matter to an effective and robust method and are still open problems in this field. Deep learning techniques, extracting the high-level features, have successfully addressed many tasks [42]–[44]. In this paper, we introduce two state-of-the-art deep learning techniques for this task. We propose to employ a very deep ResNet to extract discriminative visual features from frames and exploit a LSTM network to learn the temporal dependencies among frames to model the motions and capture the transition frames. Experiments have sufficiently demonstrated effectiveness of these two techniques.

More importantly, we seamlessly integrate the ResNet and the LSTM network together to form the proposed SV-RCNet and train it in an end-to-end manner. We propose a novel joint loss function to guide the training process so that the visual representations and sequential dynamics can be jointly optimized in the whole training process. The generated spatio-temporal features are, in general, more discriminative than the features produced by traditional CNNs which only take visual information into account. Moreover, it is worth emphasizing that comparing with separately trained CNN-LSTM architectures [29], [45], [46], our SV-RCNet with utilizing joint learning mechanism enables networks to take full advantage of complementary of visual and temporal information and therefore produces more discriminative spatio-temporal features. Overall, we integrate a deep ResNet and a LSTM to form an end-to-end network with a joint loss function and a set of tailored training schemes to solve the challenging problem of workflow recognition from surgical videos. The performance improvement demonstrates the advantage of end-to-end training in the surgical workflow recognition task. In addition, such an attempt can inspire researchers to develop

Authorized licensed use limited to: Johns Hopkins University. Downloaded on February 18,2022 at 04:18:34 UTC from IEEE Xplore. Restrictions apply.

more powerful architectures for the analysis of surgical videos, as well as other time-series medical signals.

By powerful spatio-temporal modeling capability, the SV-RCNet can achieve highly-consistent and accurate recognition results and in particular, precisely identify phase transition frames, which is not only essential for this task but also important for many computer-assisted procedures and even surgical robotics. For example, in case that we can accurately recognize the transition points between consecutive surgical phases, we can automatically adjust the configurations and parameters of a surgical robot to go into next phase. Based on the high-quality predictions of SV-RCNet and considering that most of surgical videos have well-ordered and structure contents, we develop a simple yet effective inference strategy, i.e. PKI, to manage the recognition process. In principle, the PKI can be considered as a transition monitor, to determine the phase prior according to transition points between consecutive phases detected by SV-RCNet. The phase prior, in turn, can help correct some wrong predictions existing in the intermediate of each phase and hence further improve the recognition performance. It is worthwhile to note that the success of PKI heavily relies on the high consistency and accurate predictions of SV-RCNet. From Fig. 6 and Fig. 7, we can find that it is difficult for traditional CNNs to precisely locate the transition points in a surgical video due to the lack of temporal information, while our SV-RCNet can overcome this shortcoming.

As main concerns of the proposed SV-RCNet, the spatio and temporal depths of network should be carefully designed after comprehensively considering network performance, computational resource, training difficulty and so on. For spatio depth, we find that when integrating a 101-layer ResNet with a LSTM network, the required computational resource is not affordable even by an advanced GPU card. In addition, while the training and testing time of a 101-layer ResNet is around twice as long as that of a 50-layer ResNet, the performance gains are quite limited. In this case, we decide to employ 50-layer ResNet to implement the SV-RCNet, achieving satisfactory results while managing the computational resource and training time in a reasonable range. Similarly, for temporal depth, the computational resource and training time are the main constraints to do the exploration of clip length increase. Multi-GPUs and other distributed architectures can enlarge the memory capacity and address the main limitation about computational memory. Moreover, by leveraging the distributed architectures, the exploration about utilizing convolutional LSTM as temporal model can be implemented. Convolutional LSTM can preserve spatial structures as well as model temporal information [30]. However, it consumes large computational memory compared with the conventional LSTM used in our current SV-RCNet, and would be more difficult to train jointly with a 50-layer deep ResNet. In the future, we may first modify our framework into multi-GPU version and then practically find effective training strategies, but we should carefully considered if the computational setting of distributed architectures is suitable for operating rooms before deployment.

The proposed automatic surgical workflow recognition method has the great significance in clinical practice. It can be used as a powerful tool to facilitate surgeon skill evaluation, improve the efficiency of documenting surgical reports and automatically index the surgical video databases [11]. More importantly, our approach can cope with the recognition task online. The quick processing speed (0.1s per frame) can help to develop live context-aware surgical assistance system, including staff scheduling, surgical process monitoring and so on. Note that such online action recognition systems are highly demanded and will be the key component of the operating rooms in the future [47]. As the proposed SV-RCNet is a data-driven approach and utilizes almost no domain-specific knowledge, it is general for various surgical videos though we take cholecystectomy as an example in our study. In addition, most of other types of surgeries have the prior phase order information as well, such as cataract surgery, pituitary surgery as well as robotic surgery [6], [48], [49]. In this regard, our entire framework, including PKI strategy, can also be extended to recognize the workflow of other types of surgical videos. We believe the proposed SV-RCNet and PKI can find many applications in the field of medical video analysis and inspire more and further investigations on how to effectively analyze medical videos based on deep learning techniques.

V. Conclusion

We present a novel and effective recurrent convolutional network, i.e. SV-RCNet, to automatically recognize workflow from surgical videos. We exploit a deep ResNet and a LSTM network the extract visual features and temporal dependencies. Compared with previous methods based on either hand-crafted engineering or traditional CNNs, the deep ResNet is capable of extracting more discriminative visual features. We integrate the ResNet and the LSTM network into the SV-RCNet and train it in an end-to-end manner so that the visual representations and sequential dynamics can be jointly and effectively optimized in the training process. The generated high quality spatio-temporal features from the SV-RCNet can more accurately recognize the frames of difference phases and, in particular, the transition frames between phases. However, it is a difficult task for the features learned from traditional CNNs due to the lack of temporal information and other separately trained CNN-LSTM architectures because of no implicit interplay between the visual and temporal features. Thanks to the comprehension of natural characteristic of surgical videos, we further propose an inference scheme, i.e. the PKI, which leverages prior knowledge to further improve the recognition performance. Extensive experiments on the dataset of *M2CAI Workflow Challenge* demonstrate the superior performance of our method, surpassing all the other participants by a significant margin. Our approach has also outperformed state-of-the-art methods on the *Cholec80* dataset, further verifying the effectiveness of our surgical video recognition framework. The proposed SV-RCNet is inherently general and can be utilized in other medical video analysis tasks.

References

[1] N. Bricon-Souf and C. R. Newman, "Context awareness in health care: A review," *Int. J. Med. Informat.*, vol. 76, no. 1, pp. 2–12, 2007.

Authorized licensed use limited to: Johns Hopkins University. Downloaded on February 18,2022 at 04:18:34 UTC from IEEE Xplore. Restrictions apply.

[2] O. Dergachyova, D. Bouget, A. Huaulmé, X. Morandi, and P. Jannin, "Automatic data-driven real-time segmentation and recognition of surgical workflow," *Int. J. Comput. Assist. Radiol. Surgery*, vol. 11, no. 6, pp. 1–9, 2016.

[3] L. Dazzi, C. Fassino, R. Saracco, S. Quaglini, and M. Stefanelli, "A patient workflow management system built on guidelines," in *Proc. AMIA Annu. Fall Symp. Amer. Med. Informat. Assoc.*, 1997, p. 146.

[4] B. Bhatia, T. Oates, Y. Xiao, and P. Hu, "Real-time identification of operating room state from video," in *Proc. Assoc. Adv. Artif. Intell.*, vol. 2. 2007, pp. 1761–1766.

[5] G. Forestier, L. Riffaud, and P. Jannin, "Automatic phase prediction from low-level surgical activities," *Int. J. Comput. Assist. Radiol. Surgery*, vol. 10, no. 6, pp. 833–841, 2015.

[6] G. Quellec, M. Lamard, B. Cochener, and G. Cazuguel, "Real-time task recognition in cataract surgery videos using adaptive spatiotemporal polynomials," *IEEE Trans. Med. Imag.*, vol. 34, no. 4, pp. 877–887, Apr. 2015.

[7] N. Padoy, T. Blum, S.-A. Ahmadi, H. Feussner, M.-O. Berger, and N. Navab, "Statistical modeling and recognition of surgical workflow," *Med. Image Anal.*, vol. 16, no. 3, pp. 632–641, 2012.

[8] J. E. Bardram, A. Doryab, R. M. Jensen, P. M. Lange, K. L. Nielsen, and S. T. Petersen, "Phase recognition during surgical procedures using embedded and body-worn sensors," in *Proc. IEEE Int. Conf. Pervasive Comput. Commun. (PerCom)*, Mar. 2011, pp. 45–53.

[9] M. S. Holden *et al.*, "Feasibility of real-time workflow segmentation for tracked needle interventions," *IEEE Trans. Biomed. Eng.*, vol. 61, no. 6, pp. 1720–1728, Jun. 2014.

[10] H. C. Lin, I. Shafran, T. E. Murphy, A. M. Okamura, D. D. Yuh, and G. D. Hager, "Automatic detection and segmentation of robot-assisted surgical motions," in *Proc. Int. Conf. Med. Image Comput.-Assist. Intervent.*, 2005, pp. 802–810.

[11] 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.

[12] T. Blum, H. Feußner, and N. Navab, "Modeling and segmentation of surgical workflow from laparoscopic video," in *Proc. Int. Conf. Med. Image Comput.-Assist. Intervent.*, 2010, pp. 400–407.

[13] F. Lalys, L. Riffaud, D. Bouget, and P. Jannin, "A framework for the recognition of high-level surgical tasks from video images for cataract surgeries," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 4, pp. 966–976, Apr. 2012.

[14] U. Klank, N. Padoy, H. Feussner, and N. Navab, "Automatic feature generation in endoscopic images," *Int. J. Comput. Assist. Radiol. Surgery*, vol. 3, nos. 3–4, pp. 331–339, 2008.

[15] H. Greenspan, B. van Ginneken, and R. M. Summers, "Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1153–1159, May 2016.

[16] D. Shen, G. Wu, and H.-I. Suk, "Deep learning in medical image analysis," *Annu. Rev. Biomed. Eng.*, vol. 19, pp. 221–248, Jun. 2017.

[17] G. Litjens *et al.* (Feb. 2017). "A survey on deep learning in medical image analysis." [Online]. Available: https://arxiv.org/abs/1702.05747

[18] L. Tao, L. Zappella, G. D. Hager, and R. Vidal, "Surgical gesture segmentation and recognition," in *Proc. Int. Conf. Med. Image Comput.-Assist. Intervent.*, 2013, pp. 339–346.

[19] N. Padoy, T. Blum, H. Feussner, M.-O. Berger, and N. Navab, "On-line recognition of surgical activity for monitoring in the operating room," in *Proc. Assoc. Adv. Artif. Intell.*, 2008, pp. 1718–1724.

[20] F. Lalys, L. Riffaud, X. Morandi, and P. Jannin, "Surgical phases detection from microscope videos by combining SVM and HMM," in *Proc. Int. MICCAI Workshop Med. Comput. Vis.*, 2010, pp. 54–62.

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

[22] J. Håstad and M. Goldmann, "On the power of small-depth threshold circuits," *Comput. Complexity*, vol. 1, no. 2, pp. 113–129, 1991.

[23] J. Håstad, *Computational Limitations of Small-Depth Circuits*. Cambridge, MA, USA: MIT Press, 1987.

[24] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1–9.

[25] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Training very deep networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 2377–2385.

[26] L. Yu, H. Chen, Q. Dou, J. Qin, and P.-A. Heng, "Automated melanoma recognition in dermoscopy images via very deep residual networks," *IEEE Trans. Med. Imag.*, vol. 36, no. 4, pp. 994–1004, Apr. 2017.

[27] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1110–1118.

[28] K. Andrej and F.-F. Li, "Deep visual-semantic alignments for generating image descriptions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3128–3137.

[29] H. Chen *et al.*, "Automatic fetal ultrasound standard plane detection using knowledge transferred recurrent neural networks," in *Proc. Int. Conf. Med. Image Comput.-Assist. Intervent.*, 2015, pp. 507–514.

[30] T. Zeng, B. Wu, J. Zhou, I. Davidson, and S. Ji, "Recurrent encoder-decoder networks for time-varying dense prediction," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2017, pp. 1165–1170.

[31] A. Graves. (2013). "Generating sequences with recurrent neural networks." [Online]. Available: https://arxiv.org/abs/1308.0850

[32] J. Donahue *et al.*, "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 2625–2634.

[33] C. Olah, "Understanding LSTM networks," in *Proc. GitHub Blog*, Aug. 2015. [Online]. Available: http://colah.github.io/posts/2015-08-Understanding-LSTMs/

[34] B. Kong, Y. Zhan, M. Shin, T. Denny, and S. Zhang, "Recognizing end-diastole and end-systole frames via deep temporal regression network," in *Proc. Int. Conf. Med. Image Comput.-Assist. Intervent.*, 2016, pp. 264–272.

[35] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 675–678.

[36] R. Cadène, T. Robert, N. Thome, and M. Cord. (2016). "M2CAI workflow challenge: Convolutional neural networks with time smoothing and hidden Markov model for video frames classification." [Online]. Available: https://arxiv.org/abs/1610.05541

[37] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2921–2929.

[38] X. Feng, J. Yang, A. F. Laine, and E. D. Angelini, "Discriminative localization in CNNs for weakly-supervised segmentation of pulmonary nodules," in *Proc. Int. Conf. Med. Image Comput.-Assist. Intervent.*, 2017, pp. 568–576.

[39] M. Sahu, A. Mukhopadhyay, A. Szengel, and S. Zachow. (2016). "Tool and phase recognition using contextual CNN features." [Online]. Available: https://arxiv.org/abs/1610.08854

[40] O. Dergachyova, D. Bouget, A. Huaulmé, X. Morandi, and P. Jannin, "Data-driven surgical workflow detection: Technical report for M2CAI 2016 surgical workflow challenge," *M2CAI Challenge*, to be published. [Online]. Available: http://camma.u-strasbg.fr/m2cai2016/reports/Dergachyova-Workflow.pdf

[41] A. P. Twinanda, D. Mutter, J. Marescaux, M. de Mathelin, and N. Padoy. (2016). "Single- and multi-task architectures for surgical workflow challenge at M2CAI 2016." [Online]. Available: https://arxiv.org/abs/1610.08844

[42] H.-C. Shin *et al.*, "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1285–1298, May 2016.

[43] H. Chen, Y. Zheng, J.-H. Park, P.-A. Heng, and S. K. Zhou, "Iterative multi-domain regularized deep learning for anatomical structure detection and segmentation from ultrasound images," in *Proc. Int. Conf. Med. Image Comput.-Assist. Intervent.*, 2016, pp. 487–495.

[44] Q. Dou *et al.*, "3D deeply supervised network for automated segmentation of volumetric medical images," *Med. Image Anal.*, vol. 41, pp. 40–54, Oct. 2017.

[45] J. Chen, L. Yang, Y. Zhang, M. Alber, and D. Z. Chen, "Combining fully convolutional and recurrent neural networks for 3D biomedical image segmentation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 3036–3044.

[46] H. Chen *et al.*, "Ultrasound standard plane detection using a composite neural network framework," *IEEE Trans. Cybern.*, vol. 47, no. 6, pp. 1576–1586, Jun. 2017.

[47] A. P. Twinanda, E. O. Alkan, A. Gangi, M. de Mathelin, and N. Padoy, "Data-driven spatio-temporal RGBD feature encoding for action recognition in operating rooms," *Int. J. Comput. Assist. Radiol. Surgery*, vol. 10, no. 6, pp. 737–747, 2015.

[48] F. Lalys, L. Riffaud, X. Morandi, and P. Jannin, "Automatic phases recognition in pituitary surgeries by microscope images classification," in *Proc. Int. Conf. Inf. Process. Comput.-Assist. Intervent.*, 2010, pp. 34–44.

[49] R. DiPietro *et al.*, "Recognizing surgical activities with recurrent neural networks," in *Proc. Int. Conf. Med. Image Comput.-Assist. Intervent.*, 2016, pp. 551–558.

Authorized licensed use limited to: Johns Hopkins University. Downloaded on February 18,2022 at 04:18:34 UTC from IEEE Xplore. Restrictions apply.