EndoNet: A Deep Architecture for Recognition Tasks on Laparoscopic Videos

Andru P. Twinanda*, Sherif Shehata, Didier Mutter, Jacques Marescaux, Michel de Mathelin, and Nicolas Padoy

***Abstract*— Surgical workflow recognition has numerous potential medical applications, such as the automatic indexing of surgical video databases and the optimization of real-time operating room scheduling, among others. As a result, surgical phase recognition has been studied in the context of several kinds of surgeries, such as cataract, neurological, and laparoscopic surgeries. In the literature, two types of features are typically used to perform this task: visual features and tool usage signals. However, the used visual features are mostly handcrafted. Furthermore, the tool usage signals are usually collected via a manual annotation process or by using additional equipment. In this paper, we propose a novel method for phase recognition that uses a convolutional neural network (CNN) to automatically learn features from cholecystectomy videos and that relies uniquely on visual information. In previous studies, it has been shown that the tool usage signals can provide valuable information in performing the phase recognition task. Thus, we present a novel CNN architecture, called EndoNet, that is designed to carry out the phase recognition and tool presence detection tasks in a multi-task manner. To the best of our knowledge, this is the first work proposing to use a CNN for multiple recognition tasks on laparoscopic videos. Experimental comparisons to other methods show that EndoNet yields state-of-the-art results for both tasks.**

***Index Terms*— Laparoscopic videos, cholecystectomy, convolutional neural network, tool presence detection, phase recognition.**

I. Introduction

IN THE community of computer-assisted interventions (CAI), recognition of the surgical workflow is an important topic because it offers solutions to numerous demands of the modern operating room (OR) [1]. For instance, such recognition is an essential component to develop context-aware systems that can monitor the surgical processes, optimize OR and staff scheduling, and provide automated assistance to the clinical staff. With the ability to segment surgical workflows, it would also be possible to automate the indexing of surgical video databases, which is currently a time-consuming manual process. Indexed databases allow an effortless navigation for video browsing, which is particularly interesting for training purposes and post-operative review. In the long run, through finer analysis of the video content, such context-aware systems could also be used to alert the clinicians to probable upcoming complications.

Various types of features have been used in the literature to carry out the phase recognition task. For instance, in [2], [3], binary tool usage signals were used to perform phase recognition on cholecystectomy procedures. In more recent studies [4], [5], surgical triplets (consisting of the utilized tool, the anatomical structure, and the surgical action) were used to represent the frame at each time step in a surgery. However, these features are typically obtained through a manual annotation process, which is virtually impossible to perform at test time. Despite existing efforts [6], it is still an open question whether such information can be obtained reliably in an automatic manner.

Another feature type that is typically used to perform the phase recognition task is visual features, such as pixel values and intensity gradients [7], spatio-temporal features [8], and a combination of features (color, texture, and shape) [9]. However, these features are handcrafted, i.e., they are *empirically* designed to capture certain information from the images, leading to the loss of other possibly significant characteristics during the feature extraction process.

In this paper, we present a novel method for phase recognition that overcomes the afore-mentioned limitations.

First, instead of using handcrafted features, we propose to learn inherent visual features from surgical (specifically cholecystectomy) videos to perform phase recognition. We focus on visual features because videos are typically the only source of information that is readily available in the OR. In particular, we propose to learn the features using a convolutional neural network (CNN), because CNNs have dramatically improved the results for various image recognition tasks in recent years, such as image classification [10] and object detection [11]. In addition, it is advantageous to automatically learn the features from laparoscopic videos because of the visual challenges inherent in them, which make it difficult to design suitable features. For example, the camera in laparoscopic procedures is not static, resulting in motion blur and high variability of the observed scenes along the surgery. The lens is also often stained by blood which can blur or completely occlude the scene captured by the laparoscopic camera.

Manuscript received May 13, 2016; revised June 21, 2016 and July 14, 2016; accepted July 14, 2016. Date of publication July 22, 2016; date of current version December 29, 2016. This work was supported by French state funds managed by the ANR within the Investissements d'Avenir program under references ANR-11-LABX-0004 (Labex CAMI), ANR-10-IDEX-0002-02 (IdEx Unistra) and ANR-10-IAHU-02 (IHU Strasbourg). *Asterisk indicates corresponding author.*

*A. P. Twinanda is with ICube, University of Strasbourg, CNRS, IHU Strasbourg, France (e-mail: twinanda@unistra.fr).

S. Shehata, M. de Mathelin, and N. Padoy are with the ICube, University of Strasbourg, CNRS, IHU Strasbourg 67091, France.

D. Mutter and J. Marescaux are with the University Hospital of Strasbourg, IRCAD and IHU Strasbourg, France.

Digital Object Identifier 10.1109/TMI.2016.2593957

Second, based on our and others' promising results of using tool usage signals to perform phase recognition [3], [12], we hypothesize that tool information can be additionally utilized to generate more discriminative features for the phase recognition task. This has also been shown in [7], where the tool usage signals are used to reduce the dimension of the handcrafted visual features through canonical correlation analysis (CCA) in order to obtain more semantically meaningful and discriminative features. To incorporate the tool information, we propose to implement a multi-task framework in the feature learning process. The resulting CNN architecture, that we call EndoNet, is designed to jointly perform the phase recognition and tool presence detection tasks. The latter is the task of automatically determining all types of tools present in an image. In addition to helping EndoNet learn more discriminative features, the tool presence detection task itself is also interesting to perform because it could be exploited for many applications, for instance to automatically index a surgical video database by labeling the tool presence in the videos. Combined with other signals, it could also be used to identify a potential upcoming complication by detecting tools that should not appear in a certain phase. It is important to note that this task does not require tool localization, thus it differs from the usual tool detection task [13]. In addition, the tool presence is solely determined by the visual information from the laparoscopic videos. Thus, it does not result in the same tool information as the one used in [3], which cannot always be obtained from the laparoscopic videos alone. For example, the presence of trocars used in [3] is not always apparent in the laparoscopic videos. Automatic presence detection for such tools would require another source of information, e.g., an external video.

Training CNN architectures requires a substantial capacity of parallel computing and a large amount of labeled data. In the domain of medicine, labeled data is particularly difficult to obtain due to regulatory restrictions and the cost of manual annotation. Girshick et al. [11] recently showed that transfer learning can be used to train a network when labeled data is scarce. Inspired by [11], we perform transfer learning to train the proposed EndoNet architecture.

To validate our method, we build a large dataset of cholecystectomy videos containing 80 videos recorded at the University Hospital of Strasbourg. In addition, to demonstrate that our proposed (i.e., EndoNet) features are generalizable, we carry out additional experiments on the EndoVis 2015 challenge dataset[1] containing seven cholecystectomy videos recorded at the Hospital Klinikum Rechts der Isar in Munich [12]. Through comparisons, we also show that EndoNet outperforms other state-of-the-art methods. Moreover, we also demonstrate that training the network in a multi-task manner results in a better network than training in a single-task manner.

In summary, the contributions of this paper are five-fold: (1) for the first time, CNNs are utilized to extract visual features for recognition tasks on laparoscopic videos, (2) we design a CNN architecture that jointly performs the phase recognition and tool presence detection tasks, (3) we present a wide range of comparisons between our method and other approaches, (4) we show state-of-the-art results for both tasks on cholecystectomy videos using solely visual features, and (5) we demonstrate the feasibility of using EndoNet in addressing several practical CAI applications.

[1] http://grand-challenge.org/site/endovissub-workflow/data/

II. Related Work

A. Tool Presence Detection

The literature addressing the problem of automatic tool *presence* detection in the CAI community is still limited. The approaches typically focus on other tasks, such as tool detection [13], [14], tool pose estimation [15], and tool tracking [16], [17]. In addition, most of the methods are only tested on short sequences, while we carry out the task on the complete procedures.

In recent studies [18], [19], radio frequency identification (RFID)-tagged surgical tools have been proposed for tool detection and tracking. Such an active tracking system can be used to solve the tool presence detection problem, but this system is complex to integrate into the OR. Thus, it is interesting to investigate other features that are already available in the OR, e.g., visual cues from the videos. For instance, in [20], Speidel et al. presented an approach to automatically recognize the types of the tools that appear in laparoscopic images. However, the method consists of many steps, such as tool segmentation and contour processing. In addition, it also requires the 3D models of the tools to perform the tool categorization. In a more recent work [9], Lalys et al. proposed to use an approach based on the Viola-Jones object detection framework to automatically detect the tools in cataract surgeries, such as the knife and Intra Ocular Lens instruments. However, the tool presence detection problem on laparoscopic videos poses other challenges that do not appear in cataract surgeries where the camera is static and the tools are not articulated. In this paper, we propose a more direct approach to perform the tool presence detection task by using only visual features without localization steps.

B. Phase Recognition

The phase recognition task has been addressed in several types of surgeries, ranging from cataract [9], [21], neurological [5], to laparoscopic surgeries [4], [7], [22]. Multiple types of features have also been explored to carry out the task, such as tool usage signals [3], [5], surgical action triplets [4], [23], and visual features [7], [24]. Since we propose to carry out the task relying solely on the visual features, we focus the literature discussion on methods that use the visual features.

In [25], Padoy et al. proposed an online phase recognition method based on Hidden Markov Model (HMM) that combines the tool usage signals and two visual cues from the laparoscopic images. The first and second cues respectively indicate whether the camera is inside the patient's body and whether clips are in the field of view. However, to recognize the phase, this method requires the tool signals which are not always immediately available in the OR. Instead, Blum et al. [7] proposed to use the tool usage signals to perform dimensionality reduction on the visual features using CCA. Once the projection function is obtained, the tool information is not required anymore to estimate the surgical

phase. At test time, the visual features are mapped to the common space and then later used to determine the phase. The method performed well, resulting in an accuracy of 76%. However, it has only been tested on a dataset of 10 videos. In addition, the method is potentially limited by the choice of handcrafted features that are used: horizontal and vertical gradient magnitudes, histograms and the pixel values of the downsampled image.

In a more recent work [9], Lalys et al. presented a framework to recognize high-level surgical tasks for cataract surgeries using a combination of visual information: shape, color, texture, and mixed information. The features also contain the tool presence information which is automatically extracted from the microscopic videos, as mentioned in Subsection II-A By using HMM on top of the features, the method yields 91% accuracy. However, the method was evaluated on cataract surgeries, which are substantially different from cholecystectomy surgeries. Cholecystectomy surgeries are generally longer than cataract surgeries. In addition, cholecystectomy videos have visual challenges that are not present in cataract surgeries, such as rapid camera motions, the presence of smoke, and the presence of more articulated tools. In [26], Lea et al. used skip-chain conditional random field on top of kinematic and image features to segment and recognize fine-grained surgical activities, such as needle insertion and tying knot. However, the method is tested on a dataset that contains short sequences (around two minutes). Furthermore, the visual features that are utilized in the afore-mentioned methods are handcrafted.

In [27], Klank et al. proposed to learn automatically the visual features from cholecystectomy videos to carry out the phase recognition task. The approach is based on genetic programming that mutates and crosses the features using predefined operators. The method is therefore limited by the set of predefined operators. In addition, the learnt features failed to give better recognition results than the handcrafted features in some cases.

C. Convolutional Neural Networks

In the computer vision community, convolutional neural networks (CNNs) are currently one of the most successful feature learning methods in performing various tasks. For instance, Krizhevsky et al. [10] addressed the image classification problem on the massive ImageNet database [28] by proposing to use a CNN architecture, referred to as AlexNet. They showed that the features learnt by the CNN dramatically improve the classification results compared to the state-of-the-art handcrafted features, e.g., Fisher Vector on SIFT [29]. Furthermore, in [30], it has been shown that the network trained in [10] is so powerful that it can be used as a black-box feature extractor (without any modification) to successfully perform several tasks, including scene classification and domain adaptation.

CNNs are hard to train because they typically contain a high number of unknowns. For instance, the AlexNet architecture contains over 60M parameters. It is essential to have a high computational power and a huge amount of annotated data to train the networks. Recently, Girshick et al. [11] showed that a new network can be learnt despite the scarcity of labeled data by performing transfer learning. They proposed to take a pre-trained CNN model as initialization and fine-tune the model to obtain a new network. It is shown that the fine-tuned network yielded a state-of-the-art performance for object recognition task, despite being fine-tuned on a network trained for image classification.

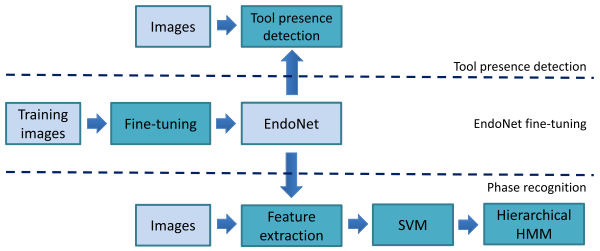


Fig. 1. Full pipeline of the proposed approach.

III. Methodology

The complete pipeline of our proposed approach is shown in Fig. 1. The first step is to train the EndoNet architecture via a fine-tuning process. Once the network is trained, it is used for both the tool presence detection and phase recognition tasks. For the former, the confidence given by the network is directly used to perform the task. For the latter, the network is used to extract the visual features from the images. These features are then passed to the Support Vector Machine (SVM) and Hierarchical HMM to obtain the final estimated phase.

A. EndoNet Architecture

The EndoNet architecture is designed based on two assumptions, which will be confirmed by the experiments presented in Section V:

- more discriminative features for the phase recognition task can be learnt from the dataset if the network is fine-tuned in a multi-task manner, i.e., if the network is optimized to carry out not only phase recognition, but also tool presence detection;
- since the tool signals have been successfully used to carry out phase recognition in previous work [3], [5], [9], the inclusion of automatically generated tool detection signals in the final feature can improve the recognition.

The proposed EndoNet architecture is shown in Fig. 2. The architecture is an extension of the AlexNet architecture [10], which consists of an input layer (in green), five convolutional layers (in red, `conv1-conv5`), and two fully-connected layers (in orange, `fc6-fc7`). The output of layer `fc7` is connected to a fully-connected layer `fc_tool`, which performs the tool presence detection. Since there are seven tools defined in the dataset used to train the network, the layer `fc_tool` contains 7 nodes, where each node represents the confidence for a tool to be present in the image. This confidence is later concatenated with the output of layer `fc7` in layer `fc8` to construct the final feature for the phase recognition. Ultimately, the output of layer `fc8` is connected to layer `fc_phase` containing 7 nodes, where each node represents the confidence that an image belongs to the corresponding phase. The surgical tool types and the surgical phases are described in Subsection IV-A.

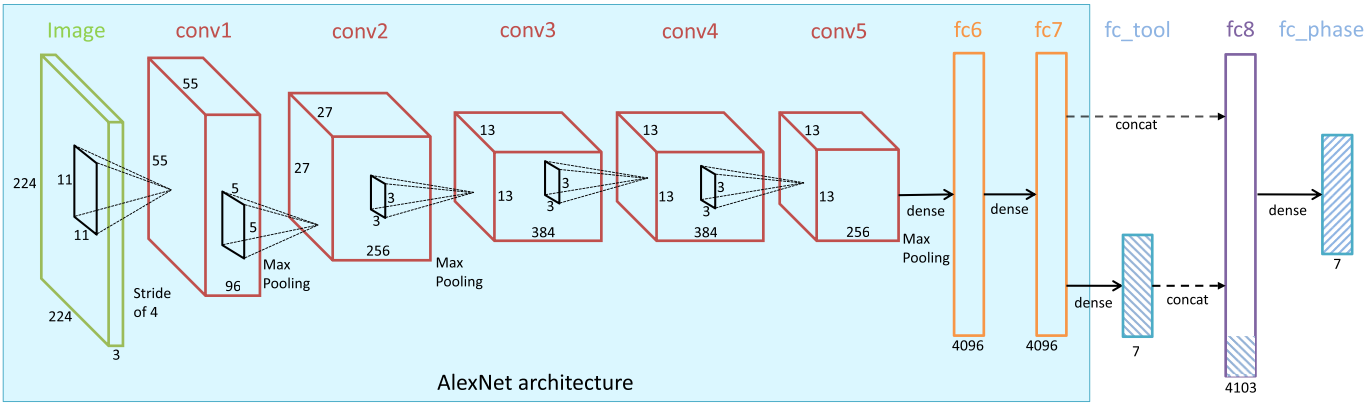


Fig. 2. EndoNet architecture (best seen in color). The layers shown in the turquoise rectangle are the same as in the AlexNet architecture.

B. Fine-Tuning

The network is trained using stochastic gradient descent with two loss functions defined for the tasks. The tool presence detection task is formulated as N_t binary classification tasks, where $N_t = 7$ is the number of tools. For each binary classification task, the cross-entropy function is used to compute the loss. Thus for N_i images in the batch, the complete loss function of the tool presence detection task for all tools is defined as:

$$\mathcal{L}_{\mathcal{T}} = \frac{-1}{N_i} \sum_{t=1}^{N_t} \sum_{i=1}^{N_i} \left[k_t^i \log\left(\sigma\left(v_t^i\right)\right) + \left(1 - k_t^i\right) \log\left(1 - \sigma\left(v_t^i\right)\right)\right], \quad (1)$$

where $i \in \{1, \ldots, N_i\}$ and $t \in \{1, \ldots, N_t\}$ are respectively the image and tool indices, $k_t^i \in \{0, 1\}$ and v_t^i are respectively the ground truth of tool presence and the output of layer `fc_tool` corresponding to tool t and image i, and $\sigma(\cdot) \in (0, 1)$ is the sigmoid function.

Phase recognition is regarded as a multi-class classification task. The softmax multinomial logistic function, which is an extension of the cross-entropy function, is utilized to compute the loss. The function is formulated as:

$$\mathcal{L}_{\mathcal{P}} = \frac{-1}{N_i} \sum_{i=1}^{N_i} \sum_{p=1}^{N_p} l_p^i \log\left(\varphi\left(w_p^i\right)\right), \quad (2)$$

where $p \in \{1, \ldots, N_p\}$ is the phase index and $N_p = 7$ is the number of phases, $l_p^i \in \{0, 1\}$ and w_p^i are respectively the ground truth of the phases and the output of layer `fc_phase` corresponding to phase p and image i, and $\varphi(\cdot) \in [0, 1]$ is the softmax function.

The final loss function is the summation of both losses: $\mathcal{L} = a \cdot \mathcal{L}_{\mathcal{T}} + b \cdot \mathcal{L}_{\mathcal{P}}$, where a and b are weighting coefficients. In this work, we set $a = b = 1$ as preliminary experiments have shown no improvement when varying these parameters. One should note that assigning either $a = 0$ or $b = 0$ is equivalent to designing a CNN that is optimized to carry out only the phase recognition task or the tool presence detection task, respectively.

C. SVM and Hierarchical HMM

The output of layer `fc8` is taken as the image feature. These features are used to compute confidence values $\mathbf{v}_p \in \mathbb{R}^7$ for phase estimation using a one-vs-all multi-class SVM. Since the confidence $\mathbf{v}_p$ is obtained without taking into account any temporal information, it is necessary to enforce the temporal constraint of the surgical workflow. Here, we use use an extension of HMM, namely a two-level Hierarchical HMM (HHMM) [31]. The top-level contains nodes that model the inter-phase dependencies, while the bottom-level nodes model the intra-phase dependencies. We train the HHMM adopting the learning process presented in [31]. Here, the observations are given by the confidence $\mathbf{v}_p$ from the SVM. For offline recognition, the Viterbi algorithm [32] is used to find the most likely path through the HHMM states. As for online recognition, the phase prediction is computed using the forward algorithm.

One can observe that EndoNet already provides confidence values through the output of layer `fc_phase`, thus it is not essential to pass EndoNet features to the SVM to obtain the confidence values $\mathbf{v}_p$. Furthermore, in preliminary experiments, we observed that there was only a slight difference of performance between $\mathbf{v}_p$ and `fc_phase` in recognizing the phases both before and after applying the HHMM. However, this additional step is necessary in order to provide a fair comparison with other features, which are passed to the SVM to obtain the confidence. In addition, using the output of layer `fc_phase` as the phase estimation confidence is only applicable to datasets that share the same phase definition as the one in the fine-tuning dataset. Thus, this step is also required for the evaluation of the network generalizability to other datasets that might have a different phase definition.

IV. Experimental Setup

A. Dataset

We have constructed a large dataset, called *Cholec80,*[2] containing 80 videos of cholecystectomy surgeries performed by 13 surgeons at the University Hospital of Strasbourg. The videos are captured at 25 fps. For faster processing and to

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

Fig. 3. List of the seven surgical tools used in the Cholec80 dataset.

TABLE I
LIST OF PHASES IN THE (a) CHOLEC80 AND (b) EndoVis DATASETS, INCLUDING THE MEAN ± std OF THE DURATION OF EACH PHASE IN SECONDS

ID	Phase	Duration (s)
P1	Preparation	125±95
P2	Calot triangle dissection	954±538
P3	Clipping and cutting	168±152
P4	Gallbladder dissection	857±551
P5	Gallbladder packaging	98±53
P6	Cleaning and coagulation	178±166
P7	Gallbladder retraction	83±56

(a)

ID	Phase	Duration (s)
P0	Placement trocars	180±118
P12	Preparation	419±215
P3	Clipping and cutting	390±194
P4	Gallbladder dissection	563±436
P5	Retrieving gallbladder	391±246
P6	Hemostasis	336±62
P7	Drainage and closing	171±128

(b)

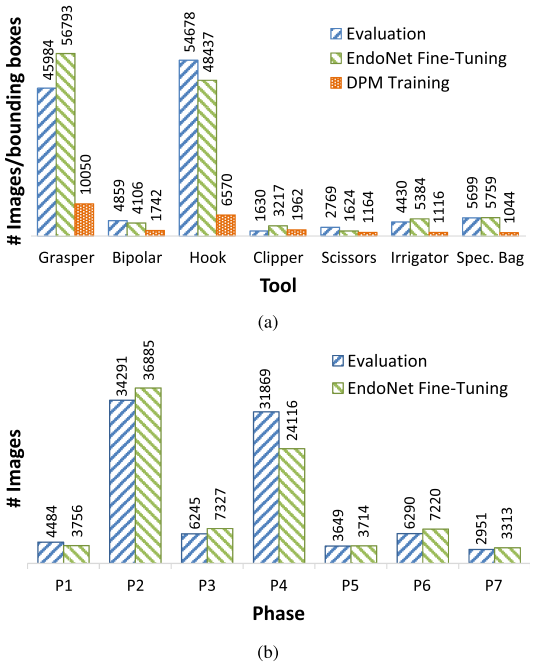


Fig. 4. Distribution of annotations in the Cholec80 dataset for (a) tool presence detection and (b) phase recognition tasks.

reduce redundancy, we downsample the videos to 1 fps by taking the first frame from every 25 frames. In a preliminary study, we also experimented with sequences downsampled to 5 fps, but did not obtain any improvement. The whole dataset is labeled with the phase and tool presence annotations. The phases have been defined by a senior surgeon in our partner hospital. Since the tools are sometimes hardly visible in the images and thus difficult to be recognized visually, we define a tool as present in an image if at least half of the tool tip is visible. The tool and the phase lists can be found in Fig. 3 and Tab. I-a, respectively.

The Cholec80 dataset is split into two subsets of equal size (i.e., 40 videos each). The first subset (i.e., the fine-tuning subset) contains ∼86K annotated images. From this subset, 10 videos have also been fully annotated with the bounding boxes of tools. These are used to train Deformable Part Models (DPM) [33]. Because the grasper and hook appear more often than other tools, their bounding boxes reach a sufficient number from the annotation of three videos. The second subset (i.e., the evaluation subset) is used to test the methods for both tool presence detection and phase recognition. The dataset is split equally in order to provide enough data for both the fine-tuning and evaluation processes. The statistics of the complete dataset can be found in Fig. 4.

The second dataset is a public dataset released for the EndoVis workflow challenge at MICCAI 2015. The dataset contains seven cholecystectomy videos collected at Klinikum Rechts der Isar [12] that are captured at 25 fps and downsampled to 1 fps for processing. Even though the challenge had been cancelled due to insufficient participation and no results have been published, the data is still very useful to demonstrate the generalisation of our approach. We therefore also provide results and comparisons for phase recognition on this dataset (see Section IV-C). We only perform phase detection on this dataset, because the types and the visual appearances of the tools are different from the tools that EndoNet is designed to detect. The list of phases in the EndoVis dataset is shown in Tab. I-b.

It can be seen that phase P3 is longer in Endovis than in Cholec80. This is due to the fact that in Cholec80, P3 is typically started when the calot triangle is clearly exposed. Yet, this is not the case in EndoVis. As a result, extra dissection steps are included in P3, leading to a longer P3 in EndoVis.

Some phases in EndoVis have been defined differently from the phases in Cholec80. For instance, a phase *placement trocars* is defined in the EndoVis dataset, even though it should be noted that this phase is not always visible from the laparoscopic videos. Additional sources of information (e.g., external videos), which are not available in the dataset, are required to label this phase correctly. Another difference is in the definition of the *preparation* phase. In the EndoVis dataset, the *preparation* phase includes the *calot triangle*

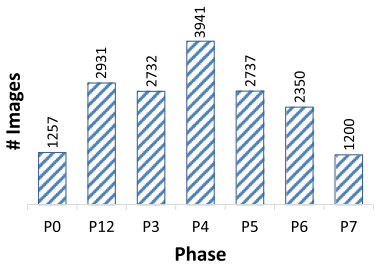


Fig. 5. Phase distribution in the EndoVis dataset.

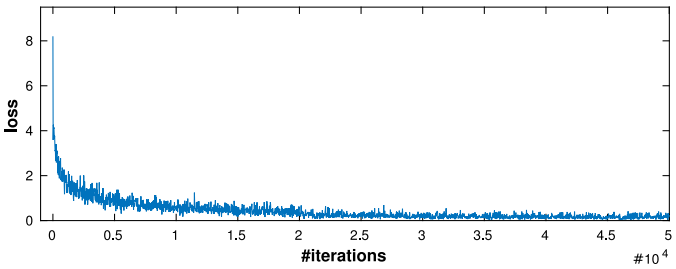


Fig. 6. Evolution of the loss function during the fine-tuning process of EndoNet.

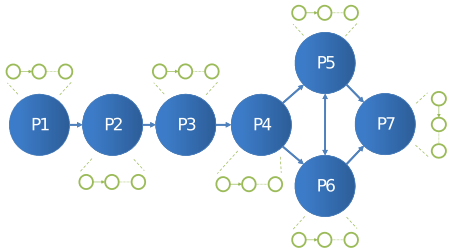
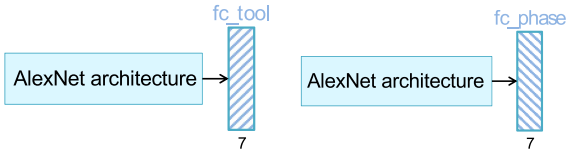


Fig. 7. Graph representation of the two-level HHMM for the surgical phases defined in Cholec80. The top-level states, representing the phases defined in the dataset, are shown in blue. The transitions for top-level states show all possible phase transitions defined in the dataset. The bottom-level states are shown in green.



(a) ToolNet (b) PhaseNet

Fig. 8. Single-task CNN architectures for the (a) tool presence detection and (b) phase recognition tasks. The AlexNet architecture is the same as the one used in EndoNet (see Fig. 2). The single-task networks are also trained via transfer learning.

dissection phase (hence the ID *P12* in Tab. I-b). The other phases are defined similarly to the phases in Cholec80. The distribution of the phases in EndoVis is shown in Fig. 5.

B. Fine-Tuning, SVM and HHMM Parameters

EndoNet is trained by fine-tuning the publicly available AlexNet network [10] which has been pre-trained on the ImageNet dataset [28]. The layers that are not defined in AlexNet (i.e., `fc_tool` and `fc_phase`) are initialized randomly. The network is fine-tuned for 50K iterations with $N_i = 50$ images in a batch. The learning rate is initialized at 10^{-3} for all layers, except for `fc_tool` and `fc_phase`, whose learning rate is set higher at 10^{-2} because of their random initialization. The learning rates for all layers decrease by a factor of 10 for every 20 K iterations. The fine-tuning process is carried out using the Caffe framework [34]. The evolution of the loss function $\mathcal{L}$ during the fine-tuning process is shown in Fig. 6. The graph shows the convergence of the loss, indicating that the network is successfully optimized to learn the optimal features for the phase recognition and tool presence detection tasks.

The networks are trained using an NVIDIA GeForce Titan X graphics card. The training process takes ∼80 seconds for 100 iterations, i.e., roughly 11 hours per network. The feature extraction process takes approximately 0.2 second per image. The computational time for SVM training depends on the size of the features, ranging from 0.1 to 90 seconds, while the HHMM training takes approximately 15 seconds using our MATLAB implementation.

To carry out phase recognition, all features are passed to a one-vs-all *linear* SVM, except the handcrafted features, which are passed through a histogram intersection kernel beforehand. We tried to use non-linear kernels for other features in our preliminary experiments, but this did not yield any improvements.

For the HHMM, we set the number of top-level states to seven (equal to N_p), while the number of bottom-level states is data-driven (as in [31]). To model the output of the SVM, we use a mixture of five Gaussians for every feature, except for the binary tool signal, where one Gaussian is used. The type of covariance is diagonal. In Fig. 7, the graph representation of the HHMM used to recognize the phases in Cholec80 is shown.

C. Baselines

For tool presence detection, we compare the results given by EndoNet (i.e., the output of layer `fc_tool`) with two other methods. The first method is DPM [33], since it is an ubiquitous method for object detection that is available online. In the experiments, we use the default parameters, model each tool using three components and represent the images using HOG features. The second method is a network trained in a single-task manner that solely performs the tool presence detection task (ToolNet). ToolNet is fine-tuned using the same parameters as the ones mentioned in Subsection IV-B. We compare the ToolNet results with the EndoNet results in order to show that performing the fine-tuning process in a multi-task manner yields a better network than in a single-task manner. The architecture of this network can be seen in Fig. 8-a.

For phase recognition, we compare the phase recognition results using the following features as input:

- binary tool information generated from the manual annotation; this is a vector depicting the presence of the tools in an image, i.e. $\mathbf{v}_t \in \{0,1\}^7$ and $\mathbf{v}_t \in \{0,1\}^{10}$ for the Cholec80 and EndoVis datasets, respectively;
- handcrafted visual features: bag-of-word of SIFT, HOG, RGB and HSV histograms; these features are chosen

because they have been successful in carrying out classification [35] on laparoscopic videos;
- the afore-mentioned handcrafted visual features + CCA, similar to the approach suggested in [7];
- the output of layer `fc7` of AlexNet trained on the ImageNet dataset (i.e., the initialization of the fine-tuning process); this is an interesting feature to compare to, because it has been shown that this network can be used for various computer vision tasks [30];
- the output of layer `fc7` from a network that is fine-tuned to carry out phase recognition in a single-task manner, shown in Fig. 8-b (PhaseNet); PhaseNet is also fine-tuned using the parameters mentioned in Subsection IV-B;
- our proposed features, i.e., the output of layer `fc8` from EndoNet.

We also include features called EndoNet-GTbin for phase recognition on the Cholec80 dataset. These features consist of the output of layer `fc7` from EndoNet concatenated with binary tool information obtained from the ground-truth annotations. This evaluation allows us to investigate whether the tool information automatically extracted from EndoNet, which is included in our proposed features, is sufficient for the phase recognition task.

D. Evaluation

As previously mentioned, tool presence detection and phase recognition are two different classification tasks. The former is a binary classification task, while the latter is a multi-class one. Therefore, different evaluation metrics are used. The performance of the tool presence detection is measured by the average precision (AP) metric. It is obtained by computing the area under the precision-recall curve.

For phase recognition, several evaluation metrics are used, i.e., precision, recall, and accuracy as defined in [3]. Precision and recall show the quality of the recognition results for each phase. They compute the number of correct detections divided by the length of the complete detections and by the length of the ground truth, respectively. In contrast, accuracy represents the percentage of correct detections in the complete surgery. Because some phases are short, it is informative to observe the precision and recall per phase in addition to the overall accuracy. All these metrics are then averaged over the surgeries.

For tool presence detection, the learnt models (i.e., DPM, ToolNet, and EndoNet) can be directly used to perform the task without any additional training step. In contrast, the phase recognition task requires an additional step to train the SVM and HHMM after the networks are trained. Because of this reason, the tool presence detection task is performed directly on the evaluation subset of Cholec80; while for phase recognition, we run a 4-fold cross-validation on the evaluation subset of Cholec80 and a full cross-validation on the EndoVis dataset. This additional training step for phase recognition contains random initializations, which can result in different outcomes for each experimental run. Thus, for each validation fold, we perform five experimental runs and average the evaluation metrics over all experimental runs (shown in Table III, Table IV, and Table V).

TABLE II
AVERAGE PRECISION (AP) FOR ALL TOOLS, COMPUTED ON THE 40 VIDEOS FORMING THE EVALUATION DATASET OF CHOLEC80. THE BEST AP FOR EACH TOOL IS WRITTEN IN BOLD

Tool	DPM	ToolNet	EndoNet
Bipolar	60.6	85.9	**86.9**
Clipper	68.4	79.8	**80.1**
Grasper	82.3	84.7	**84.8**
Hook	93.4	95.5	**95.6**
Irrigator	40.5	73.0	**74.4**
Scissors	23.4	**60.9**	58.6
Specimen bag	40.0	86.3	**86.8**
MEAN	58.4	80.9	**81.0**

In order to show the improvements that the proposed features yield, we compute the evaluation metrics for phase recognition on the results before and after applying HHMM. To provide a deeper analysis of the results, we also present in Section VI the performance of EndoNet on two medical applications.

V. EXPERIMENTAL RESULTS

A. Cholec80 Dataset

1) Tool Presence Detection: The results of the tool presence detection task are shown in Tab. II It can be seen that the networks yield significantly better results than DPM. It might be due to the fact that the number of images used for fine-tuning the networks is higher than the number of bounding boxes used for DPM training, but this may only partly explain this large difference. To provide a fairer comparison, we compare the performance of DPM with ToolNet and EndoNet models that are trained only with the 10 videos used to train DPM (see also Section V-A3 and Fig. 11-a for the influence of the fine-tuning subset size). As expected, the performance of the networks is lower compared to the networks trained on the full fine-tuning subset. However, the mean APs are still better than the one of DPM: 65.9 and 62.0 for ToolNet and EndoNet, respectively. Note that, the networks are only trained using binary annotations (present vs. not-present), while DPM uses bounding boxes containing specific localization information. Furthermore, the networks contain a much higher number of unknowns to optimize than DPM. In spite of these facts, with the same amount of training data, the networks perform the task better than DPM.

From Tab. II, it can be seen that EndoNet gives the best results for this task. This shows that training the network in a multi-task manner does not compromise the EndoNet's performance in detecting the tool presence. For all methods, there is a decrease in performance for scissors detection. This might be due to the fact that this tool has the smallest amount of training data (see Fig. 4-a), as it only appears shortly in the surgeries. In addition, it could be confused with the grasper since they share many visual similarities. Over the seven tools and 40 complete surgeries in the evaluation subset of Cholec80, EndoNet obtains 81% mean AP for tool presence detection. The success of this network suggests that binary annotations are sufficient to train a model for this task. This is particularly interesting, since tagging the images

TABLE III

PHASE RECOGNITION RESULTS BEFORE APPLYING THE HHMM (MEAN ± std) ON: (A) CHOLEC80 AND (B) ENDOVIS THE RESULTS FROM OUR PROPOSED FEATURES (EndoNet) ARE WRITTEN IN BOLD. THE BEST RESULT FOR EACH EVALUATION METRIC IS WRITTEN IN ITALIC

Feature	Cholec80		
	Avg. Precision	Avg. Recall	Accuracy
Tool binary	42.8±33.9	41.1±32.3	48.2±2.7
Handcrafted	22.7±28.8	17.9±28.9	44.0±1.8
Handcrafted+CCA	21.9±14.1	18.7±23.3	39.0±0.6
AlexNet	50.4±12.0	44.0±22.5	59.2±2.4
PhaseNet	67.0±9.3	63.4±11.8	73.0±1.6
EndoNet	**70.0±8.4**	**66.0±12.0**	**75.2±0.9**
EndoNet+GTBin	*70.1±9.1*	*66.7±11.1*	*75.3±1.1*

(a)

EndoVis		
Avg. Precision	Avg. Recall	Accuracy
44.3±32.5	48.5±39.3	49.0±9.7
35.7±6.6	33.2±10.5	36.1±2.6
31.1±4.6	31.6±22.6	32.6±5.3
60.2±8.0	57.8±9.3	56.9±4.1
63.5±5.7	63.2±9.3	62.6±4.9
64.8±7.3	***64.3±11.8***	***65.9±4.7***

(b)

TABLE IV

PHASE RECOGNITION RESULTS AFTER APPLYING THE HHMM (MEAN ± std) ON: (a) CHOLEC80 AND (b) EndoVis. THE RESULTS FROM OUR PROPOSED FEATURES (EndoNet) ARE WRITTEN IN BOLD. THE BEST RESULT FOR EACH EVALUATION METRIC IS WRITTEN IN ITALIC

Feature	Overall-Offline (%)			Overall-Online (%)		
	Avg. Precision	Avg. Recall	Accuracy	Avg. Precision	Avg. Recall	Accuracy
Binary tool	68.4±24.1	75.7±13.6	69.2±8.0	54.5±32.3	60.2±23.8	47.5±2.6
Handcrafted	40.3±20.4	40.0±17.8	36.7±7.8	31.7±20.2	38.4±19.2	32.6±6.4
Handcrafted+CCA	54.6±23.8	57.2±21.2	61.3±8.3	39.4±31.0	41.5±21.6	38.2±5.1
AlexNet	70.9±12.0	73.3±16.7	76.2±6.3	60.3±21.2	65.9±16.0	67.2±5.3
PhaseNet	82.5±9.8	86.6±4.5	89.1±5.4	71.3±15.6	76.6±16.6	78.8±4.7
EndoNet	**84.8±9.1**	**88.3±5.5**	**92.0±1.4**	**73.7±16.1**	**79.6±7.9**	**81.7±4.2**
EndoNet-GTbin	*85.7±9.1*	*89.1±5.0*	*92.2±3.5*	*75.1±15.6*	*80.0±6.7*	*81.9±4.4*

(a)

Feature	Overall-Offline (%)			Overall-Online (%)		
	Avg. Precision	Avg. Recall	Accuracy	Avg. Precision	Avg. Recall	Accuracy
Binary tool	81.4±16.1	79.5±12.3	73.0±21.5	80.3±18.1	77.5±18.8	69.8±21.7
Handcrafted	49.7±15.6	33.2±21.5	46.5±24.6	46.6±16.2	48.0±18.5	43.4±21.6
Handcrafted+CCA	66.1±22.3	64.7±22.1	61.1±17.3	52.3±22.2	49.4±21.5	44.0±22.3
AlexNet	85.7±13.2	80.8±10.4	79.5±11.0	78.4±14.1	73.9±11.4	70.6±12.3
PhaseNet	86.8±14.2	83.1±10.6	79.7±12.2	79.1±15.0	75.7±15.3	71.0±9.2
EndoNet	***91.0±7.7***	***87.4±10.3***	***86.0±6.3***	***83.0±12.5***	***79.2±17.5***	***76.3±5.1***

(b)

with binary information of tool presence is much easier than providing bounding boxes. It also shows that the networks can successfully detect tool presence without any explicit localization pre-processing steps (such as segmentation and ROI selection).

2) Phase Recognition: In Tab. III-a, the results of phase recognition on Cholec80 before applying HHMM are shown. These are the results after passing the image features to the SVM. The results show that the CNNs are powerful tools to extract visual features: despite being trained on a completely unrelated dataset, the AlexNet features outperform the handcrafted visual features (without and with CCA) and the binary tool annotation. Furthermore, the fine-tuning step significantly improves the results: the PhaseNet features yield improvements for all metrics compared to the AlexNet features. In addition to yielding the tool presence detection as a by-product, the multi-task framework applied in EndoNet further improves the features for the phase recognition task. It is also interesting to observe that the phase recognition results using the EndoNet-GTbin features are only slightly better than the ones using the EndoNet features, with approximately 0.1% improvement in accuracy. In other words, the tool information generated from the ground-truth does not bring more information than the EndoNet features and the visual features extracted by EndoNet alone are sufficient to carry out the phase recognition task.

In Tab IV-a, the phase recognition results after applying HHMM are shown. Due to the nature of offline phase recognition, where the algorithm can see the complete video, the offline results are better than the online counterparts. However, when we compare the feature performance, the trend is consistent across the offline and online modes. By comparing the results from Tab. III-a and Tab IV-a, we can see the improvement that the HHMM brings, which is consistent across all features.

In Fig. 9, we show confusion matrices to visualize how well the EndoNet features distinguish the phases from one another. The confusion matrices are generated using the results from one (randomly chosen) experimental run, both before and after HHMM (offline mode). It can be seen that before applying the temporal model, some images are misclassified by the SVM. This is to be expected since the SVM only relies on per-frame visual features and most laparoscopic images look similar to each other. Once the temporal model is incorporated, the recognition results are significantly improved. It can be seen that the phases are typically confused with the neighboring phases. However, the misclassification rates are significantly lower, except for the last few phases, which is due to the non-sequential transitions among these phases (see Fig. 7). It is important to note that confusion matrices for tool presence

TABLE V
PRECISION AND RECALL OF PHASE RECOGNITION FOR EACH PHASE ON CHOLEC80 USING THE EndoNet FEATURES

Feature	Metric	P1	P2	P3	P4	P5	P6	P7
EndoNet - offline	Prec.	83.5±9.6	97.1±2.0	81.0±7.7	97.3±2.1	73.1±8.0	79.7±10.4	81.9±11.8
	Rec.	90.9±5.7	80.8±4.3	88.1±7.4	94.7±1.0	83.7±5.6	79.6±8.8	86.7±11.8
EndoNet - online	Prec.	90.0±5.6	96.4±2.0	69.8±10.7	82.8±6.2	55.5±11.9	63.9±10.5	57.5±11.0
	Rec.	85.5±3.9	81.1±8.9	71.2±9.7	86.5±4.3	75.5±3.8	68.7±9.1	88.9±7.5

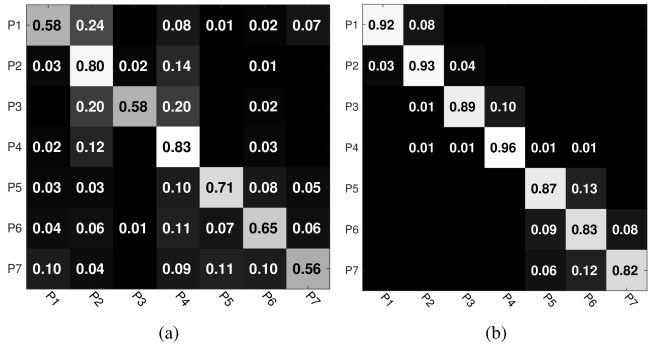


Fig. 9. Confusion matrices for phase recognition on Cholec80: (a) before HHMM and (b) after HHMM (offline mode).

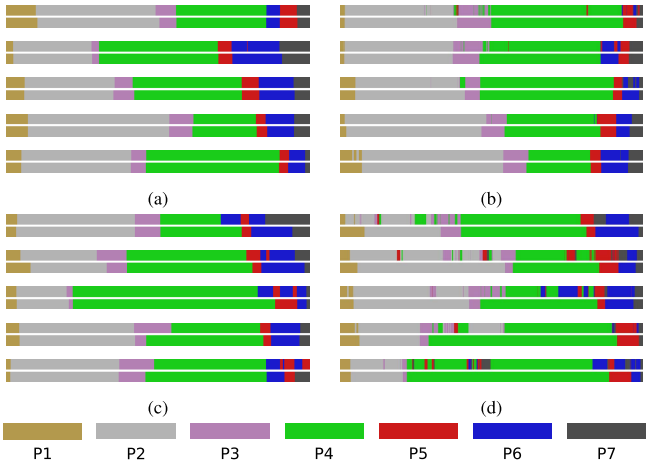


Fig. 10. Phase recognition results vs. ground truth on the evaluation subset of Cholec80 in a color-coded ribbon illustration. The horizontal axis of the ribbon represents the time progression in a surgery. The top ribbon is the estimated phase and the bottom ribbon is the ground truth. (a) Top-5 offline. (b) Top-5 online. (c) Bottom-5 offline. (d) Bottom-5 online.

detection cannot be generated since we perform *tool presence* detection only. Contrary to tool localization approaches, our approach cannot determine if a certain tool is being confused with another tool.

In Fig. 10, we visualize the confusions temporally on top-5 and bottom-5 recognition results obtained from the same experimental run used to generate Fig. 9. In offline mode, it can be seen that the top-5 results are very good, resulting in over 98% accuracies. In addition, the bottom-5 results in offline mode are comparable to the ground truth. The drop of accuracy for the bottom-5 are caused by the jumps that can happen between P5 and P6, which are shown by the alternating blue and red in Fig. 10-c. These jumps occur also because of the non-sequential transitions among these phases (see Fig. 7).

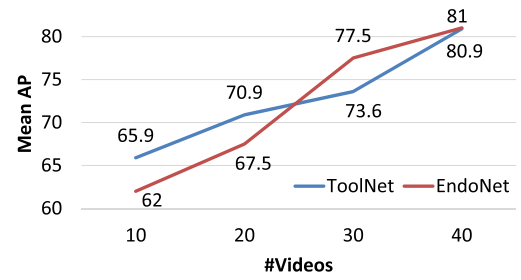
In online mode, one can observe more frequent jumps in the phase estimations. This is due to the nature of recognition in online mode, where future data is unavailable, so that the model is allowed to correct itself after making an estimation. Despite these jumps, the top-5 online results are still very close to the ground-truth, resulting in accuracies above 92%.

In order to provide more comprehensive information regarding the performance of EndoNet over the whole dataset, we present the recognition results for all phases in both offline and online modes in Tab. V. It can be seen that the EndoNet features perform very well in recognizing all the phases. A decrease in performance can be observed for the recognition of P5 and P6. This is likely due to the fact that the transitions between these phases are not sequential and that there is not always a clear boundary between them, especially as some images sometimes do not show any activity. This creates some ambiguity in the phase estimation process.

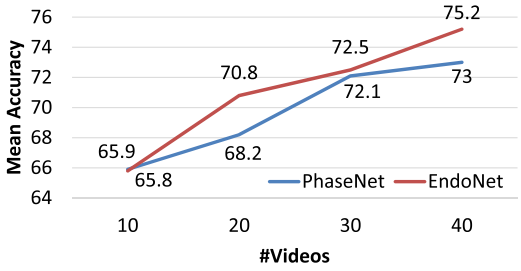
3) Effect of Fine-Tuning Subset Size: In order to show the importance of the amount of training data for the fine-tuning process, we fine-tune our networks using fine-tuning subsets with gradually increasing size: 10, 20, 30, and ultimately 40 videos. We perform both tool presence detection and phase recognition tasks on the evaluation subset of Cholec80 using the trained networks. The results are shown in Fig. 11. As expected, the performance of the networks increase proportionally to the amount of data in the fine-tuning subset. It can also be seen that EndoNet performs better than the single-task networks (i.e., PhaseNet and ToolNet), except for the tool presence detection task where fewer videos are used to train the networks. This indicates that EndoNet takes more advantage of the big dataset compared to ToolNet.

B. EndoVis Dataset

Similar results for phase recognition are obtained from the EndoVis dataset, as shown in Tab. III and Tab. IV-b. It is interesting to see that there is quite a difference in the performance of the binary tool features between the two datasets. After HHMM, the feature is shown to perform better on EndoVis than on Cholec80. However, if we compare the results before HHMM (in Table III), the feature performs similarly on both datasets (48.2% vs. 49% accuracy). This suggests that the improvement comes from the temporal model. Note that the binary tool features from EndoVis consist of 10 binary signals, while the ones from Cholec80 consist



(a)



(b)

Fig. 11. Evolution of network performance on Cholec80 with respect to the number of videos in the fine-tuning subset. (a) Tool presence detection. (b) Phase recognition.

of 7 binary signals. Thus, it is possible that the confidence values given by the SVM for EndoVis are better (i.e., easier to smooth temporally by the temporal models) since the binary tool features from EndoVis provide more information than the ones from Cholec80.

In terms of learnt features, it can be observed that the improvements obtained by PhaseNet and EndoNet on EndoVis are not as high as the result improvements on Cholec80, which is expected since these networks are fine-tuned using the videos from Cholec80. In spite of this fact, the results on the EndoVis dataset also show that the EndoNet features improve the phase recognition results significantly. It indicates that the multi-task learning results in a better network than the single-task counterpart. The fact that the features from EndoNet yield the best results for all cases also shows that EndoNet is generalizable to other datasets.

One should note that we use the output of layer `fc8` from EndoNet as the image feature, which includes confidence values for tool presence. Because the tools used in EndoVis dataset are not the same tools as the ones in the Cholec80 dataset (which is used to train EndoNet), these confidence values can simply be regarded as 7 additional scalar features appended to the feature vector. The results show that these values help to construct more discriminative features.

VI. Medical Applications

In Section V, we have discussed the phase recognition results in both offline and online modes. These results already demonstrate the feasibility of using EndoNet for various medical tasks, such as context-aware support (online recognition) and reporting (offline recognition). Here, we further demonstrate the applicability of EndoNet for other practical CAI applications.[3]

TABLE VI
Number of Phases That Are Correctly Identified in Offline Mode Within the Defined Tolerance Values in the 40 Evaluation Videos of Cholec80.

Tolerance (s)	Phase						
	P1	P2	P3	P4	P5	P6	P7
<30	40	34	34	34	40	30	33
30-59	0	0	0	0	0	0	0
60-89	0	4	1	0	0	1	3
90-119	0	0	1	2	0	0	2
≥120	0	2	4	4	0	4	2
TOTAL	40	40	40	40	40	35	40

We present the results from the same experimental run that is used to generate Fig. 10. First, to show the feasibility of using EndoNet as the basis for automatic surgical video indexing, we show the error of the phase estimation in seconds to indicate how precise the phase boundary estimations from EndoNet are. Second, we investigate further how accurately EndoNet detects the presence of two tools: clipper and bipolar. These tools are particularly interesting because: (1) the appearance of the clipper typically marks the beginning of the *clipping and cutting* phase, which is the most delicate phase in the procedure, and (2) the bipolar tool is generally used to stop haemorrhaging, which could lead to possible upcoming complications.

A. Automatic Surgical Video Database Indexing

For automatic video indexing, the task corresponds to carrying out phase recognition in offline mode. From the results shown in Fig. 10-a, c, one can already roughly interpret how accurate the phase recognition results are. To give a more intuitive evaluation, we present the number of phase boundaries that are detected within defined temporal tolerance values in Tab. VI. One might notice that the number of P6 occurrences is not 40 since not all surgeries go through the cleaning and coagulation phase. We can see that EndoNet generally performs very well for all the phases, resulting in 89% of the phase boundaries being detected within 30 seconds. It can also be seen that only 6% of the phase boundaries are detected with an error over 2 minutes. It is also important to note that this error is computed with respect to the strict phase boundaries defined in the annotation. In practice, these boundaries are not as harsh or visually obvious. Thus, this error is acceptable in most cases. In other words, it indicates that the results from EndoNet do not require a lot of corrections, which will make surgical video indexing a lot faster and easier.

It should be noted that similar metrics could be computed for phase recognition in online mode. However, this is an ill-posed problem because there is no clear boundary

[3]This paper has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the authors. This video visualizes the results of tool presence detection and phase recognition (in both offline and online modes). It also presents a medical application, namely surgical video database indexing, which is built on top of the proposed method. The size of the video is 33 MB.

TABLE VII
APPEARANCE BLOCK DETECTION RESULTS FOR BIPOLAR AND CLIPPER, INCLUDING THE NUMBER OF CORRECTLY CLASSIFIED BLOCKS AND MISSED BLOCKS, AND THE FALSE POSITIVE RATE OF THE DETECTION

Tolerance (s)	Bipolar	Clipper
<5	114	49
6-29	9	10
30-59	1	0
≥60	0	1
Missed	0	1
False positives	3.8%	8.3%

for the detected phases obtained from the recognition in online mode since the detections may jump between phases (see Fig. 10-d). In practice, additional post-processing steps, such as smoothing on a temporal window, would be required for this application, which is not in the scope of this work.

B. Bipolar and Clipper Detection

In addition to showing the AP for detection of both tools in Tab. II, we present a more intuitive metric to measure the reliability of EndoNet for the bipolar and clipper presence detections. We define a *tool block* as a set of consecutive frames in which a certain tool is present. Since the tools might not always be visible in an image even though they are currently being used, we merge the blocks (of the same tool) in the ground-truth data that have a gap that is less than 15 seconds. Then, we define a tool block as identified if EndoNet can detect the tool in at least one of the frames inside the block. To show the performance of EndoNet in terms of temporal precision, we also present the time difference between the first frame of the tool block and the first frame of the detection. In this experiment, we determine the tool presence by taking a confidence threshold that gives a high precision for each tool, so that the system can obtain the minimal amount of false positives and retain the sensitivity in correctly detecting the tool blocks. Since the false positive rate is measured using the tool block definition, we also close the gaps between the tool presence detections that are less than 15 seconds.

We show the block detection results in Tab. VII. It can be seen that all the bipolar blocks are detected very well by EndoNet. Over 90% of the blocks are detected under 5 seconds. EndoNet also yields a very low false positive rate (i.e., 3.8%) for the bipolar. This excellent performance is obtained thanks to the distinctive visual appearance that the bipolar has (e.g., the blue shaft). For the clipper, it can be seen that the false positive rate is higher than for the bipolar. This could be due to the fact that it has the second lowest amount of annotations in the dataset, because, similarly to the scissors, the clipper only appears shortly in the surgeries. However, EndoNet still performs very well for clipper detection, showing that 80% and 97% of the blocks are detected under 5 and 30 seconds, respectively.

VII. DISCUSSION AND CONCLUSIONS

In this paper, we address the problem of phase recognition in laparoscopic surgeries and propose a novel method to learn visual features directly from raw images. This method is based on a convolutional neural network (CNN) architecture, called EndoNet, which is designed to perform two tasks simultaneously: tool presence detection and phase recognition. We show through experiments that this approach overcomes the inherent visual challenges in the dataset and subsequently yields visual features that outperform both previously used features and the features obtained from architectures designed for a single task. Interestingly, the EndoNet visual features also perform significantly better in the phase recognition task than binary tool signals indicating which tools can be seen in the image, even though these signals are obtained from ground truth annotations. These results therefore suggest that the images contain additional characteristics useful for recognition in addition to simple tool presence information and that these characteristics are successfully retrieved by EndoNet. Additionally, we have shown that EndoNet also performs well on another smaller dataset, namely EndoVis, and is therefore generalizable.

To train and evaluate EndoNet, we constructed a large dataset containing 80 videos of cholecystectomy procedures performed by 13 surgeons. Even though the cholecystectomy procedure is a common focus for surgical workflow analysis, to the best of our knowledge, the cholecystectomy datasets used in previous work are limited to less than 20 surgeries. This is therefore the first large-scale study performed for these recognition tasks. This is also the first comparison of the features that can be used to perform phase recognition on laparoscopic surgeries. Furthermore, it is shown by the *std* of the phase durations in Tab-I-a that the dataset in itself contains a high variability. The state-of-the-art results from EndoNet indicates that our proposed method can cope with such complexity.

The results of varying sizes of the fine-tuning subset suggest that taking more videos from Cholec80 to fine-tune the networks will lead to better performance. However, it should be noted that the videos in Cholec80 come from one hospital, thus the complexity of the data is limited to the variability of procedure executions by surgeons from the same institution. Training a CNN network with such a dataset can lead to over-fitting and subsequently reduce the generalizability of the network. To obtain more generalizable networks, videos from other medical institutions should be included to ensure a higher variability in the dataset. The success of EndoNet in carrying out the tool presence detection and phase recognition tasks should be considered as a call for action in the community to open their data to accelerate the development of generalizable solutions for these tasks.

We have shown the applicability of EndoNet for two different applications. These applications focus on video database management, which is one of the demands from our clinical partners. In future work, other related applications should be addressed, such as context-aware assistance during live surgeries. It will also be interesting to explore whether the features generated by EndoNet can be used to perform other tasks in laparoscopic videos, such as the estimation of the completion time of the procedure [3], the classification of surgical videos [35], and the recognition of the anatomy.

Despite yielding state-of-the-art results, the presented phase recognition pipeline still has some limitations. For example, the phase recognition still relies on the HHMM, which is required to enforce the temporal constraints in the phase estimation. Thus, the features learnt by EndoNet do not include any temporal information present in the videos. In addition, since the HHMM is trained separately from the EndoNet fine-tuning process, the EndoNet features are not optimized on the entire phase recognition task. With additional training data, these limitations could be solved by using long short term memory (LSTM) architectures. Such an approach will form part of future efforts to improve phase recognition.

ACKNOWLEDGEMENTS

The authors would like to thank the IRCAD audio-visual team for their help in generating the dataset. The authors would also like to acknowledge the support of NVIDIA with the donation of the GPU used in this research.

REFERENCES

[1] K. Cleary, H. Y. Chung, and S. K. Mun, "Or2020 workshop overview: Operating room of the future," *Int. Congr. Ser.*, vol. 1268, pp. 847–852, Jun. 2004.

[2] L. Bouarfa, P. P. Jonker, and J. Dankelman, "Discovery of high-level tasks in the operating room," *J. Biomed. Informat.*, vol. 44, no. 3, pp. 455–462, 2011.

[3] N. Padoy *et al.*, "Statistical modeling and recognition of surgical workflow," *Med. Image Anal.*, vol. 16, no. 3, pp. 632–641, 2012.

[4] D. Katić *et al.*, "Knowledge-driven formalization of laparoscopic surgeries for rule-based intraoperative context-aware assistance," in *Information Processing in Computer-Assisted Interventions* (Lecture Notes in Computer Science), vol. 8498. 2014, pp. 158–167.

[5] G. Forestier *et al.*, "Multi-site study of surgical practice in neurosurgery based on surgical process models," *J. Biomed. Informat.*, vol. 46, no. 5, pp. 822–829, 2013.

[6] F. Lalys, D. Bouget, L. Riffaud, and P. Jannin, "Automatic knowledge-based recognition of low-level tasks in ophthalmological procedures," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 8, no. 1, pp. 39–49, 2012.

[7] T. Blum, H. Feußner, and N. Navab, "Modeling and segmentation of surgical workflow from laparoscopic video," in *Medical Image Computing and Computer-Assisted Intervention* (Lecture Notes in Computer Science), vol. 6363. 2010, pp. 400–407.

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

[9] 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.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[11] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. CVPR*, 2014, pp. 580–587.

[12] R. Stauder *et al.*, "Random forests for phase detection in surgical workflow analysis," in *Information Processing in Computer-Assisted Interventions* (Lecture Notes in Computer Science), vol. 8498. 2014, pp. 148–157.

[13] R. Sznitman, C. Becker, and P. Fua, "Fast part-based classification for instrument detection in minimally invasive surgery," in *Medical Image Computing and Computer-Assisted Intervention* (Lecture Notes in Computer Science), 2014, pp. 692–699.

[14] D. Bouget *et al.*, "Detecting surgical tools by modelling local appearance and global shape," *IEEE Trans. Med. Imag.*, vol. 34, no. 12, pp. 2603–2617, Dec. 2015.

[15] M. Allan *et al.*, "Image based surgical instrument pose estimation with multi-class labelling and optical flow," in *Medical Image Computing and Computer-Assisted Intervention* (Lecture Notes in Computer Science), vol. 9349. 2015, pp. 331–338.

[16] N. Rieke *et al.*, "Surgical tool tracking and pose estimation in retinal microsurgery," in *Medical Image Computing and Computer-Assisted Intervention* (Lecture Notes in Computer Science), vol. 9349. 2015, pp. 266–273.

[17] A. Reiter, P. K. Allen, and T. Zhao, "Feature classification for tracking articulated surgical tools," in *Medical Image Computing and Computer-Assisted Intervention* (Lecture Notes in Computer Science), vol. 7511. 2012, pp. 592–600.

[18] M. Kranzfelder *et al.*, "Real-time instrument detection in minimally invasive surgery using radiofrequency identification technology," *J. Surg. Res.*, vol. 185, no. 2, pp. 704–710, 2013.

[19] T. Neumuth and C. Meißner, "Online recognition of surgical instruments by information fusion," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 7, no. 2, pp. 297–304, 2012.

[20] S. Speidel *et al.*, "Automatic classification of minimally invasive instruments based on endoscopic image sequences," *Proc. SPIE*, vol. 7261, p. 72610A, Mar. 2009.

[21] 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.

[22] B. P. L. Lo, A. Darzi, and G.-Z. Yang, "Episode classification for the analysis of tissue/instrument interaction with multiple visual cues," in *Medical Image Computing and Computer-Assisted Intervention* (Lecture Notes in Computer Science), vol. 2878. 2003, pp. 230–237.

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

[24] C. Lea, J. Facker, G. Hager, R. Taylor, and S. Saria, "3D sensing algorithms towards building an intelligent intensive care unit," in *Proc. AMIA Summits Transl. Sci.*, 2013, p. 136.

[25] 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. 20th IAAI*, 2008, pp. 1718–1724.

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

[27] U. Klank, N. Padoy, H. Feussner, and N. Navab, "Automatic feature generation in endoscopic images," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 3, no. 3, pp. 331–339, 2008.

[28] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[29] J. Sanchez and F. Perronnin, "High-dimensional signature compression for large-scale image classification," in *Proc. CVPR*, Washington, DC, USA, 2011, pp. 1665–1672.

[30] J. Donahue *et al.*, "Decaf: A deep convolutional activation feature for generic visual recognition," in *Proc. ICML*, 2013, pp. 647–655.

[31] N. Padoy, D. Mateus, D. Weinland, M.-O. Berger, and N. Navab, "Workflow monitoring based on 3D motion features," in *Proc. ICCV Workshops*, 2009, pp. 585–592.

[32] A. J. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Trans. Inf. Theory*, vol. 13, no. 2, pp. 260–269, Apr. 1967.

[33] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.

[34] Y. Jia *et al.* (2014). "*Caffe: Convolutional architecture for fast feature embedding*." [Online]. Available: http://arxiv.org/abs/1408.5093

[35] A. P. Twinanda, J. Marescaux, M. De Mathelin, and N. Padoy, "Towards better laparoscopic video database organization by automatic surgery classification," in *Proc. IPCAI*, 2014, pp. 186–195.