SEGCROP: SEGMENTATION-BASED DYNAMIC CROPPING OF ENDOSCOPIC VIDEOS TO ADDRESS LABEL LEAKAGE IN SURGICAL TOOL DETECTION

Adnan Qayyum[1,2], *Muhammad Bilal*[3,*], *Junaid Qadir*[4], *Massimo Caputo*[5], *Hunaid Vohra*[5], *Taofeek Akinosho*[3], *Ilhem Berrou*[3], *Faatihah Niyi-Odumosu*[3], *Michael Loizou*[3], *Anuoluwapo Ajayi*[3], and *Sofiat Abioye*[3]

[1]University of Glasgow, Glasgow, United Kingdom
[2]Information Technology University, Lahore, Pakistan
[3]University of the West of England, Bristol, England
[4]Qatar University, Doha, Qatar
[5]Bristol Heart Institute, University of Bristol, Bristol, England

ABSTRACT

In recent times, surgical data science has emerged as an important research discipline in interventional healthcare. There are many potential applications for analysing endoscopic surgical videos using machine learning (ML) techniques such as surgical tool classification, action recognition, and tissue segmentation. However, the efficacy of ML algorithms to learn robust features drastically deteriorates when models are trained on noise-affected data [1]. Appropriate data preprocessing for endoscopic videos is thus crucial to ensure robust ML training. To this end, we demonstrate the presence of label leakage when surgical tool classification is performed naively and present *SegCrop*, a dynamic U-Net model with an integrated attention mechanism to dynamically crop the arbitrary field of view (FoV) in endoscopic surgical videos to remove spurious label-related information from the data. In addition, we leverage explainability techniques to demonstrate how the presence of spurious correlations influences the model's learning capability.

Index Terms— Surgical Data Science, Surgical Tool Detection, Image Segmentation, Robust ML, Explainable AI

1. INTRODUCTION

Over the past few years, surgical data science has received increasing attention from the research community. Various vision-related tasks have been modelled using endoscopic surgical videos, such as surgical task detection [2] and surgical tool detection and tracking [3]. Particularly, surgical tool detection is a critical problem that can be used to model high-level image semantics and dynamics for applications such as surgical skills assessment. For such vision-oriented tasks, deep learning (DL) algorithms now represent the state of the art due to their superior performance compared to classical image processing techniques.

Corresponding Author: muhammad.bilal@uwe.ac.uk

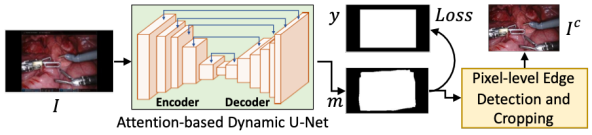

Fig. 1. Our proposed method for dynamic cropping of endoscopic surgical images using attention-based U-Net.

Endoscopic surgical videos are afflicted with numerous problems such as data imbalance (that could lead to the development of a biased model) and different noise types that could result in spurious correlations being learned rather than robust features. A key issue that significantly misleads the training process is the presence of *leaky labels*. In label leakage, input data used for training models also contain target labels directly or indirectly, which the learning algorithm may use as a shortcut [4]. Consequently, the model ends up picking spurious correlations instead of robust features, which ultimately results in impressive models (in terms of accuracy), however, they are not robust. It is therefore critical to thoroughly analyse surgical ML models to confirm that they are learning robust features before clinical deployment.

In this paper, we attempt to empirically investigate the effect of spurious correlations that are caused by label leakage when a DL-based surgical tool detection system is developed. Our proposed method is illustrated in Figure 1. We used the dataset of a recent Medical Image Computing and Computer-Assisted Intervention (MICCAI) challenge on surgical tool detection and localisation. The names of surgical tools are included in the lower part of the videos (shown on the robotic dashboard for guiding surgeons), which can mislead the model learning. To this end, we empirically demonstrate this leaked tool information significantly influences the learning capabilities of our multi-class ConvNext classifier (a famous vision model [5]). We took corrective measures to reduce the field of view (FoV) by dynamically cropping regions

leaking the tool names using a U-Net-based data preprocessing strategy.

In summary, the main contributions of this paper are:

1. We leverage IntegratedGrad and GradCAM for interpretability to demonstrate the effect of label leakage on the model trained for surgical tool detection.
2. We present a dataset of 24,694 images along with corresponding human-generated masks.
3. We present the use of a dynamic U-Net model with an attention mechanism for dynamically reducing FoV in surgical endoscopic videos.

2. METHODOLOGY

We have formulated the dynamic cropping of the FoV as a segmentation problem. The intuition is based on the need to crop each video with a different size depending on the location where the tool information appears within the video. This can be a time-consuming and labour-intensive task if done manually. Similarly, it is crucial to ensure efficiency in a system developed to eradicate this form of label leakage. For a given image I with a spatial resolution of $m \times n$, where m and n denote the number of rows and columns, our objective is to crop the FoV in I to a variable dimension of i and j such that redundant tool information gets completely truncated, thus $I(i,j) \in I(m,n)$. Note that the tool information could appear at either the top or bottom of a video with different sizes (see Figure 2). In addition, redundant image pixels (i.e., black borders) can also result in wasted computations while not contributing to the model's learning. Therefore, the development of an automatic mechanism to learn dynamic cropping is desirable to reduce human effort, time, and cost.

2.1. Data Description

We used the dataset provided by MICCAI 2022 challenge on surgical tool detection in endoscopic videos.[1] It contains surgical videos acquired at a frame rate of 60fps and annotations for 14 different surgical tools. This dataset also has significant noise in training labels that is introduced by extracting tool information from robot system data directly. We extracted and manually created segmentation masks for selective frames from each video using CVAT (a widely adopted annotation tool for images). As a result, we created $24,694$ mask images required for training a generative model to dynamically crop the robotic UI and borders from unseen videos. Therefore, we have a paired data $\mathcal{D}(x,y)$, where x is the image to be cropped and y is the corresponding ground truth mask.

2.2. Proposed SegCrop Method for Dynamic Cropping

We trained a dynamic cropping model using a variant of the U-Net model that contains an attention mechanism and our paired data $\mathcal{D}$. In the case of image segmentation, the attention mechanism highlights the relevant information during model training, thus forcing the U-Net model to learn better features while reducing computations and increasing the model's generalizability [6]. The generative network $\mathcal{G}$ learns the mapping between the input x and ground truth y to classify the content of the image into the foreground (i.e., FoV) and background (i.e., redundant to be eliminated) and generates the candidate mask m to crop the image. The algorithm to dynamically crop the endoscopic videos using a predicted mask from the network is described in Algorithm 1.

[1] https://surgtoolloc.grand-challenge.org/

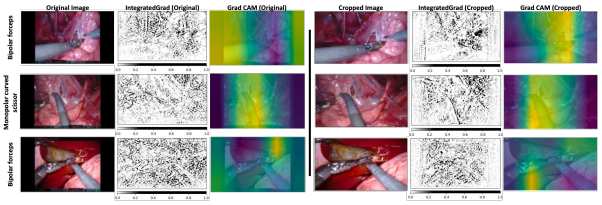

Fig. 2. Examples images requiring dynamic cropping (tool panel width varies across images). *Note that a few images have a top banner while others do not (cf. see second row, first image). Also, black borders also incur wasted computation.*

Algorithm 1 SegCrop algorithm for dynamic cropping of endoscopic images.

Input: Generative model $\mathcal{G}$, paired data $\mathcal{D}\{I(m,n), y(i,j)\}$, epochs ϵ, bunch size $\mathcal{B}$, learning rate η, batch size β
Output: $I^c(i,j)$
Initialize: ϵ, $\mathcal{B}$, η, β
for $e = 1, ..., \epsilon$ **do**
  train $\mathcal{G}$ using $\mathcal{D}\{I(m,n), y(i,j)\}$ and β
end
Return: Trained model $\mathcal{G}_{trained}$
Create: Bunches of $\mathcal{D}$ each having size $\mathcal{B}$ and return set $B_n = \{b1, b2, b3, ..., b_n\}$
for $b = b1, ..., B_n$ **do**
  Get: generated mask m using $\mathcal{G}_{trained}$
  Perform: Pixel-level boundary detection using m
  Return: Coordinates (x_1, y_1) and (x_2, y_2) for detected boundary
  Perform: Cropping of images in bunch b using the (x_1, y_1) and (x_2, y_2)
  Return: Cropped image $I^c(i,j)$
end

Segmentation Model's Architecture: We have used an hourglass U-Net architecture with skip connections as shown in Figure 1. To augment the capabilities of the network we integrated an attention mechanism within the U-Net architecture. The network has a total of 6 layers in the encoder part and 14 layers in the decoder part. There are three 1D convolutions layers in the bottleneck. We employed batch normalization and ReLU activation functions in the encoder and decoder, whereas ReLU is only used after the last 1D convolutional layer in the bottleneck. We used ResNet34 as the backbone network in the segmentation model that learns

the classification of foreground and background. The network has approximately 40M parameters out of which 20M are trainable and 21M are non-trainable parameters.

Evaluation Strategy: To empirically evaluate the effect of label leakage due to the presence of tool information at the bottom of the videos, we trained a famous ConvNext model for surgical tool classification. We then demonstrate the label leakage using two widely adopted interpretability techniques known as gradient class activation map (GradCAM) and IntegratedGrad. The results are described in the next section.

3. EXPERIMENTS AND RESULTS

3.1. Experimental Setup

We partitioned the data into training and testing sets using a split of 80% and 20%, respectively. To augment the efficacy of the segmentation network, we employed data augmentation using the squish method. The best learning rate (LR) for each model was selected using the LR scheduler that scales the magnitude of weight updates to minimize the network's loss function. The training progresses slowly if the LR is too low since small updates are made to model weights. It can cause undesirable divergent behaviour if it is too high. We employed a cyclic LR for training the models as explained in [7], where stochastic gradient descent with warm restarts is used to combine an aggressive annealing schedule with periodic "restarts" to create an LR schedule. This approach significantly improved the model's learning abilities. We adopted a similar strategy to choose the optimal LR to train the surgical tool detection models to evaluate the label leakage phenomena. To ensure fair comparison we train the model with original data and the model using cropped data using an LR of 1e-4. Each model was trained for a maximum of 12 epochs. We used a batch size of 32 for the training segmentation model and a batch size of 64 for training tool detection models (i.e., models using original and cropped data). All implementation was performed using the `fastai` ML library [8].

3.2. Results and Discussions

Dynamic Cropping: Proposed SegCrop provided an average accuracy of 99.29% on the validation data. In addition, the SegCrop outperformed ground-truth human-generated segmentation masks while generating candidate masks for cropping (Figure 3). The figure demonstrates that the segmentation network efficiently adapts to the semantics of the input image while dynamically cropping the images (retaining the FoV). Finally, we employ pixel-level edge detection to refine the generated segmentation masks.

Evaluating the Effect of Cropping on Tool Classification: As described above, we evaluated the effect of label leakage for surgical tool detection in endoscopic videos. Specifically, we trained the classification model on both the uncropped and cropped data after performing cropping using

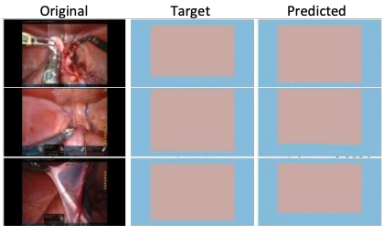

Fig. 3. Candidate masks generated by the dynamic U-Net that are further refined at the pixel level. *Model effectively learns to predict FoV boundary while surpassing target masks.*

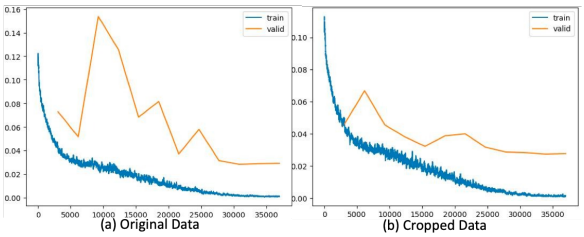

Fig. 4. Depiction of tool detection model training using original data (a) and using cropped data (b).

the SegCrop method. The learning curves for both models are shown in Figure 4. It is evident that the model shows a fluctuating behaviour on validation loss for the original uncropped data (Figure 4 (a)), while comparatively stable and smoother curve on cropped data (Figure 4 (b)). Despite this, surprisingly, we have got comparatively similar average performance in terms of accuracy and F1-score for both models. The model trained on original data has an average accuracy and F1-score of 99.57% and 99.4%, respectively. Likewise, the model trained on the cropped data provided an average accuracy and F1-score of 99.56% and 95.45%, respectively. Therefore, we decided to deeply investigate the tool detection model to identify whether or not label leakage in endoscopic frames is influencing the model.

Using Explainability to Interpret Models' Predictions: We used two explainable ML methods (i.e., GradCAM and IntegratedGrad) to interpret the model decision for surgical tool classification (Figure 5). These methods identified regions (i.e., pixels) used by the model to inform the prediction. Figure 5 provides a comparison of surgical tool detection for a single tool label with the ConvNext model trained on original and cropped data. It also depicts the focused regions of the two models informing their predictions. It is evident from Figure 5 that the label leakage in the original images (without cropping) influences the model to focus on more pixels that do not constitute or are closer to the tool under consideration. Whereas, we see that when the ConvNext model is trained on cropped images using SegCrop is effectively able to see relevant regions for the desired tool (see the darker points in IntegratdGrad plots and dark yellow regions in GradCAM

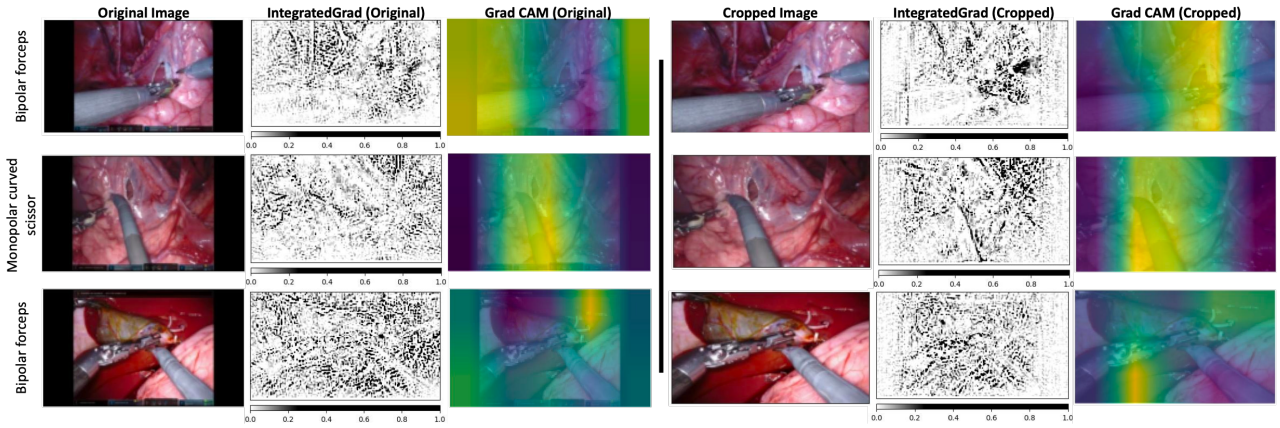

Fig. 5. Demonstrating the effect label leakage using IntegratedGrad (IG) and GradCAM (GC) for ConvNext trained using original data (left) and cropped data (right). *The darker pixels in IG and yellow in GC plots shows regions used by the model for predicting tools. Note label leakage in IG plots (left), where the entire image is being focused (as darker pixels are spread across the whole image instead of the tool under consideration).*

plots highlight the most focused pixels during the model's prediction). It is worth noting that underlying data has enormous label noise. The tool classifier is trained for 14 classes while only three tools can be present in a specific video. Many videos are said to have three tools from provided labels but in actuality, only fewer than three tools are present. This noise gets translated to individual video frames when labels are extrapolated from videos to frame level. As we are only showing the interpretability plots for one class (so they are not only focusing on the specific regions) and label noise might influence interpretability in some cases.

4. CONCLUSIONS

We present an empirical approach for investigating a model to explain features informing its predictions. In our endoscopic tools detection task, the model is revealed to use regions in videos that leak tool information. Such models certainly fail in production. We overcome label leakage by a dynamic U-Net model for cropping arbitrary pixels. The U-Net model created cropped images at a human-level performance we are used for training the multi-class tools detection model. The revised tool classification model started to focus on more relevant regions to make predictions. This approach not only tackles data leakage but also avoids unnecessary computation (wasted in processing irrelevant pixels). In the future, we will extend our research toward creating a systematic methodology for tackling noisy labels from endoscopic videos.

5. REFERENCES

[1] Adnan Qayyum, Junaid Qadir, Muhammad Bilal, and Ala Al-Fuqaha, "Secure and robust machine learning for healthcare: A survey," *IEEE Reviews in Biomedical Engineering*, vol. 14, pp. 156–180, 2020.

[2] Andru P Twinanda, Sherif Shehata, Didier Mutter, Jacques Marescaux, Michel De Mathelin, and Nicolas Padoy, "Endonet: A deep architecture for recognition tasks on laparoscopic videos," *IEEE transactions on medical imaging*, vol. 36, no. 1, pp. 86–97, 2016.

[3] Lena Maier-Hein, Swaroop S Vedula, Stefanie Speidel, Nassir Navab, Ron Kikinis, Adrian Park, Matthias Eisenmann, Hubertus Feussner, Germain Forestier, Stamatia Giannarou, et al., "Surgical data science for next-generation interventions," *Nature Biomedical Engineering*, vol. 1, no. 9, pp. 691–696, 2017.

[4] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann, "Shortcut learning in deep neural networks," *Nature Machine Intelligence*, vol. 2, no. 11, pp. 665–673, 2020.

[5] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie, "A convnet for the 2020s," in *Proceedings of the IEEE CVPR*, 2022, pp. 11976–11986.

[6] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al., "Attention u-net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999*, 2018.

[7] Leslie N Smith, "A disciplined approach to neural network hyper-parameters: Part 1–learning rate, batch size, momentum, and weight decay," *arXiv preprint*, 2018.

[8] Jeremy Howard and Sylvain Gugger, "Fastai: A layered API for deep learning," *Information*, vol. 11, no. 2, pp. 108, 2020.