Biomedical Signal Processing and Control 68 (2021) 102801

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

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

A deep learning spatial-temporal framework for detecting surgical tools in laparoscopic videos

Tamer Abdulbaki Alshirbaji [a,c,*], Nour Aldeen Jalal [a,c], Paul D. Docherty [a,b], Thomas Neumuth [c], Knut Möller [a]

[a] Institute of Technical Medicine (ITeM), Furtwangen University, Villingen-Schwenningen, Germany
[b] Department of Mechanical Engineering, University of Canterbury, Christchurch, New Zealand
[c] Innovation Center Computer Assisted Surgery (ICCAS), University of Leipzig, Leipzig, Germany

ARTICLE INFO

Keywords:
Surgical tool presence detection
Spatial-temporal information
CNN
LSTM
Endoscopic video

ABSTRACT

Background and objective: Image-based surgical tool presence detection is an indispensable component for developing various intelligent applications in future operating rooms (ORs). To date, tool presence detection in laparoscopic videos has been investigated, and some recent studies tackled it in a spatial-temporal manner. The promising performance demonstrates the value of temporal information to develop robust methods for surgical tool detection. Therefore, a deep learning framework that considers spatial and temporal information for detecting surgical tools in laparoscopic videos is proposed.

Methods: The proposed approach consists of a hierarchical organised neural architecture consisting of a convolutional neural network (CNN) with two long short-term memory (LSTM) models. The CNN model was used to learn spatial features from laparoscopic images. Since the data was sparsely labelled at 1 Hz, an LSTM network (LSTM-clip) -based on the CNN output- was employed to learn temporal dependencies from short intermediate partially labelled video clips. Finally, temporal dependencies along the complete surgical videos were modelled using another LSTM (LSTM-video). The models were trained and validated using six-fold Monte Carlo cross-validation (MCCV).

Results: Six-fold cross-validation experiments on the large publicly available dataset (Cholec80) explicate the advantage of temporal information to the tool detection task by improving the mean average precision (mAP) by 3.00 %. The proposed approach achieved a mAP of 94.74 % that exceeds the state-of-the-art methods.

Conclusion: The overall approach demonstrates the value of modelling temporal dependencies across consecutive laparoscopic images to enhance surgical tool presence detection.

1. Introduction

Recent technological advances have played an effective role in the transformation of surgery by bringing new medical devices and improvements in surgical practice [1–4]. Nevertheless, this evolution increased surgical workflow complexity and has thus triggered active research in computer-assisted intervention (CAI) in the next generation of operating room environments (ORs) [4]. Future ORs will use computer-aids to enable knowledge-based analyses of available data to allow novel decision-support or context-aware systems (CASs). CAS aims to optimise surgical treatment and improve surgical practice by analysing and relating information from different medical disciplines and communicating relevant knowledge to human operators during surgical interventions. Analysis of surgical tool usage is a fundamental goal of CAS as it will allow surgical phase recognition [5–8]. Moreover, automatic recognition of surgical tools may be utilised to generate automatic reports during surgical procedures.

Several approaches for detecting surgical tools have been proposed. These approaches include sensor-based methods like radiofrequency identification (RFID) [9] or endoscopic image-based methods [10–22]. Early image-based approaches focused on extracting *a-priori* defined features such as colour [13,19] or image gradient [15]. With the emergence of deep learning in object detection tasks, most recent approaches utilise convolutional neural networks (CNNs) [10–12,14,16–18,20–22]. Video-based tool detection is highly preferable as endoscopic videos in minimally invasive surgery provide an easily

* Corresponding author at: Institute of Technical Medicine, Furtwangen University, Jakob-Kienzle-Strasse 17, 78054, Villingen-Schwenningen, Germany.
E-mail address: Tamer.Abdulbaki.Alshirbaji@hs-furtwangen.de (T. Abdulbaki Alshirbaji).

https://doi.org/10.1016/j.bspc.2021.102801
Received 11 January 2021; Received in revised form 26 April 2021; Accepted 15 May 2021
Available online 26 May 2021

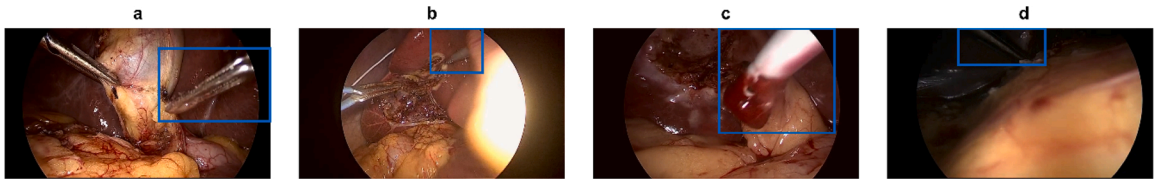

Fig. 1. Examples of the obscured nature of laparoscopic images. (a) partial appearance of the bipolar (the tool in the right side of the image), (b) light reflection affects detection of the irrigator (the tool in the image background), (c) blood covering the irrigator tip. (d) dark background due to tissues near to the laparoscopic light source. The respective tool is marked with a blue rectangle.

recorded source of information. However, video-based tool recognition remains a challenging task. In particular, surgical tool detection is a multi-object detection task with many possible tool usage combinations. Laparoscopic images can be obscured by rapid movement of the laparoscopic camera, smoke, variability of biological tissues, and blood covering tools (see Fig. 1).

To overcome the aforementioned obstacles, several studies took advantage of relationships between surgical tools and surgical phases and adopted CNN architectures to perform both tool and phase recognition in a multi-task manner [21,22]. Twinanda et al. introduced a baseline model, called EndoNet [21]. EndoNet performs phase recognition and surgical tool presence detection in cholecystectomy videos. Similarly, Jin et al. developed a multi-task deep learning framework, but they also defined a new correlation loss to exploit tool-phase relation [22]. Sahu et al. addressed the imbalanced dataset problem, and they resampled the data to generate balanced training set based on tool occurrences [12]. Abdulbaki Alshirbaji et al. employed resampling techniques and weighted loss to counter this problem [14]. Additionally, since the surgical video represents sequential data, recurrent neural networks (RCNNs) have been used to consider the temporal aspect of the surgical procedure and refine the CNN classification [10,11,20]. Mishra et al. proposed the use of a long short-term memory (LSTM) network to learn temporal dependencies across adjacent frames [11]. In a similar fashion, Chen et al. explored using 3D CNN to learn spatiotemporal features from short video clips [17]. Al Hajj et al. applied a CNN-RNN pipeline to detect tool usage in surgical videos. Instead of training both networks in an end-to-end manner, they introduced a boosting strategy that relied on using weak classifiers to supervise the CNN training according to the RNN output [10]. Nwoye et al. proposed a deep learning framework that was trained using tool binary annotations to perform tool presence detection and tool tracking [20]. They employed convolutional LSTM (ConvLSTM) to learn spatiotemporal features in surgical videos. Recently, Wang et al. demonstrated the feasibility of using graph convolutional networks (GCN) to learn temporal relationship across consecutive frames [18]. Their approach involved utilising video sequences of labelled frame and its neighbouring unlabelled frames and achieved a significant improvement compared to reference methods.

Although the promising performance achieved by the various CNN methods, potential to improve the temporal modelling of the surgical tool presence in laparoscopic videos remains. Some of the previously proposed methods have relied on capturing temporal dependencies in short video sequences [10,17,18]. In contrast, other methods proposed modelling temporal information along the surgical video [20]. Since the

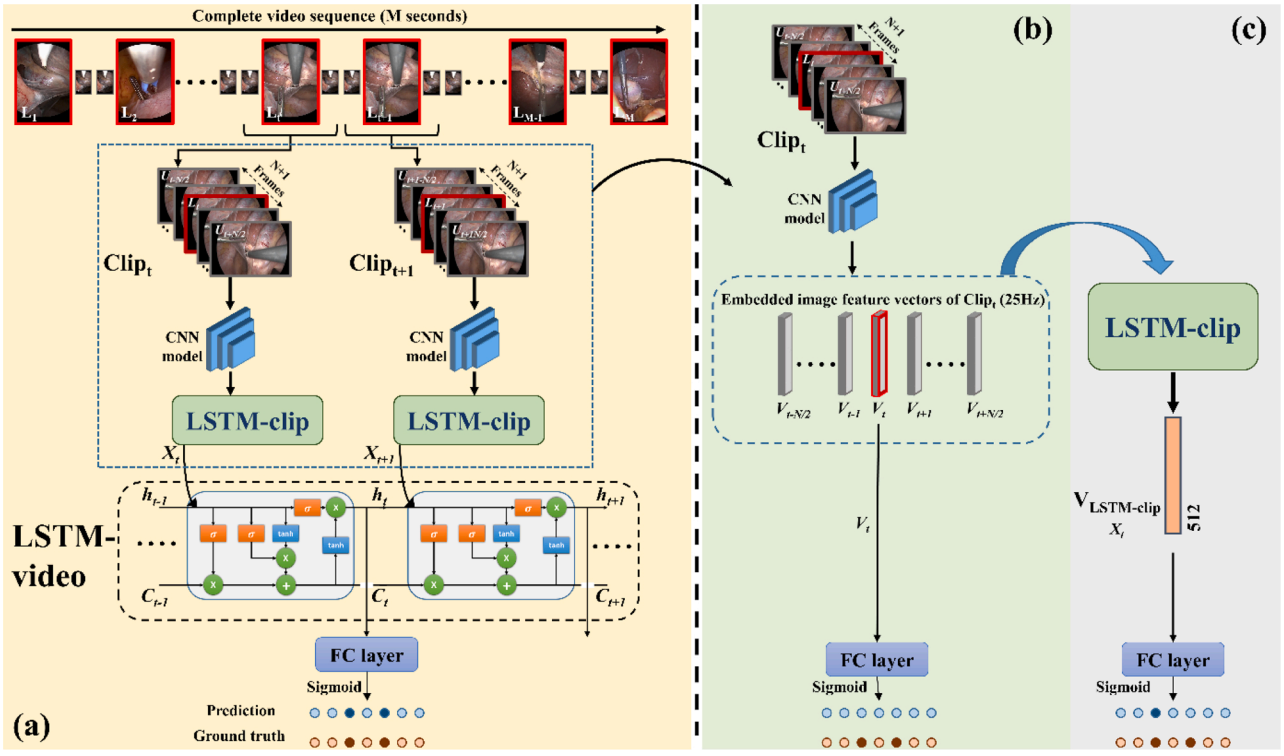

Fig. 2. The proposed framework for tool presence detection in laparoscopic videos. (a) Overview of the full pipeline, (b) the CNN approach, (c) the LSTM-clip approach. Red rectangles and grey rectangles refer to labelled (L) and unlabelled (U) frames, respectively. $V_{t\pm N/2}$ represents visual feature vectors extracted using the CNN model where N = 20, $V_{t\pm N/2}$ dimension is 4096 or 2048 when the VGG-16 or ResNet-50 was used, X_M represents feature vectors extracted using LSTM-clip ($V_{LSTM\text{-}clip}$) where M is complete video length, X_M dimension is 512.

available datasets are mostly sparsely annotated (e.g. the Cholec80 dataset was labelled at 1 fps), there is a potential to model fine-level temporal information in short sequences of the labelled frame and its neighbouring unlabelled frames. Hence, it is possible that misclassifications caused by the obscured images can be revised (see Fig. 1). Moreover, many surgeries consist of several procedures that are performed in a systematic way, where the surgeon uses a specific toolset to undertake each surgical phase. Thus, learning temporal information along the complete surgical video may enable more accurate tool detection.

In this paper, tool presence detection in laparoscopic videos was considered as a spatial-temporal problem. A cascade of two LSTM models were utilised to explicitly model temporal information firstly in short video sequences and then along the entire surgical video. Initially, a CNN model was trained to perform surgical tool detection and phase recognition solely based on images. This model was then used as visual feature extractor from laparoscopic images. Then, an LSTM model, termed as LSTM-clip, was employed to learn temporal dependencies from visual feature vectors of short video clips consisting of labelled and unlabelled frames. On top of the LSTM-clip, another LSTM network, termed as LSTM-video, was added to learn temporal information along the complete video sequence. Finally, the proposed approach was extensively evaluated on the publicly available dataset Cholec80. The detection performance of these approaches was evaluated using the average precision (AP) and compared with the state-of-the-art methods.

2. Method

The proposed method consists of two main approaches: CNN and RNN. The CNN encodes spatial information of endoscopic images, while the RNN captures temporal dependencies across sequential data. Two LSTM models (LSTM-clip and LSTM-video), which are both a type of RNN, were used to model temporal information in short video clips and complete endoscopic videos, respectively. A flow diagram showing the methods is shown in Fig. 2. The CNN, the LSTM-clip and the LSTM-video models were trained separately.

2.1. CNN-feature extraction

Two state-of-the-art CNN models, namely VGG-16 [23] and ResNet-50 [24], were applied to the same dataset to provide comparator statistics (see Section 2.5 for ablation study). The models were trained after modulating their architectures as in [21] to recognise surgical tools and phases in a multi-task manner. Hence, the CNN models learnt high-level features that are discriminative descriptors for surgical tools in the spatial dimension. The VGG-16 model consists of a stack of five convolutional blocks followed by three fully-connected layers. The last layer was replaced by a fully-connected layer (fc_{tools}) that performed tool detection. The output of the fc_{tools} layer was concatenated with the output of the previous layer and passed to another fully-connected layer (fc_{phase}) that performed surgical phase recognition. Each of the fc_{tools} and fc_{phase} layers has a number of nodes equals to the number of the defined tools and phases in the dataset, respectively. Similarly, the ResNet-50 model was adopted by substituting the fully-connected layer by the fc_{tools} layer and concatenating the fc_{tools} output with the output of the average pooling layer in the fc_{phase} layer. The trained VGG-16 and ResNet-50 models were utilized as feature extractors to represent images as 4096-dimensional or 2048-dimensional feature vectors, respectively. The feature vectors of short video clips were forwarded to the first LSTM model.

2.2. Temporal models

Endoscopic videos are a large source of static and sequential information. CNN can only spatially model visual information of static images. Two LSTM units were introduced to also capture temporal tool occurrence in sequential images. The first LSTM, termed as LSTM-clip, has a sequence-to-one configuration. LSTM-clip considers a sequence of N unlabelled frames surrounding the target frame F_t (labelled frame) to predict tools presence in F_t. Accordingly, the training videos were segmented into short clips and thereafter a sequence of CNN-feature vectors was extracted for each video clip and forwarded to LSTM-clip. The LSTM-clip outputs C dimensional feature vector $V_{LSTM\text{-}clip}$ which is passed to a fully-connected layer with seven nodes to perform tool classification, where C is the number of memory cells in the LSTM unit.

The second LSTM unit, termed as LSTM-video, was employed to incorporate sequential information along consecutive clips of a complete video. To this end, $V_{LSTM\text{-}clip}$ was extracted from every clip in a video using LSTM-clip to form a sequence of feature vectors. The sequences produced by LSTM-clip form the input of the LSTM-video model. This model has, similar to the LSTM-clip model, a fully-connected layer for tool classification.

Table 1
Description for applied approaches in the conducted experiments.

Approach	Model		Description
	Exp. 1	Exp. 2	
CNN	VGG	ResNet	The CNN model was trained to perform tool and phase recognition simultaneously.
CNN → LSTM-clip	VGG-LC	ResNet-LC	The LSTM-clip was trained with CNN features to encode temporal information in short sequences.
CNN → LSTM-clip → LSTM-video	VGG-LC-LV	ResNet-LC-LV	The LSTM-video was employed to model temporal information along the complete video sequence

2.3. Evaluation

2.3.1. Dataset

The Cholec80 dataset, constructed by A. P. Twinanda et al. [21], was used in this work. It contains endoscopic videos of 80 cholecystectomy procedures executed by 13 surgeons at University Hospital of Strasbourg. The videos were recorded at 25 frames per second (fps) and had surgical phases labelled continuously and surgical tools labelled at 1 Hz. Thus, every tool-labelled frame is surrounded with 48 unlabelled frames. The surgeons used seven tools to perform the endoscopic intervention; the surgical tools are: Grasper, Bipolar, Hook, Clipper, Scissors Irrigator, and Specimen bag. The videos had a median length of 2095 s (min 739, max 5993, first quartile 1641, and third quartile 2882)

2.3.2. Training setup

Six-fold Monte Carlo cross-validation (MCCV) was employed to evaluate the models. For each cross-validation fold, 40 randomly selected, full videos were used for training the CNN, LSTM-clip and LSTM-video models. The remaining 40 videos were used for testing and validation of the predictive performance of the models.

Two experiments were performed using different CNN architectures as a foundation for the more advanced RNN implementations. The VGG-16 and ResNet-50 were employed in the first and second experiment, respectively (see Section 2.5 for ablation study). LSTM-clip and LSTM-video models were initially trained using either the VGG-16 or ResNet-50 model outcomes (visual features). These different approaches are termed ***VGG-LC*** and ***VGG-LC-LV*** for the LSTM models that utilised information from the VGG-16 model, and ***ResNet-LC*** and ***ResNet-LC-LV*** for the LSTM models that utilised information from the ResNet-50 model (see Table 1). Each of the approaches were trained separately six times using the training data of respective MCCV fold.

The CNN models were fine-tuned using the transfer learning approach. Each model was initialized with weights learnt from pre-training on ImageNet dataset [25]. The CNN models were trained for 10 epochs with a learning rate starting at $2{\cdot}10^{-3}$ and a weight decay of

$9{\cdot}10^{-4}$. Training images were shuffled at the beginning of every epoch, and a batch of 50 images was passed to the CNN model every training iteration.

Each of the temporal models consists of one LSTM layer with 512 cells for LSTM-clip and with 4096 cells for LSTM-video. They were trained for 30 epochs separately. The initial learning rate was set to 10^{-4} with a decaying factor of 10^{-3}. The length of video-clip was set to 21 frames, that includes the labelled frame F_t, 10 preceding and 10 succeeding unlabelled frames (see Section 2.5 for ablation study). The LSTM-clip model was trained on batches of CNN-features extracted from 50 video-clips. The video-clips of all training videos were shuffled every epoch. The trained LSTM-clip was used to extract 512-dimensional $V_{LSTM\text{-}clip}$ for every clip in a video. For every video, a sequence was constituted from the extracted feature vectors. The LSTM-video was trained using a batch of one sequence every iteration, and, therefore, padding shorter sequences was not required.

All approaches were trained to perform multi-label binary classification for surgical tools. Therefore, the fully-connected layer, performing surgical tool classification in each model, has a sigmoid activation function. However, the layer that predicts the surgical phase in the CNN models has a softmax activation function. The binary cross-entropy function was employed to compute the loss for surgical tool prediction as shown in Eq. 1.

$$loss_t = \frac{-1}{N}\sum_{n=1}^{N}[l_n\ log(\sigma(\delta_n)) + (1-l_n)\ log(1-\sigma(\delta_n))] \quad (1)$$

where $loss_t$ is the loss for tool t, N is the batch size, l_n is the binary label of the particular tool, σ is a sigmoid function, and δ_n is the confidence of tool presence. The loss for the phase recognition task was computed using the softmax multinomial logistic loss function defined in Eq. 2.

$$loss_p = \frac{-1}{N}\sum_{n=1}^{N}\sum_{i=1}^{P} l_n^i \log\left(\varphi\left(w_n^i\right)\right) \quad (2)$$

where $loss_P$ is the loss for surgical phases, N is the batch size, P is the number of phases, l_n^p is the phase label, w is the output of the fc_{phase} layer and φ is the softmax function. An Adam optimizer [26] was used to minimize the loss function during the training process.

The implementation was carried out using the Keras framework. Training the CNN and LSTM models in both experiments took roughly 10 h for every MCCV-fold. Additional time was required to extract feature vectors from the CNN and LSTM-clip models. The feature extraction rate was approximately 100 fps using the CNN model (VGG-16 or ResNet-50) and 1000 fps using the LSTM-clip. The experiments were conducted on a computer with an Intel Xeon 2.20 GHz CPU and an NVIDIA GeForce RTX 2080Ti GPU.

2.3.3. Evaluation metric

Average precision (*AP*) was used as an evaluation metric for the tool classification task. It is defined as the area under the precision-recall curve and can be calculated according to Eq. 3.

$$AP_t = \int_0^1 P_t \cdot\ d[R_t] \quad (3)$$

where P_t is the precision and R_t is the recall of tool t as a function of the confidence threshold.

2.4. Visualising class activation maps

For further analysis, the VGG-16, VGG-LC and VGG-LC-LV models were used. Cases where VGG-LC showed improvement over VGG-16 were investigated, and gradient weighted class activation maps were visualised for the labelled frame and its adjacent unlabelled frames. A class activation map shows regions of the image most distinctly engaged in the classification by using gradient information backpropagated into the last convolutional layer of the CNN model [27]. The class activation maps for some examples, where the VGG-LC revised detection results of the VGG-16 model, are presented in the Results section.

Table 2
Tool presence detection results of different CNN architectures on the Cholec80 dataset (bold values indicate best performance).

Architecture	VGG-16 [23]	ResNet-50 [24]	Densnet-121 [28]	EfficientNet-B0 [29]
Grasper	93.87	93.88	93.85	**94.90**
Bipolar	93.44	**94.70**	94.45	94.02
Hook	99.54	99.51	99.50	**99.55**
Scissors	69.46	78.18	**78.24**	73.71
Clipper	91.56	**94.26**	94.23	90.37
Irrigator	83.84	**89.60**	89.37	85.87
Specimen Bag	92.50	93.85	93.84	**94.23**
Mean	89.17	**92.00**	91.93	90.38

Table 3
Computation times of the CNN models. An NVIDIA GeForce RTX 2080Ti GPU was used for implementation.

Model	Training (h)	Test (ms/im)
VGG-16 [23]	6.34	17
ResNet-50 [24]	5.09	19
Densnet-121 [28]	7.48	22
EfficientNet-B0 [29]	7.34	23

Table 4
Ablative testing results of the ResNet-LC approach for increasing length of the video clip. Np and Ns represent the number of preceding and succeeding unlabelled frames respectively.

Clip length	5 Np = 5, Ns = 0	10 Np = 5, Ns = 5	10 Np = 10, Ns = 0	20 Np = 10, Ns = 10	40 Np = 20, Ns = 20
Grasper	93.78	94.56	93.84	94.70	94.80
Bipolar	95.23	95.83	95.46	96.05	96.21
Hook	99.54	99.61	99.56	99.68	99.70
Scissors	81.01	82.78	82.13	84.20	84.41
Clipper	94.97	95.63	95.46	96.32	96.42
Irrigator	89.95	90.84	90.08	91.51	91.53
Specimen Bag	94.59	94.90	94.74	95.48	95.56
Mean	92.73	93.45	93.04	93.99	94.09
Training time (minutes)	48	58	58	83	123

2.5. Ablation study

To choose a suitable feature extractor for our approach, ablation experiments with four CNN architectures (VGG-16 [23], ResNet-50 [24], DensNet-121 [28] and EfficientNet-B0 [29]) were performed. The results showed that ResNet-50 gave the best performance, while VGG-16 gave the lowest performance among other networks (see Table 2). The training times of the four CNNs are presented in Table 3. Since the proposed framework aimed at detecting surgical tool presence by employing temporal information, the networks with highest and lowest performances (ResNet-50 and VGG-16) were utilised as feature extractors to evaluate the proposed framework.

Additionally, ablation experiments were conducted to evaluate the effectiveness of different video clip lengths in the LSTM-clip approach. Table 4 lists the AP of each tool for ResNet-LC with increasing clip lengths. Indeed, increasing the clip length produced improvements for all tools, and the LSTM-clip approach gave almost equal performance with clip lengths in the range [20,40]. The training times of the LSTM-clip with video clip length of 40 was 123 min compared to 83 min with a

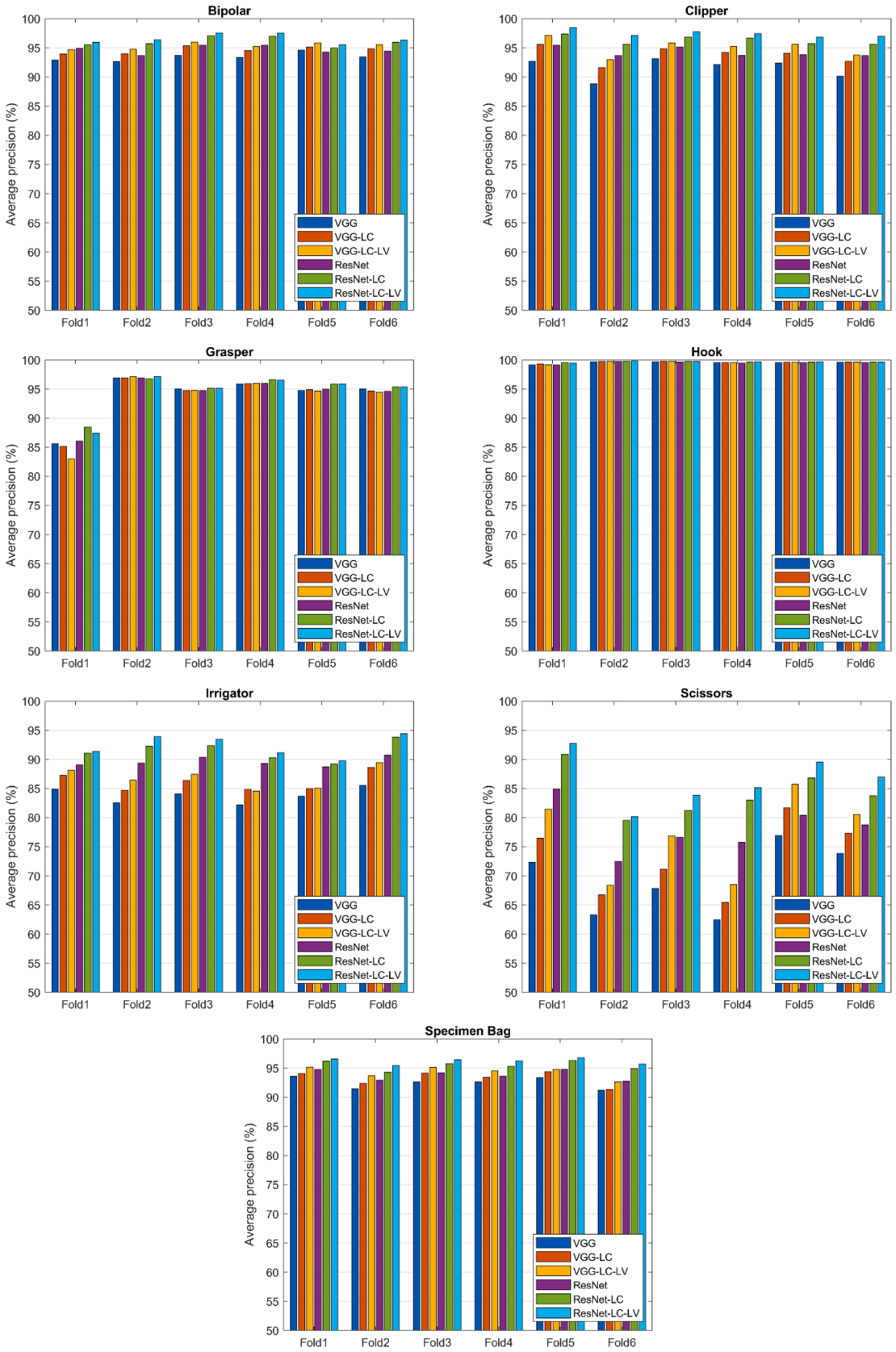

Fig. 3. Average precision of the tool presence detection for all tools (each with 6 fold models). Note the truncated scale of the y-axis.

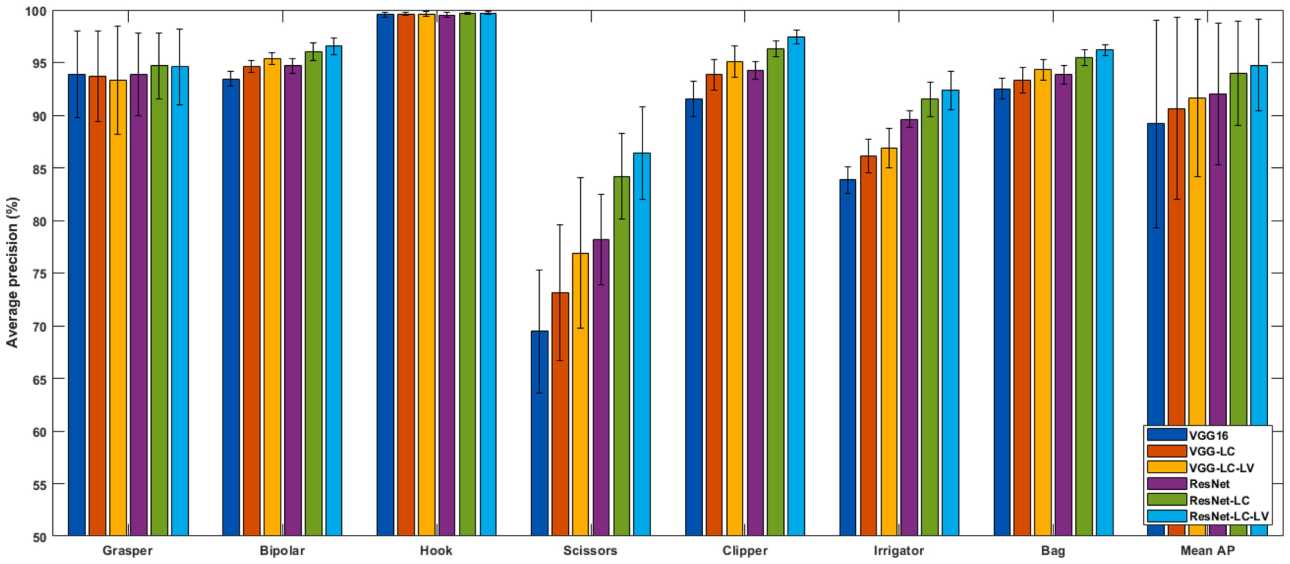

Fig. 4. Mean average precision (mAP) and standard deviation of the tool presence detection for all models. Note the truncated scale of the y-axis.

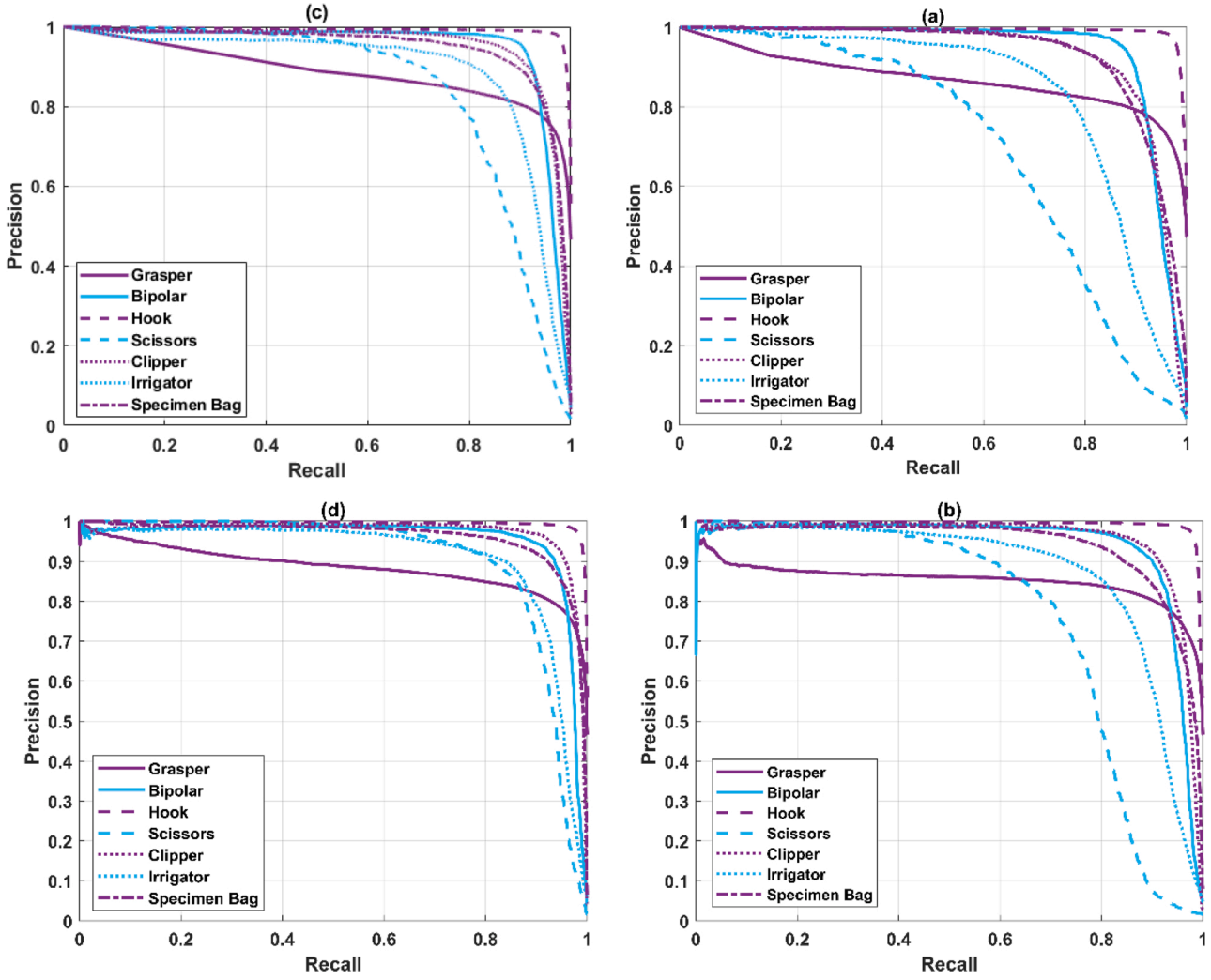

Fig. 5. Precision-Recall curves of the first fold. Each figure represents the performance of these models: (a) VGG-16; (b) VGG-LC; (c) ResNet-50; (d) ResNet-LC.

Table 5
Average precision of tool presence detection obtained from Fold 1 models; and a comparison with the state-of-the-art methods (bold values indicate best performance of each tool).

Tool	EndoNet [21]	Endo3D [17]	GCN [18]	Jin [22]	VGG-LC-LV	ResNet-LC-LV
Grasper	84.8	71.32	–	84.7	83.00	**87.40**
Bipolar	86.9	69.72	–	90.1	94.70	**95.95**
Hook	95.6	87.81	–	95.6	99.18	**99.45**
Scissors	58.6	87.33	–	86.7	81.44	**92.73**
Clipper	80.1	95.12	–	89.8	97.16	**98.50**
Irrigator	74.4	96.43	–	88.2	88.16	**91.37**
Specimen Bag	86.8	94.97	–	88.9	95.17	**96.58**
Mean	81.02	86.1	90.13	89.1	91.26	**94.57**

Table 6
Average precision of tool presence detection obtained from fold 1 models where the evaluation set is only the last thirty videos, and a comparison with the methods proposed in Vardazaryan et al. [16] and Nwoye et al. [20] (bold values indicate best performance of each tool).

Tool	Vardazaryan [16]	Nwoye [20]	ResNet-LC-LV
Grasper	96.8	**99.7**	85.54
Bipolar	94.2	95.6	**95.88**
Hook	99.6	**99.8**	99.36
Scissors	49.8	86.9	**92.39**
Clipper	83.0	97.5	**98.72**
Irrigator	93.3	74.7	**95.89**
Specimen Bag	94.0	96.1	**96.83**
Mean	87.2	92.9	**94.95**

video clip length of 20. Consequently, a video clip length of 20 unlabelled frames was chosen for the assessment of the LSTM-clip and LSTM-video models.

The proposed framework consists of a CNN followed by the LSTM-clip and then the LSTM-video. In order to show the advantages of each approach, the tool detection performance of each approach (CNN, CNN-LC and CNN-LC-LV) was evaluated and presented in Figs. 3 and 4.

3. Results

Fig. 3 shows AP of each tool obtained by the CNN, LSTM-clip and LSTM-video for the six folds. In this six-fold validation, results show the value of temporal information for tool classification in laparoscopic videos (see Figs. 4 and 5). The average precision of all tools except the grasper and hook enhanced after using the LSTM-clip approaches, and the most notable improvement reached after employing the LSTM-video. Fig. 4 shows the mean average precision (mAP) and the standard deviation of all tools for all models. Tables 5 and 6 present the comparison results of tool presence detection with reference the methods. The results shown in both tables were achieved by models that were trained with the first 40 videos. However, to achieve consistency with prior papers [16–18,20–22], the evaluation dataset for Table 5 was the last forty videos and only the last thirty videos for Table 6.

Figs. 6 and 7 show four examples of the performance improvement achieved by the LSTM-clip over the CNN model. The class activation maps of consecutive frames were also visualised to determine the image regions involved in the prediction process performed by the CNN. The short video clips represent the input of the LSTM-clip to perform tool presence detection in the form of sequence-to-one classification. Every video clip contains the labelled frame (stamped by t) and its adjacent frames (stamped by $t \pm n$). The activation maps are also labelled with the prediction probability of the examined tool class obtained by the CNN model. The examined tool class and the revised prediction probability obtained by the LSTM-clip are shown on the top of each clip.

4. Discussion

This study presents a deep learning framework for detecting surgical tool presence in laparoscopic videos by aggregating spatial and temporal features. Initially, a CNN model was fine-tuned to learn visual spatial features from images. Two LSTM networks were then sequentially employed to learn temporal information from short video sequences and the entire surgical videos, respectively. The framework was trained and tested on the Cholec80 dataset [21].

A six-fold experimental validation show both LSTM-clip and LSTM-video improved the total classification performance obtained by the established models. ***ResNet-LC-LV*** and ***VGG-LC-LV*** yielded mean mAP values of 94.74 % and 91.64 %, respectively. These values improved on the established VGG-16 and ResNet-50 models mAP values of 89.17 % and 92.00 %, respectively (Fig. 4). LSTM-clip contributed effectively to revising and improving predictions obtained by the CNN model in almost all tool cases by modelling temporal information from unlabelled adjacent frames. Similarly, using temporal information along the complete video sequence achieved the most notable improvement of all tools except the grasper and hook (see Figs. 3 and 4).

Fig. 6 shows the CNN sometimes failed to detect the correct tool, even in cases when the activation map matched the tool location in the image. For the first video clip in Fig. 6, the bipolar is in-frame. Attention maps from the CNN model are shown for the labelled frame and three previous unlabelled frames. The bipolar partially appears in the labelled image, hence tool presence prediction relying on the single frame information gives low prediction probability. In contrast, the previous three frames more clearly show the bipolar, and led to much higher probabilities from the CNN. Hence, the CNN-clip and CNN-video models were able to use the information from the unlabelled images to increase the confidence in prediction of the labelled image. Similarly, the CNN struggles to achieve high prediction confidence of the grasper in the second video clip (see frame t in Fig. 6b). The tissues come near to the laparoscopic light source, so the image background is dark, and the visual appearance of the grasper in the background is unclear. Conversely, prediction confidences from the CNN in the previous frames are higher. Again, the LSTM-clip and LSTM-video models were able to leverage this confidence.

For the third video clip (Fig. 6c), blood covering the tool tip caused misclassification by the CNN for the irrigator. However, the CNN is capable of detecting the irrigator in previous frames where more parts of the irrigator appear. The attention maps (Fig. 6) show that the CNN is not reliable for detecting surgical tools in some abnormal cases due to the obscured nature of laparoscopic images. Furthermore, these abnormalities last for a short period, and the LSTM-clip model is therefore relevant to generate more discriminative features by modelling temporal dependencies between adjacent frames. Indeed, temporal clues from previous frames are sufficient for predicting correct tool category in the video clips shown in Fig. 6, but are not enough in the video clip shown in Fig. 7 where information from following frames is also required. Hence, prior to this summative assessment of the LSTM-clip and LSTM-video models, a sensitivity analysis was carried out in the beginning of this work to investigate the effect of video clip length on improving tool detection performance. Ultimately, results not presented here indicated a video clip of 20 unlabelled frames gave optimal precision (see Section 2.5 for ablation study).

Typically, a surgical procedure can be described as a sequence of events that occur in specific surgical phases. These phases are performed in a specific order through the entire procedure, and each phase is accomplished using corresponding surgical tools. In other words, there is a correlation between surgical tool presence and surgical phase. Fig. 8 shows mean surgical tool occurrences in every surgical phase in the Cholec80 dataset. Therefore, modelling temporal dependencies across the entire procedure using the LSTM-video improved classification accuracies of almost all tools. However, LSTM-video has negligible effect on improving classification accuracy of the grasper since it appears in all

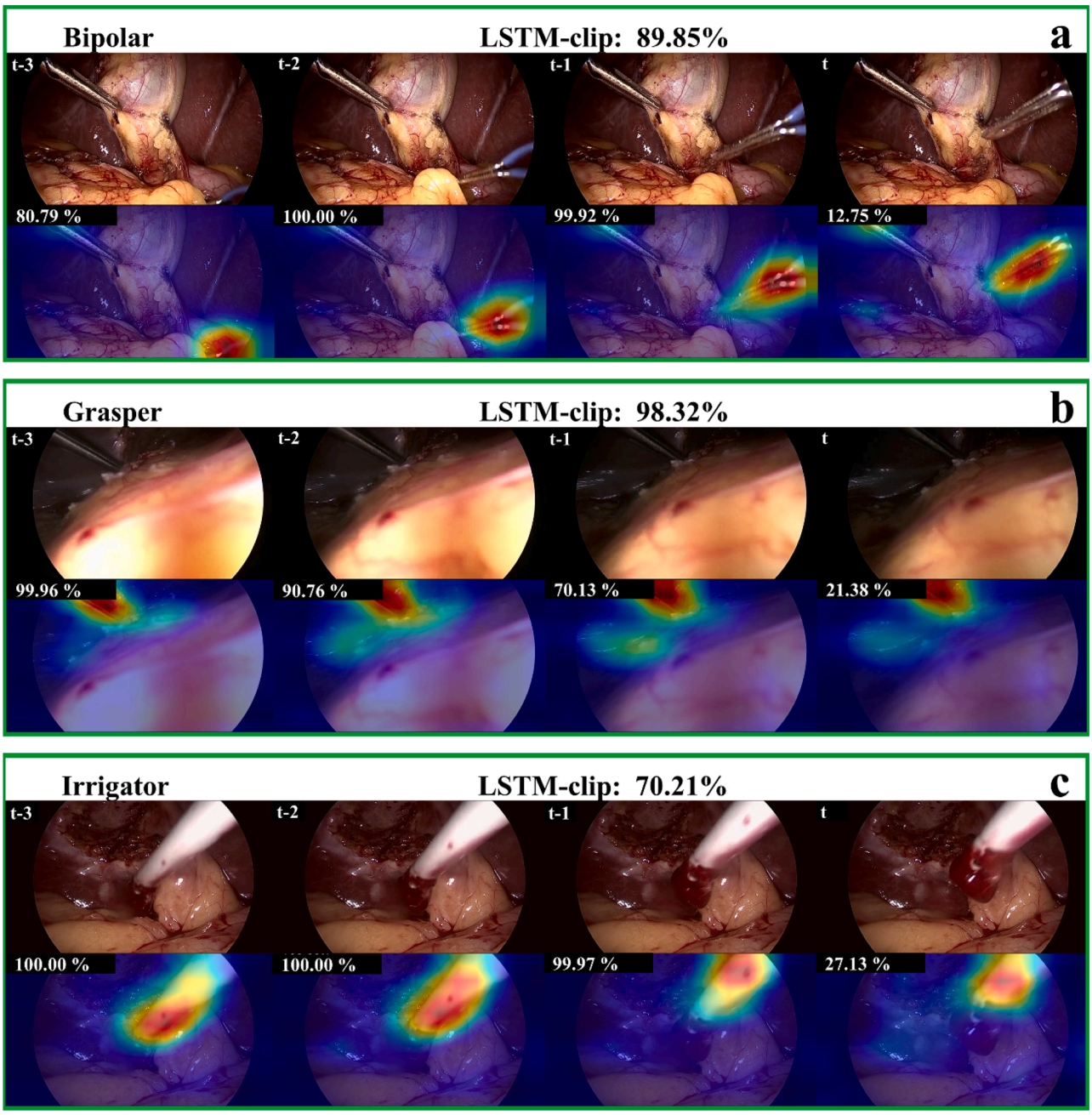

Fig. 6. Visualisation of class activation maps of three video clips indicating the parts of the images involved in the classification performance. (a) partial appearance of the surgical tool (the bipolar in the right side), (b) dark image background caused by tissue near to the light source, (c) blood covering the tool tip. All examples represent cases where temporal information from only the previous frames are required to revise classification by the CNN. For each video clip, the examined tool class is noted on the top-left corner, and the prediction probability towards the examined tool of frame t using LSTM-clip is presented on the top-mid. Labelled frame is indicated by t, and other frames are indicated by t ± n (n: [1–6]). Each attention map is labelled by the prediction probability of the tool obtained by the VGG-16 model.

surgical phases (Fig. 8) and therefore no discriminative temporal information can be learnt for this tool.

Table 5 shows the leading methods and their corresponding detection precision. Twinanda et al. introduced a CNN model called EndoNet that was trained in a multi-task manner to perform tool presence detection and surgical phase recognition and achieved mean AP of 81.02 % [21]. A more recent study Jin et al. addressed the problem in a similar fashion to Twinanda, but they also introduced a correlation loss to model the relation between phase recognition and tool presence detection, and they reported a mAP of 89.10 % [22]. However, both approaches [21,22] did not consider temporal information to detect surgical tool presence. In this study, the CNN models were trained to perform both surgical phase recognition and tool presence detection tasks similarly to Twinanda. Furthermore, a cascade of two LSTMs was employed to capture valuable temporal information that was not directly considered by the current state-of-the-art methods. It is the incorporation of temporal information that allowed the LSTM-clip and LSTM-video models to exceed the state-of-the-art in tool recognition (Tables 5 and 6).

Wang et al. [18] and Chen et al. [17] tackled tool detection in laparoscopic videos in a similar manner to this study. Wang et al. proposed a deep learning frame work consisting of a 3D inflated DensNet to

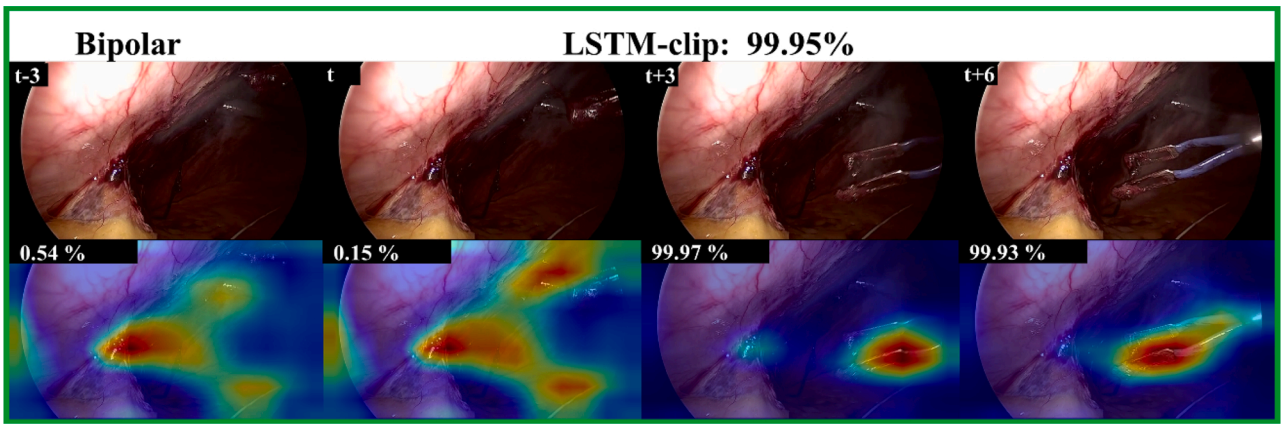

Fig. 7. Visualisation of class activation maps of a video clip where temporal information from previous and following frames are required to revise classifications obtained by the CNN model at t.

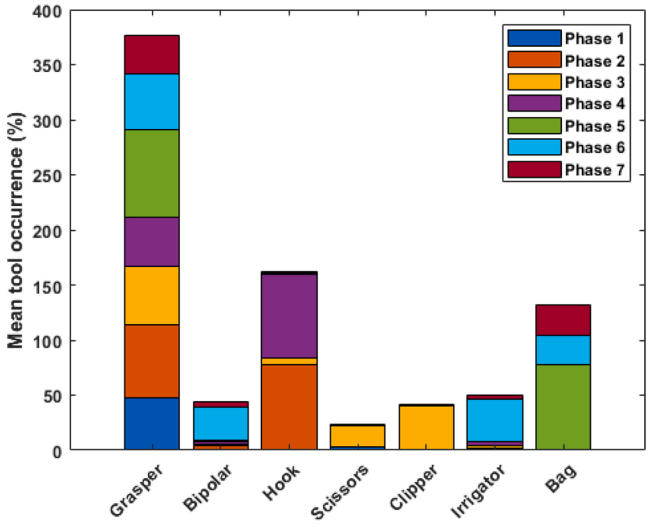

Fig. 8. Mean tool occurrence in different surgical phases in the Cholec80 dataset.

learn visual features from images and a Graph Convolutional Network (GCN) that learns temporal dependencies across short video clips [18]. Chen et al. suggested using a 3D convolutional network to learn spatial-temporal features from short video sequences [17]. While Wang et al. reported a value of 90.13 %, and Chen et al. reported 86.10 % for mAP, the models presented here (VGG-LC-LV and ResNet-LC-LV) achieved higher mAP values than [17,18] with 91.26 % and 94.57 %, respectively. This added precision was enabled by incorporating information from the unlabelled frames and also temporal dependencies across the entire surgical procedure.

Vardazaryan et al. proposed a deep architecture that performs both tool presence detection and tool localisation [16]. Their method relied on using ResNet-18 as a base model followed by a convolutional layer to generate localisation maps of the seven tools. These maps were transformed using a pooling operation into a vector that represents tool confidences. Nwoye et al. used similar architecture, but they added a Convolutional LSTM (ConvLSTM) to perform surgical tool tracking by learning spatiotemporal features [20]. Vardazaryan et al. and Nwoye et al. reported mAP values of 87.20 % and 92.90 %, respectively. Their findings highlight the value of temporal information for tool presence detection. However, the methods presented in this paper achieved a higher value on the same evaluation set when ResNet-50 was used as a base model with 94.95 % mAP (Table 6). The detection precision of the scissors was significantly improved by the proposed approach. This may be due to the deeper CNN model (ResNet-50) and temporal information within the current approach. However, in contrast to Vardazaryan et al. and Nwoye et al., the proposed approach performs only tool presence detection but not tool localisation or tool tracking. Hence, the present study optimises a different goal. Nonetheless, the class activation maps presented in Figs. 6 and 7 demonstrate the potential for augmentation of the proposed approach to enable tool localisation in addition to tool presence detection.

The dataset was divided into two equal datasets to train and test the models. Each dataset contains equal number of surgical videos (n = 40), but these videos had different lengths and different tools distributions. Therefore, high variance in performance was noted between the models trained with different training sets. For instance, the average precision of the grasper obtained by the VGG-16 or ResNet-50 models of the first fold is lower than that obtained by the second fold same model. This high variance is interpreted by the different distributions of the grasper between the training sets of these two folds as seen in Fig. 9. The grasper appears in around 57k images in the first subset compared to 46k images in the second subset. In this case, there was a higher false positive rate for the grasper in the first case lowering the precision.

Fig. 4 shows the level of variability from the six-fold validation. In contrast, almost all established methods presented results from a singular data split for training and testing. This is understandable given the constraints of the computationally intensive optimisation process. However, such approaches can lead to results that are most applicable in the specific division of training and testing datasets. In contrast, this study used a six-fold cross-validation experiment to show the expected variability of results. Furthermore, it showed that the improvements offered by the approach were not obtained via false optimisation to the unique characteristics of the training and validation data.

Despite the high classification performance of the models presented, the study had some limitations. The three models were trained separately, and there is potential for fusing them in an end-to-end framework to generate more discriminative spatiotemporal features. In addition, to reduce the effect of the imbalanced dataset on the training process, data augmentation can be considered, and the models could potentially be trained with a weighted loss [12,14]. Furthermore, the proposed approach only predicted surgical tool presence but not tool localization or segmentation. It is possible that future work could adapt the current models to achieve surgical tool segmentation and localization, as well as the tool presence detection undertaken in this study. Finally, the proposed approach was only evaluated on cholecystectomy surgery where only seven tools were utilised to perform the surgical procedure. It may be interesting to evaluate the method on more complex surgeries such as the sigmoid resection where a greater number of tools are used.

In conclusion, this study proposed a deep learning approach to detect surgical tool presence in laparoscopic videos. The proposed approach

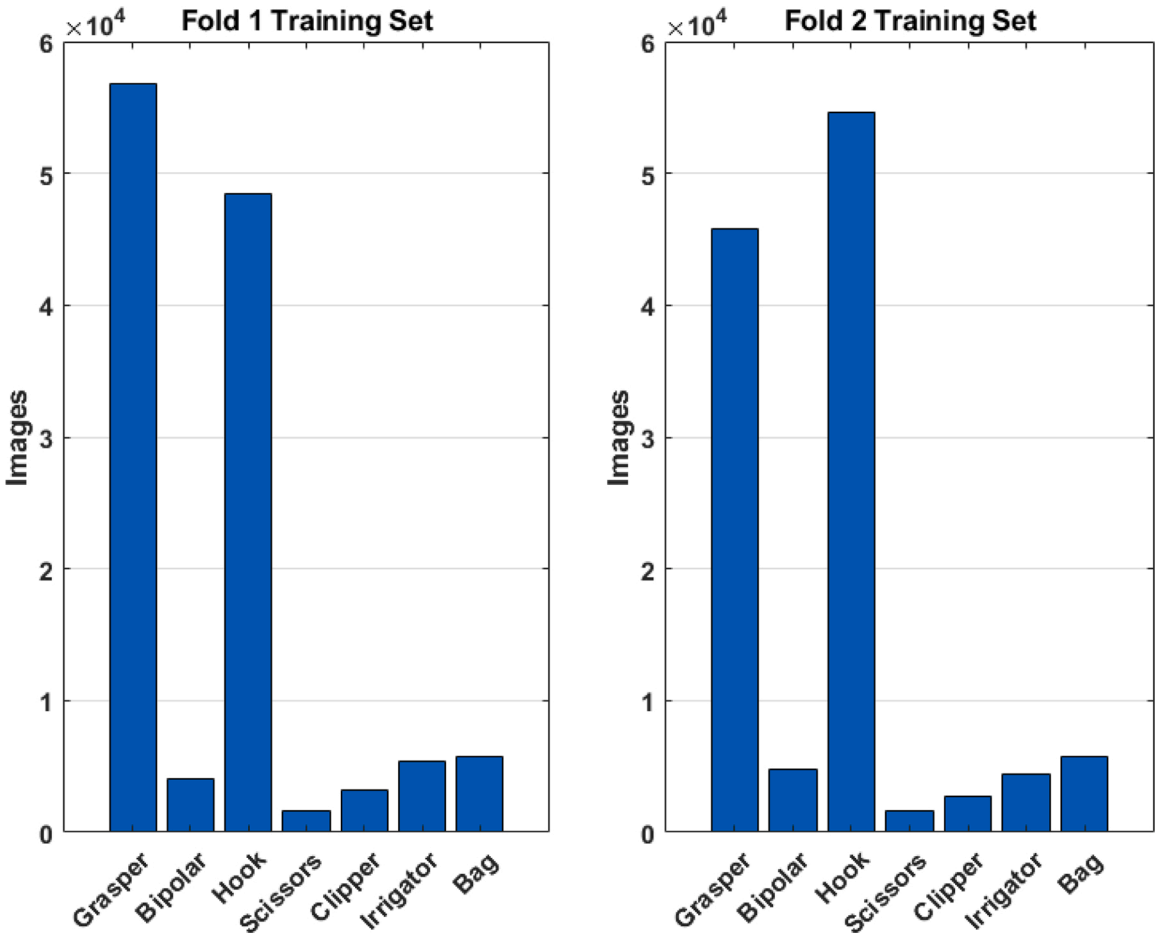

Fig. 9. Distribution of surgical tools in the training set for the first and second folds.

models temporal dependencies across short video clips around the classified frames and through the entire video. Experimental results of this method showed high tool detection performance that exceeds the state-of-the-art methods, indicating that this approach is very promising for developing intelligent systems inside ORs.

CRediT authorship contribution statement

Tamer Abdulbaki Alshirbaji: Conceptualization, Methodology, Software, Formal analysis, Investigation, Writing - original draft. **Nour Aldeen Jalal:** Conceptualization, Methodology, Formal analysis, Investigation, Writing - original draft, Visualization. **Paul D. Docherty:** Writing - review & editing. **Thomas Neumuth:** Writing - review & editing, Supervision. **Knut Möller:** Conceptualization, Writing - review & editing, Supervision, Project administration, Funding acquisition.

Acknowledgments

This work was supported by the German Federal Ministry of Research and Education (BMBF under grant CoHMed/IntelliMed grant no. 13FH5I01IA and 13FH5I05IA) and H2020 MSCA Rise (#872488—DCPM).

Declaration of Competing Interest

The authors declare no conflicts of interest.

References

[1] F. Lalys, P. Jannin, Surgical process modelling: a review, Int. J. Comput. Assist. Radiol. Surg. 9 (3) (2014) 495–511.
[2] N. Padoy, Machine and deep learning for workflow recognition during surgery, Minim. Invasive Ther. Allied Technol. 28 (2) (2019) 82–90.
[3] D. Bouget, M. Allan, D. Stoyanov, P. Jannin, Vision-based and marker-less surgical tool detection and tracking: a review of the literature, Med. Image Anal. 35 (2017) 633–654.
[4] L. Maier-Hein, S. Vedula, S. Speidel, N. Navab, R. Kikinis, A. Park, M. Eisenmann, H. Feussner, G. Forestier, S. Giannarou, Surgical Data Science: Enabling Next-Generation Surgery, 2017 arXiv preprint arXiv:170106482.
[5] N. Padoy, T. Blum, S.A. Ahmadi, H. Feussner, M.O. Berger, N. Navab, Statistical modeling and recognition of surgical workflow, Med. Image Anal. 16 (3) (2012) 632–641.
[6] O. Dergachyova, D. Bouget, A. Huaulme, X. Morandi, P. Jannin, Automatic data-driven real-time segmentation and recognition of surgical workflow, Int. J. Comput. Assist. Radiol. Surg. 11 (6) (2016) 1081–1089.
[7] S.-A. Ahmadi, T. Sielhorst, R. Stauder, M. Horn, H. Feussner, N. Navab, Recovery of surgical workflow without explicit models, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2006, pp. 420–428.
[8] N.A. Jalal, T.A. Alshirbaji, K. Möller, Predicting surgical phases using CNN-NARX neural network, Curr. Dir. Biomed. Eng. 5 (1) (2019) 405–407.
[9] M. Kranzfelder, A. Schneider, A. Fiolka, E. Schwan, S. Gillen, D. Wilhelm, R. Schirren, S. Reiser, B. Jensen, H. Feussner, Real-time instrument detection in minimally invasive surgery using radiofrequency identification technology, J. Surg. Res. 185 (2) (2013) 704–710.
[10] H. Al Hajj, M. Lamard, P.-H. Conze, B. Cochener, G. Quellec, Monitoring tool usage in surgery videos using boosted convolutional and recurrent neural networks, Med. Image Anal. 47 (2018) 203–218.
[11] K. Mishra, R. Sathish, D. Sheet, Learning latent temporal connectionism of deep residual visual abstractions for identifying surgical tools in laparoscopy procedures, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (2017) 58–65.
[12] M. Sahu, A. Mukhopadhyay, A. Szengel, S. Zachow, Addressing multi-label imbalance problem of surgical tool detection using CNN, Int. J. Comput. Assist. Radiol. Surg. 12 (6) (2017) 1013–1020.

[13] S. Bodenstedt, A. Ohnemus, D. Katic, A.-L. Wekerle, M. Wagner, H. Kenngott, B. Müller-Stich, R. Dillmann, S. Speidel, Real-time Image-Based Instrument Classification for Laparoscopic Surgery, 2018 arXiv preprint arXiv:180800178.
[14] T.A. Alshirbaji, N.A. Jalal, K. Möller, Surgical tool classification in laparoscopic videos using convolutional neural network, Curr. Dir. Biomed. Eng. 4 (1) (2018) 407–410.
[15] S. Haase, J. Wasza, T. Kilgus, J. Hornegger, Laparoscopic instrument localization using a 3-D Time-of-Flight/RGB endoscope, in: 2013 IEEE Workshop on Applications of Computer Vision (WACV), IEEE, 2013, pp. 449–454.
[16] A. Vardazaryan, D. Mutter, J. Marescaux, N. Padoy, Weakly-supervised learning for tool localization in laparoscopic videos. Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis, Springer International Publishing, Cham, 2018, pp. 169–179.
[17] W. Chen, J. Feng, J. Lu, J. Zhou, Endo3d: online workflow analysis for endoscopic surgeries based on 3d cnn and lstm. OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis, Springer, 2018, pp. 97–107.
[18] S. Wang, Z. Xu, C. Yan, J. Huang, Graph convolutional nets for tool presence detection in surgical videos, International Conference on Information Processing in Medical Imaging (2019) 467–478.
[19] D. Bouget, R. Benenson, M. Omran, L. Riffaud, B. Schiele, P. Jannin, Detecting surgical tools by modelling local appearance and global shape, IEEE Trans. Med. Imaging 34 (12) (2015) 2603–2617.
[20] C.I. Nwoye, D. Mutter, J. Marescaux, N. Padoy, Weakly supervised convolutional LSTM approach for tool tracking in laparoscopic videos, Int. J. Comput. Assist. Radiol. Surg. 14 (6) (2019) 1059–1067.
[21] A.P. Twinanda, S. Shehata, D. Mutter, J. Marescaux, M. de Mathelin, N. Padoy, EndoNet: a deep architecture for recognition tasks on laparoscopic videos, IEEE Trans. Med. Imaging 36 (1) (2017) 86–97.
[22] Y. Jin, H. Li, Q. Dou, H. Chen, J. Qin, C.-W. Fu, P.-A. Heng, Multi-task recurrent convolutional network with correlation loss for surgical video analysis, Med. Image Anal. 59 (2020) 101572.
[23] K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, 2014 arXiv preprint arXiv:14091556.
[24] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016) 770–778.
[25] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: a large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2009, pp. 248–255.
[26] D.P. Kingma, J. Ba, Adam: A Method for Stochastic Optimization, 2014 arXiv preprint arXiv:14126980.
[27] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: visual explanations from deep networks via gradient-based localization, Proceedings of the IEEE International Conference on Computer Vision (2017) 618–626.
[28] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2017) 4700–4708.
[29] M. Tan, Q. Le, Efficientnet: rethinking model scaling for convolutional neural networks, in: International Conference on Machine Learning, PMLR, 2019, pp. 6105–6114.