IJCARS manuscript No.
(will be inserted by the editor)

Addressing multi-label imbalance problem of Surgical Tool Detection using CNN

Manish Sahu · Anirban Mukhopadhyay · Angelika Szengel · Stefan Zachow

Received: date / Accepted: date

Abstract

Purpose: A fully automated surgical tool detection framework is proposed for endoscopic video streams. State-of-the-art surgical tool detection methods rely on supervised *one-vs-all* or *multi-class* classification techniques, completely ignoring the co-occurrence relationship of the tools and the associated class imbalance.

Methods: In this paper, we formulate tool detection as a *multi-label* classification task where tool co-occurrences are treated as separate classes. In addition, imbalance on tool co-occurrences is analyzed and stratification techniques are employed to address the imbalance during Convolutional Neural Network (CNN) training. Moreover, temporal smoothing is introduced as an online post-processing step to enhance run time prediction.

Results: Quantitative analysis is performed on the *M2CAI16 tool detection dataset* to highlight the importance of stratification, temporal smoothing and the overall framework for tool detection.

Conclusion: The analysis on tool imbalance, backed by the empirical results indicates the need and superiority of the proposed framework over state-of-the-art techniques.

Keywords Transfer Learning · Surgical Tool Detection · CNN · Laparoscopic Videos · Multi-label Learning

1 Introduction

Fast and accurate recognition of surgical workflow plays an important role in modern Computer Assisted Intervention (CAI). Modern Operating Rooms (OR) demand monitoring of surgical processes to reduce preventable errors, the absence

Manish Sahu, Anirban Mukhopadhyay, Angelika Szengel and Stefan Zachow
Zuse Institute Berlin, Berlin, Germany
E-mail: sahu@zib.de

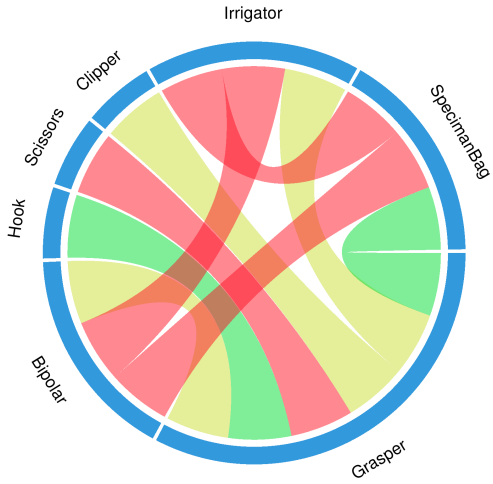

Fig. 1 *Chord diagram* [10] showing the second order co-occurrences of tools (two-tools together) in *M2CAI16 tool detection training* dataset. Tool usage frequency is color-coded in a discrete fashion: red (0-160), yellow (160-1000) and green ($>$ 1000).

of which may result in failures upto the loss of human lives [1]. In addition, multitudes of other OR procedures, for example automated clinical assistance, staff assignment etc. can benefit from surgical workflow recognition [2;3;4]. Recent CAI literature [5;6;2;7] has identified that surgical tool occurrences are closely related to the phases of surgical workflow. Moreover, tool detection and tracking on endoscopic images has the potential of controlling a robot mounted endoscopic camera holder, especially for solo surgeries [8;9]. Change of illumination, specular reflection and partial occlusion are some of the major challenges that render surgical tool detection a challenging task. This work mainly focuses on a fully automatic identification of surgical tool(s) from endoscopic video streams.

State-of-the-art methods treat surgical tool detection as a supervised *multi-class* or *one-vs-all* classification task [2]. Based on the observation that multiple tool co-occurrences happen quite often in endoscopic video frames, we have formulated the task as a generalized *multi-label* classification. The co-occurrence of tools, formally termed as *label-sets*, forms the output set of *multi-label* classification. In particular, rather than treating each tool as a stand-alone class, we've considered *label-sets* of multiple tools, along with the introduction of a no-tool (i.e. background) class. Second order co-occurrences of tools along with frequency of occurrences are visualized in Figure 1 for intuitive understanding. Interesting tool co-occurrence relationship patterns emerge from such visualization. For example, though *hook* is the most often used tool in surgical intervention, in second order, it is used only in association with grasper. Other interesting relationships involving two tools usage can also be inferred from Figure 1.

One key observation of this work is the imbalance associated with the tool usage (could be understood intuitively from Figure 1). It has already been proven that the imbalance in data affects the *binary* and *multi-class* classification accuracy [11]. However, the imbalance associated with the tool usage during a surgery

has not been quantified before. In this work, imbalance in tool usage during an endoscopic intervention is analyzed quantitatively using the measures introduced by Charte et al.[11]. Moreover, the effects of tool usage imbalance on detection accuracy is quantitatively analyzed in a novel experimental setup and specific sampling strategy to address imbalance[12] in tool usage is introduced during CNN training.

Major contributions of this work are twofold. First, surgical tool detection is performed in a generalized *multi-label* setting, with quantitative analysis focusing on handling surgical tool imbalance. To the best of our knowledge, this is the first work where tool detection is formulated as a general *multi-label* classification problem. Secondly, a novel transfer learning architecture is proposed for fine tuning and domain adaptation of AlexNet[13] towards a surgical tool identification task. In particular, weighted *uni-variate* loss (as learning objective) for joint output distribution is adopted for handling residual tool imbalance after stratification.

2 Related Work

Surgical tool detection within endoscopic videos has received increasing attention in recent years[6;14]. For the sake of brevity, we have mainly considered works where endoscopic video streams are considered as the sole input source. Tool detection procedures from video streams often consider the task as both tool identification and localization problem[6]. For example, Speidel et al.[15] suggested a tool identification pipeline that consists of segmentation and 3D model based processing. Sahu et al.[9] proposed detection and tracking of surgical tools over the virtual control interface on endoscopic stream to control a robot.

This work, however, similar to Twinanda et al.[2], considers tool detection as a tool presence detection task without explicit localization. For the rest of the paper, 'tool detection' is commonly used to refer to automatic tool identification from endoscopic video frames. Twinanda et al.[2] proposed deep learning based features to be used in *one-vs-all* classification framework for tool identification without localization. Most recently, performance of different tool detection techniques[16;17;18] is quantitatively evaluated in *M2CAI16 tool detection challenge*. In particular, Twinanda et al.[16] used ToolNet - a network very similar to AlexNet with a final tool detection layer. Raju et al.[18] used an ensemble of two networks and Sahu et[17] consisted of a *multi-label* learning approach with no stratification, followed by random forest for classification. In this work, unlike other proposed techniques, we have generalized the problem as *multi-label* classification task, quantitatively analyzed the imbalance and adopted strategies to overcome imbalance related issues.

3 Method

In this section, we first provide an overview of *multi-label* classification for tool detection task. Next we describe metrics used for defining levels of imbalance in a *multi-label* dataset in Section 3.2; followed by a description of stratification technique used to address *multi-label* imbalance in Section 3.3. In Section 3.4 ZIBNet architecture is introduced with novel design choices that are incorporated during

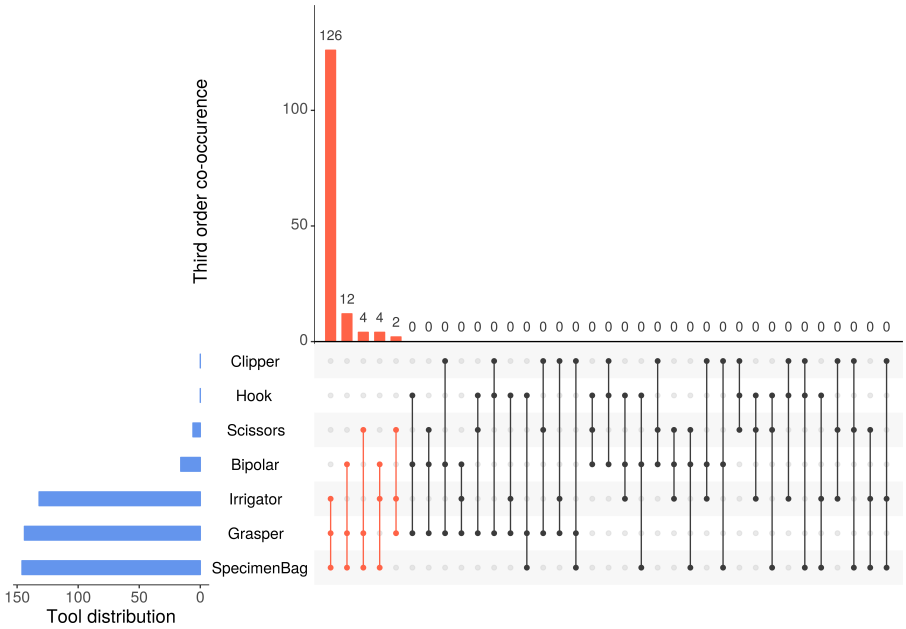

Fig. 2 *Upset* [19] visualization of the third order co-occurrence (three tools appearing together) of the tools in *M2CAI16 tool detection training* dataset. It shows possible three tool combinations (bottom right) with orange and black color representing presence and absence of these combinations in the training dataset respectively. The third order co-occurrences of the tools are shown on top with corresponding individual tool distribution on bottom left.

learning. Finally, we propose temporal smoothing as an online post processing step which suppresses false positives during prediction.

3.1 Multi-label Classification

Supervised classification based surgical tool detection have focused on formulating the problem in a *one-vs-all* or *multi-class* setting. Even though simple and intuitive, in this paper we argue that these settings do not address the problem in its general sense. In particular, due to the co-occurrence of multiple surgical tools at different endoscopic video frames, general *multi-label* classification should be used to model tool presence instead. *Multi-label* classification is the generalization of *binary* or *multi-class* classification. In this scenario, no a-priori limit on the number of tools present in the output set is imposed during classification.

For example, second order co-occurrences of surgical tools (i.e. two tools appearing together) is reported in Figure 1. In *one-vs-all* or *multi-class* classification setting, the desired chord diagram would be an empty circle, penalizing all the interconnected entries. However, plotting the ground truth annotation in Figure 1 suggests the existence of interconnected entries (more than one tool), which would have been penalized in the earlier settings. Similarly, third order tool co-occurrence with respect to instrument distribution is visualized in Figure 2 which shows distribution of three tools appearing together. The presence of the co-occurrences in

Figure 1 and 2 violates the mutually exclusive class assumption of *one-vs-all* or *multi-class* classification. This motivates us to model the problem as a *multi-label* classification one where co-occurrence entries are also considered valid and not penalized during classification.

Formally, for all N annotated video frames in our training dataset $F \in \{f_i\}$ where $i = 1, 2, ...N$, a *multi-label* classifier C learns to represent the total set of tool labels $T \in \{t_j\}$ where $j = 1, 2, ..., M$. For a testing video stream, C must produce as output a set $Z_i \subseteq T$ with predicted tool labels for the i-th video frame. Note that, this generalization results in 2^M potential combinations , which are termed as label sets. In a general setting, this might result in many practical constraints (e.g. memory for storing, representation and performing actual classification). However due to the practicalities of endoscopic intervention, where only a limited number of tools (maximum three for this particular dataset) and combinations (see Figure 1 and 2) can be present at once, the *multi-label* problem remains tractable. In particular, the seven tool *M2CAI16 tool detection* dataset has resulted in approximately twenty *label-sets* spanning from order zero to three.

3.2 Imbalance Quantification

Even though imbalance in *binary* and multi class classification is a well-studied problem, quantitative analysis of imbalance in *multi-label* dataset is proposed in very few occasions[11;12]. Conventional imbalance analysis methods, designed for *binary/multi-class* classification, assume only the ratio of majority to minority class labels as imbalance measure, therefore, not suitable for *multi-label* datasets. There exist some traditional metrics notably label cardinality and label density which characterize *multi-label* datasets. Label cardinality is the average number of active labels per sample and label density denotes average of label cardinality over the total number of labels. Mathematically, these measures can be defined as follows[11]:

$$Cardinality(F) = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} t_{ij} \tag{1}$$

$$Density(F) = \frac{Cardinality(F)}{M} \tag{2}$$

With the presence of multiple *label-sets*, special metrics are needed for analyzing imbalance in *multi-label* datasets in detail. In this work, we have exploited imbalance ratio per label (IRLbl), the mean imbalance ratio (MeanIR) and the coefficient of variation of IRLbl (CVIR) introduced in[11] to analyze the imbalance in our tool dataset:

$$MeanIR = \frac{1}{M}\sum_{i=1}^{M} IRLbl(t_i) \tag{3}$$

$$CVIR = MeanIR\sqrt{\frac{\sum_{i=1}^{M}(IRLbl(t_i) - MeanIR)^2}{M-1}} \tag{4}$$

Here IR per label, $IRLbl(T_i)$ is calculated as the ratio between the majority (most frequent) label and label T_i. As a result, the majority label will always have $IRLbl = 1$ and rest of the labels will have higher $IRLbl > 1$. MeanIR computes the average level of imbalance of the dataset while CVIR measures variation of IRLbl i.e. similarity of level of imbalance between all labels. For a perfectly balanced dataset, all IRLbl values would be 1, which results in values of MeanIR and CVIR being 1 and 0 respectively. The joint use of MeanIR and CVIR with values greater than 1 and 0 respectively denote the level of imbalance in a *multi-label* dataset. Moreover, the values of IRLbl greater than 1 can be used for measuring individual label imbalance.

3.3 Stratification

Stratification is the process of sampling, where the proportion of disjoint groups is maintained [12]. Stratification in the *multi-label* data context is a challenging task. Improper stratification might significantly reduce the performance of classifiers as demonstrated by Sechidis et al. [12].

The most intuitive stratification in the tool detection setting would be to consider a *balanced* strategy. In this setting, the occurrence frequency of the least frequent tool would be considered as the desired sample size and the rest of the tools would be sampled accordingly. However, co-occurrence of tools in different frames actually results in an unbalanced training and validation set.

A better way of handling the problem would be to consider stratification on *label-sets*. Here, the frequency of *label-sets* is considered for sampling of image frames. A stratification threshold ψ is applied, where for more frequent *label-sets*, ψ occurrences are sampled randomly, and for less frequent ($< \psi$) *label-sets*, all samples are considered for training.

3.4 ZIBNet

Our proposed CNN architecture (Fig. 3) is composed of three main parts:

- *input layer* : which accepts an input image of size 384x256 pixels.
- *convolutional layers* : which are similar to AlexNet architecture.
- *fully connected layers* : which are specific to tool detection task with size 512, 64 and 8 respectively.

Before the learning step, the convolutional layers weights are initialized with AlexNet convolutional weights and the fully connected layers are initialized with random weights. The rectify units are applied to the output of convolutional and fully connected layers except the last fully connected layer which is connected to sigmoid non-linearity. Since our network contains pre-trained convolutional weights of AlexNet which are generic for earlier convolution layers and become specific to ImageNet objects for higher layers, we assign layer specific learning rates

$$LR = c_i \eta \tag{5}$$

where η is the learning rate and $c_i = [0.1, 0.2, 0.3, 0.4, 0.5, 1.0, 1.0, 1.0]$ is the learning coefficient for each layer. The learning coefficient becomes higher with

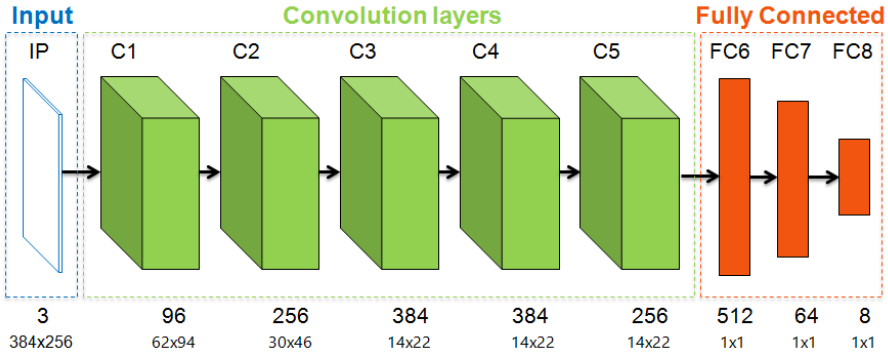

Fig. 3 Proposed CNN architecture. The input layer (blue) size and fully connected layers (orange) are adapted to tool detection task while the convolution layers are similar to the AlexNet [20] architecture.

subsequent convolutional layer, except for the fully connected layers whose coefficients remain fixed since these layers have been randomly initialized. The output layer contains 8 units for seven tools and an additional 'No Tool' label, which is added to the ground truth and represents that none of the given tools are present in the image i.e. background.

During the learning step, the network minimizes the joint label distribution through a *uni-variate* loss function $\mathcal{L}$ defined as:

$$\mathcal{L} = -\sum_{t=1}^{T} w(t)[z_t \log y_t + (1 - z_t) \log(1 - y_t)] \tag{6}$$

where $w(t)$ is a weighting function that normalizes loss in terms of output z_t, and y_t is prediction for tool t. Intuitively, even after *label-set* stratification, imbalance on the tool label would not be omitted completely. $\mathcal{L}$ is formulated as a *uni-variate* loss function where cross entropy is weighted with tool occurrence frequency in the training data, to manage the residual imbalance after stratification.

To avoid over-fitting, we perform real-time data augmentation (flipping, mirroring and cropping) during learning, apply dropout units for 'FC6' and 'FC7' layers and use a weight decay of (10^{-5}) for every layer. Finally the network is trained using stochastic gradient descent (η of (10^{-2})) with momentum of 0.9 using the holdout scheme.

3.5 Temporal Smoothing

Due to the stochastic nature of the classification process, false detections would occur during the testing step. For reducing such false detections, we have adopted a temporal smoothing (TS) approach as an online post-processing step. It assumes that each tool transition within the endoscopic videos is smooth and takes previous frame detections into account in a weighted scheme. A window of five frames (including current and four previous frames) with normalized linear weights determines the current output detection. Mathematically, TS is defined as follows:

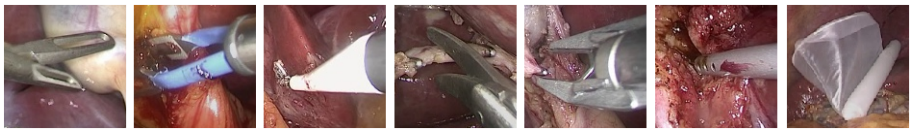

Fig. 4 Appearances of different surgical tools from *M2CAI16 tool detection dataset*. The tools used during the surgical procedure (left to right) are grasper, bipolar, hook, clipper, scissors, irrigator and specimen bag.

$$y_{ts} = \frac{\sum_{i=0}^{t} w_i y_{t-i}}{\sum_{i=0}^{t} w_i} \tag{7}$$

where y_{ts} is the temporally smooth output for the current time step, $t = 4$ and $w_i = [1.0, 0.8, 0.6, 0.4, 0.2]$ is the weight co-efficient for each time step.

4 Results

This section provides a quantitative analysis of the proposed method, as well as quantitative comparison w.r.t. state-of-the-art methods, to demonstrate its effectiveness for surgical tool detection from endoscopic video streams.

4.1 Data Preparation

All our quantitative experiments were performed on the *M2CAI Tool Detection Challenge 2016* training and testing dataset [16]. The dataset contains 15 clinical cholecystectomy surgical procedures (one video per surgical procedure), which were performed using seven surgical tools (as shown in Figure 4). The dataset is divided into 10 training and 5 testing videos. The videos were captured at 25 frames per second, but the ground truth (GT) annotation was done at 1 frames per second. A tool was considered to be present only if at least half of the tool tip was visible.

To provide a fair comparison, we only considered average precision (AP) per tool and mean average precision (mAP) for all tools [2] as the comparison metric. For all the experiments, *M2CAI Tool Detection Training* dataset was used for training and the results were reported on *M2CAI Tool Detection Testing* dataset with stratification threshold $\psi = 300$.

4.2 Comparison with state-of-the-art methods

The proposed framework (*label-set* based stratification and weighted *uni-variate* loss for training ZIBNet followed by TS) results in state-of-the-art performance when relying on mAP for evaluation. We compared the results of our proposed method with the top performing methods from *M2CAI Tool Detection Challenge*, all of which [16;17;18] relied on CNN for tool detection. All the results are taken from *M2CAI Tool Detection Challenge* website where the respective authors used their own software for training and testing. As Table 1 shows, our proposed method

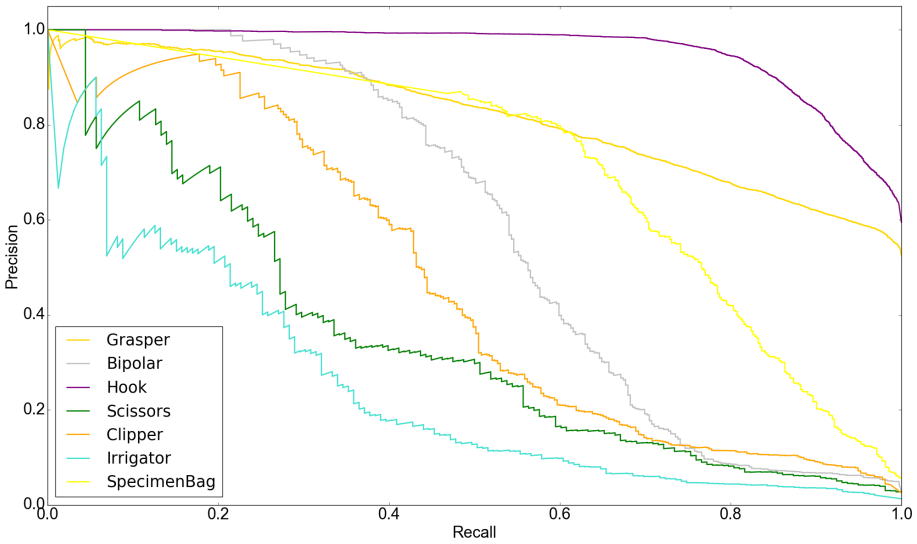

Fig. 5 Receiver operating characteristic (ROC) curve for presence detection of each tool.

Table 1 Comparison of meanAP for state-of-the-art techniques on *M2CAI Tool Detection Testing* dataset.

	ToolNet [16]	Sahu et al. [17]	Raju et al. [18]	Proposed
MeanAP	0.52	0.53	0.63	**0.65**

outperformed all other methods. We attribute the superior performance of our proposed method to our particular formulation of the problem as a *multi-label* one and our novel stratification technique.

For a detailed performance evaluation of our proposed framework, a Receiver Operating Characteristic (ROC) curve for each tool is shown in Figure 5. Apart from *Irrigator* and *Scissors*, all the other tools performed sufficiently well. Difficulties in detecting *Scissors* is already known [2;16], however, the reason behind low performance on *Irrigator* is analyzed in the remainder of this work (Section 5).

4.3 Imbalance Analysis

We have performed a detailed imbalance analysis of all the tools present in the training and testing dataset. Rather than reporting the exact number of occurrences for each tool, we concentrated on the IRLbl measure for each tool (Table 2). As described in Section 3.2 the value of the most frequent tool is 1 and rest have higher values (ideally 1, higher IRLbl means higher imbalance for the associated tool). As we can see in Table 2, the most frequent tool is *Hook* with IRLbl 1.0. It is interesting to note that *Irrigator* is almost *three times* more frequent in Training dataset compared to Testing dataset. The effects of imbalance is further reported through other metrics in Table 2 and 3.

Traditionally, for *one-vs-all* or *multi-class* classification, imbalance is reported over the whole dataset using Cardinality and Density, as shown in Equations (1)

Table 2 Imbalance ratio per label (IRLbl) of various tools in the training and testing dataset.

Tools	Training	Testing
Grasper	1.28	1.13
Bipolar	22.25	20.76
Hook	1.00	1.00
Scissors	34.37	47.17
Clipper	16.09	23.66
Irrigator	14.82	46.88
Specimen Bag	9.39	15.43

Table 3 Values of traditional metrics (label cardinality and label density) and specific metrics (MeanIR and CVIR) over the whole training and testing dataset.

Measures	Training	Testing
Cardinality	1.26	1.23
Density	0.18	0.17
MeanIR	14.17	22.29
CVIR	0.83	0.85

and (2). However, it is evident from the Equations (3) and (4) that overall imbalance of a *multi-label* dataset can be appreciated only by looking at MeanIR and CVIR together. Imbalance in the training dataset resulted in MeanIR value of approx. 14 times higher than ideal (ideally 1) with 83% variance (ideally 0%) in IRLbl values as reported in Table 3.

4.4 Analysis of Stratification techniques

We performed quantitative comparison of different sampling approaches to highlight the importance of stratification on the performance of ZIBNet for tool detection. *Label-sets* based stratification (Section 3.3) is compared to unbalanced sampling and tool-level balanced stratification (Section 3.3). Average precision for each tool and overall mAP is reported in Table 4. In particular, the baseline strategy (no stratification at all) - termed as 'Unbalanced' in Table 4, performed worst. Tool-level balanced stratification resulted in an overall increase of 3% over 'Unbalanced' whereas the proposed *label-set* stratification increased mAP by 9%, as shown in Table 4. It is worth noting that Unbalanced approach favored most frequent tools. Stratification, on the other hand, adapted the network to alleviate this bias and enhanced detection of less frequent tools.

4.5 Analysis of Temporal Smoothing

The last part of our design is to apply TS to decrease stand-alone activations as described in Section 3.5. We have run experiments to quantitatively demonstrate the importance of TS. TS is independent of the rest of the proposed framework and as shown in Table 5, consistently improved the performance of both stratification techniques by reducing false positives. In particular, the simple stratification benefited more from TS with an overall mAp increase of 8%, whereas 6% boost in performance was observed for *label-set* stratification.

Table 4 Quantitative comparison between different imbalance handling (stratification) strategies: no sampling (Unbalanced), tool-level sampling (Tool-Balanced) and label-set based sampling (Label-set) for various tools.

Tools	Unbalanced	Tool-Balanced	Label-set
Grasper	0.87	0.61	0.81
Bipolar	0.21	0.46	0.57
Hook	0.96	0.95	0.96
Scissors	0.09	0.3	0.34
Clipper	0.34	0.51	0.46
Irrigator	0.21	0.21	0.24
Specimen Bag	0.81	0.64	0.71
MeanAP	0.50	0.53	0.59

Table 5 Importance of Temporal Smoothing (TS) on overall tool detection accuracy for different tools over label based sampling (Balanced with TS) and *label-set* based sampling (Label-set with TS).

Tools	Balanced with TS	Label-set with TS
Grasper	0.62	0.83
Bipolar	0.58	0.66
Hook	0.97	0.97
Scissors	0.47	0.51
Clipper	0.65	0.54
Irrigator	0.21	0.29
Specimen Bag	0.73	0.77
MeanAP	0.61	0.65

5 Discussion and Conclusion

Detection of surgical tools in endoscopic video is an important problem requiring a rigorous understanding of the data as well as an effective handling approach. A successful surgical tool detection technique can potentially improve a multitude of CAI applications. For example, detection of surgical workflow phases can directly benefit from tool detection results. However, tool co-occurrences, change of illumination, specular reflection and partial occlusion make the detection problem significantly more difficult.

This work clearly showed that generalizing the classification problem with domain adaptation can significantly improve classification results. Our proposed method demonstrates that fully automatic tool detection results in an acceptable level of agreement with the manual annotations. By modeling tool co-occurrences as *label-sets*, we can better handle the inherent structure of surgical tool presence during interventions. Moreover, a detailed study of imbalance in *label-sets* have motivated us to develop stratification methods for CNN training.

Note that, IRLbl in Table 2 suggests that *irrigator* has significantly different imbalance ratio between training and testing dataset. Not only our results consistently lead to lowest AP as reported in Table 4 and 5, ToolNet results by Twinanda et al.[16] also reported the same. This suggests that along appearance difficulties (in case of *scissors*), label imbalance significantly challenges the performance of CNN.

An important observation of our study is the boost in performance by addition of temporal smoothing as an online post-processing method (no future information

is considered). TS consistently enhance detection results for both stratification approaches as reported in Table 5.

This study solely concentrated on tool presence identification, however, future studies of surgical workflow phase recognition can benefit from the insights. In particular, the co-relation of surgical phases with the tools being used therein can be exploited further in the *label-set* setting. The imbalance of *label-sets* also suggests special tool co-occurrences which could be used as important phase-transition cues.

In conclusion, this study motivates us to rethink about the standard assumptions regarding surgical tool presence detection. Deviating from de-facto supervised *one-vs-all* or *multi-class* techniques (the performance of which heavily depends on the co-occurrence frequencies) towards *multi-label* settings can provide multiple benefits. Finally such fully automatic techniques are expected to be instrumental in advancing the computer assistance during surgical intervention.

Compliance with Ethical Standards

Disclosure of potential conflicts of Interest

Funding: This study was funded by German Federal Ministry of Education and Research (BMBF) under the project BIOPASS (grant number - 16 5V 7257).
Conflict of Interest: The authors declare that they have no conflict of interest.

Research involving Human Participants and/or Animals

This article does not contain any studies with human participants or animals performed by any of the authors.

Informed consent

This article contains patient data from a publically available dataset.

References

1. Donaldson MS, Corrigan JM, Kohn LT (2000) To err is human: building a safer health system, vol 6. National Academies Press
2. Twinanda AP, Shehata S, Mutter D, Marescaux J, de Mathelin M, Padoy N (2016) Endonet: A deep architecture for recognition tasks on laparoscopic videos. arXiv preprint arXiv:160203012
3. Blum T, Feußner H, Navab N (2010) Modeling and segmentation of surgical workflow from laparoscopic video. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, pp 400–407
4. Allan M, Chang PL, Ourselin S, Hawkes DJ, Sridhar A, Kelly J, Stoyanov D (2015) Image based surgical instrument pose estimation with multi-class labelling and optical flow. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, pp 331–338
5. Padoy N, Blum T, Ahmadi SA, Feussner H, Berger MO, Navab N (2012) Statistical modeling and recognition of surgical workflow. Medical image analysis 16(3):632–641

6. Bouget D, Benenson R, Omran M, Riffaud L, Schiele B, Jannin P (2015) Detecting surgical tools by modelling local appearance and global shape. IEEE transactions on medical imaging 34(12):2603–2617
7. Zappella L, Béjar B, Hager G, Vidal R (2013) Surgical gesture classification from video and kinematic data. Medical image analysis 17(7):732–745
8. Voros S, Long JA, Cinquin P (2007) Automatic detection of instruments in laparoscopic images: A first step towards high-level command of robotic endoscopic holders. The International Journal of Robotics Research 26(11-12):1173–1190
9. Sahu M, Moerman D, Mewes P, Mountney P, Rose G (2016) Instrument state recognition and tracking for effective control of robotized laparoscopic systems. International Journal of Mechanical Engineering and Robotics Research 5(1):33
10. Gu Z, Gu L, Eils R, Schlesner M, Brors B (2014) circlize implements and enhances circular visualization in r. Bioinformatics p btu393
11. Charte F, Rivera AJ, del Jesus MJ, Herrera F (2015) Addressing imbalance in multilabel classification: Measures and random resampling algorithms. Neurocomputing 163:3–16
12. Sechidis K, Tsoumakas G, Vlahavas I (2011) On the stratification of multi-label data. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, pp 145–158
13. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, pp 1097–1105
14. Sznitman R, Becker C, Fua P (2014) Fast part-based classification for instrument detection in minimally invasive surgery. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, pp 692–699
15. Speidel S, Benzko J, Krappe S, Sudra G, Azad P, Müller-Stich BP, Gutt C, Dillmann R (2009) Automatic classification of minimally invasive instruments based on endoscopic image sequences. In: SPIE medical imaging, International Society for Optics and Photonics, pp 72,610A–72,610A
16. Twinanda AP, Mutter D, Marescaux J, de Mathelin M, Padoy N (2016) Single- and multi-task architectures for tool presence detection challenge at m2cai 2016. `arXiv:1610.08851`
17. Sahu M, Mukhopadhyay A, Szengel A, Zachow S (2016) Tool and phase recognition using contextual cnn features. `arXiv:1610.08854`
18. Raju A, Wang S, Huang J (2016) M2cai surgical tool detection challenge report. `http://camma.u-strasbg.fr/m2cai2016/reports/Raju-Tool.pdf`
19. Lex A, Gehlenborg N, Strobelt H, Vuillemot R, Pfister H (2014) Upset: visualization of intersecting sets. IEEE transactions on visualization and computer graphics 20(12):1983–1992
20. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, pp 1097–1105