Detection of critical structures in laparoscopic cholecystectomy using label relaxation and self-supervision

David Owen[1], Maria Grammatikopoulou[1], Imanol Luengo[1], and Danail Stoyanov[1,2]

[1] Digital Surgery, a Medtronic company, London, UK
[2] Wellcome / EPSRC Centre for Interventional and Surgical Sciences, University College London, London, UK
david.owen@medtronic.com

Abstract. Laparoscopic cholecystectomy can be subject to complications such as bile duct injury, which can seriously harm the patient or even result in death. Computer-assisted interventions have the potential to prevent such complications by highlighting the critical structures (cystic duct and cystic artery) during surgery, helping the surgeon establish the Critical View of Safety and avoid structure misidentification.
A method is presented to detect the critical structures, using state of the art computer vision techniques. The proposed label relaxation dramatically improves performance for segmenting critical structures, which have ambiguous extent and highly variable ground truth labels. We also demonstrate how pseudo-label self-supervision allows further detection improvement using unlabelled data.
The system was trained using a dataset of 3,050 labelled and 3,682 unlabelled laparoscopic cholecystectomy frames. We achieved an IoU of .65 and presence detection F1 score of .75. The model's outputs were further evaluated qualitatively by three expert surgeons, providing preliminary confirmation of our method's benefits.
This work is among the first to perform detection of critical anatomy during laparoscopic cholecystectomy, and demonstrates the great promise of computer-assisted intervention to improve surgical safety and workflow.

Keywords: surgical video · anatomy detection · self-supervised learning

1 Introduction

Laparoscopic cholecystectomy is a common surgery in which the gallbladder is removed. This involves exposing the critical structures (cystic duct and artery), clipping and dividing them, then extracting the gallbladder [7]. Complications can occur when the structures are misidentified or confused with the common bile duct, particularly as they may be difficult to distinguish without thorough dissection. Official guidance has encouraged that surgeons establish "Critical View of Safety" (CVS) before clipping and division [10]. In CVS, both structures can clearly and separately be identified, and traced as they enter the gallbladder.

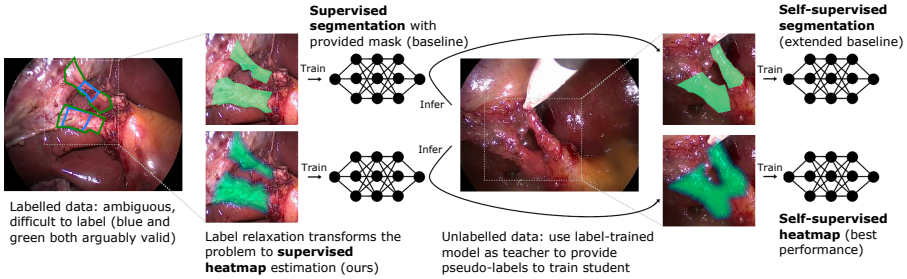


Fig. 1: Overview of our methods. Segmentation is challenged by ground truth structures with ambiguous extent (left). Label relaxation transforms the problem to heatmap estimation with down-weighting of ambiguous regions, which improves detection performance. Self-supervision feeds in unlabelled data via pseudo-labels, allowing further improvement.

Computer assistance in achieving CVS has great potential to improve surgical safety and workflow, but has only recently become possible due to advances in computer vision [7]. Namazi *et al* demonstrated a proof of principle approach using binary CVS classification [9]. Tokuyasu *et al* developed a bounding box detection system, focused on anatomical landmarks that included the common bile duct and cystic duct but not the cystic artery [12]. Most recently, Mascagni *et al* used joint segmentation of the hepatobiliary anatomy and classification of CVS [8], arguably combining the best aspects of prior work. Our work differs by focusing on the critical structures directly, as these are the structures that surgeons must identify and divide. This may be beneficial for guiding surgical workflow, providing visual cues that can help achieve CVS.

We present a novel method for detecting critical structures that outperforms conventional segmentation, using label relaxation (Section 2.1) to better handle challenging ground truth labels in images where structures are ambiguous. Subsequently, we incorporated pseudo-label self-supervision (Section 2.2), using unlabelled data to further improve performance. We trained and evaluated these methods using 3,050 labelled images from 75 videos and 3,682 unlabelled images from 90 videos for self-supervision. Finally, we gathered feedback from three experienced surgeons (Section 3.4), comparing different methods and confirming our method can improve clinical significance when detecting critical structures.

2 Methods

2.1 Critical structures identification via label relaxation

Our objective was to label the critical structures, here treated as a single foreground class, with the rest of the image considered as background. This is naturally posed as a binary segmentation problem. Standard segmentation approaches struggled to perform well in this task, because of the ambiguous and

subjective nature of critical structures annotation (see Figure 1). This problem was exacerbated by the use of conventional one-hot encoding: a given pixel is assigned as either 100% structure or 100% background class. This impairs generalisation, and led the model to struggle with false negatives.

To overcome this, we developed a technique inspired by related work in surgical tool detection [4]. Rather than segmentation, we trained a network for heatmap regression, where the ground truth heatmap is derived from the original annotations' Euclidean distance transforms.

Given a binary segmentation ground truth, x_k for structure k, we defined the relaxed label as $x_k^{'} = 1 - \exp \frac{-\operatorname{edt}(x_k \oplus t)}{d}$, where $\operatorname{edt}(\cdot)$ is the Euclidean distance transform, $\oplus t$ represents dilation with a square of t pixels and d is a parameter to control the relaxation. Each $x_k^{'}$ is then normalised by its maximum value to allow use as a probability heatmap. Where heatmaps overlap for different structures within an image, the maximum value was used.

Consequently, central pixels are assigned high confidence, and more distant pixels are assigned low confidence as shown in Figure 1. This label relaxation better reflects the ambiguity of the structure boundaries, and copes better with variation in annotations. This contrasts with pre-existing work, which largely focuses on improving segmentation results near object edges, and assumes unambiguously correct edge labels in ground truth data [15, 14].

2.2 Pseudo-label self-supervision

Labelling medical imagery is widely recognised as a bottleneck due to its difficulty, high time cost and compliance challenges [11]. This is particularly true for surgical video, which generates large amounts of unstructured data. In this work, we further improved our model by using unlabelled data via self-supervision [1]. Unlike previous work on self-supervision in endoscopic surgery [11], which uses generative models and consistency-based losses, we propose a simple pseudo-label approach that requires minimal computational overhead [1].

After training an initial model on labelled data, we used its predictions to provide pseudo-labels in unlabelled data [1]. This serves as teacher in a teacher-student architecture, where a newly initialised student model is trained on both pseudo-labelled images and the original labelled images. Previous work explored similar methods in segmentation for vehicle imagery [1] and demonstrated that the student learns a superior distillation of feature space compared to its teacher, leading to improved segmentation performance. Here we adapted the approach for heatmap regression by using softmax outputs as the pseudo-labels, rather than hard segmentation outputs. All models used the same architecture, regularisation and hyperparameters (Section 2.3).

2.3 Implementation

We used convolutional neural networks throughout, with FCN segmentation architecture [6] – a common baseline for segmentation. All networks used ResNet101

as a backbone [5]. For the segmentation and self-supervised segmentation models we trained the FCN with cross-entropy loss. We initially considered using class frequency weighted cross-entropy loss, in case class imbalance was the cause for segmentation model under-performance, but results were similar to equally weighted cross-entropy loss. To assist with comparison, the proposed heatmap model was kept similar to the segmentation model, simply using softmax to convert raw logits to a heatmap. For our proposed heatmap methods, we used soft cross-entropy loss and relaxed the ground truth label as discussed in Section 2.1.

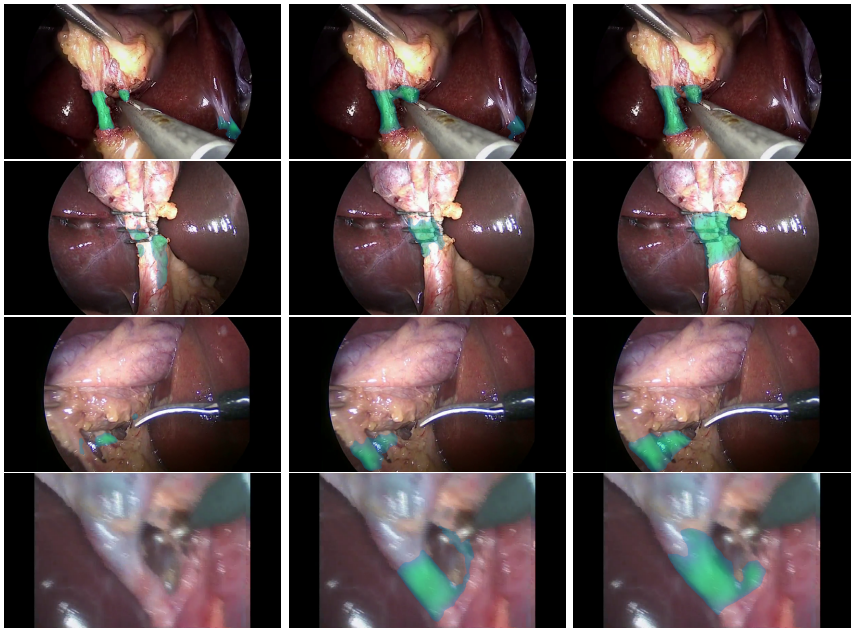
All models were implemented in PyTorch 1.5 and optimised using Adam with learning rate 1e-4 and a "poly" learning rate schedule [1], trained until convergence. During training, models used random image augmentations (padding, cropping, flipping, blurring, rotation, noising) and model regularisation via dropout. We did not perform extensive hyperparameter tuning for augmentation, nor for label relaxation parameters t and d (Section 2.1). Performance did not seem sensitive, and we used $t = 15$ and $d = 10$ throughout. For evaluation in each case, the model with lowest validation loss was used. A supervised model training takes approximately 50 epochs (10 hours) using four 16GB NVIDIA GPUs, with teacher-student self-supervision requiring approximately twice this time. For self-supervision experiments, the student model is pre-trained on teacher-generated pseudo-labels for 10 epochs, then fine tuned on ground truth labels for 50 epochs [13], again using "poly" schedule this time across the combined 60 epochs. Validation performance was not improved by pre-training for any longer, perhaps due to the relatively small size of the dataset. Similarly, we did not iterate the self-supervision as we saw no further benefit [1].

3 Experiments and results

3.1 Data and training

We used 3,050 labelled images from 75 separate laparoscopic cholecystectomy videos, frames chosen near where CVS is achieved. Frames are sampled at 1fps in a window of approximately 40s for each video. Most images contain cystic duct (90%) and/or cystic artery (87%). Labelling was performed by surgical data annotators under supervision of an anatomy specialist. Guidelines and tutorials for annotation were validated by surgeons. Labelled images were separated by video into train/val/test (60/20/20%) – with the test set held out for final evaluation of performance. We additionally used 3,682 unlabelled images derived from 90 videos for the self-supervision experiments – all used as training data.

We trained four models: a baseline segmentation method as described in Section 2.3, a heatmap method as described in Section 2.1, and variants of both methods using self-supervision to exploit the unlabelled data. We assessed performance on the validation and test sets, and then provided example model outputs from the test set for evaluation by surgeons.



(a) Segmentation (b) Heatmap (c) Heatmap+SS

Fig. 2: Example frames from the test set, with softmax model outputs imposed in blue-green. Heatmaps generally reduced false negatives e.g in the bottom two rows, and show some reduction in false positives (top row). Self-supervision led to further improvements, e.g. the false positive in the top row and false negatives in bottom three rows.

3.2 Ablation study

Table 1 shows pixel-level metrics, ordered by method (segmentation versus our proposed heatmap method) and whether self-supervision was used. To accommodate edge ambiguity, the evaluation uses VOC-style metrics, in which a 10 pixel margin around each structure is assigned as "ignore" [3] (not used during training). Heatmap detection consistently performs better than segmentation in IoU, regardless of whether self-supervision is used. Comparing without self-supervision, the IoU is higher for our proposed heatmap approach by 9.7pp/11.7pp (val/test). For both segmentation and heatmaps, our proposed self-supervision seems beneficial: it increased segmentation results by IoU by 1.5pp/0.9pp (val/test); and the IoU of heatmap methods by 3.7pp/3.1pp (val/test).

Notably, although the performance is generally best for our proposed heatmap method with self-supervision and second-best for heatmaps without self-supervision, segmentation achieved a higher pixel precision on the test set. This makes sense in light of the label relaxation, which inevitably assigns some probability mass to

Table 1: Pixel-level accuracy metrics, by method. Heatmaps typically outperformed segmentation, as shown in improvements in IoU and other metrics in val and test sets. Self-supervision (SS) generally improves models, with the possible exception of segmentation precision in the test set.

		Val			Test		
Method	SS	IoU	Precision	Recall	IoU	Precision	Recall
Segmentation	×	.547	.764	.658	.501	**.869**	.542
Segmentation	✓	.562	.807	.649	.512	.849	.563
Heatmap	×	.644	.836	.750	.618	.811	.721
Heatmap	✓	**.681**	**.867**	**.761**	**.649**	.823	**.755**

Table 2: Higher-level presence detection metrics, evaluated with IoU threshold 0.5 to count as a true positive detection. In every metric, heatmaps outperformed segmentation. Self-supervision (SS) generally improved results.

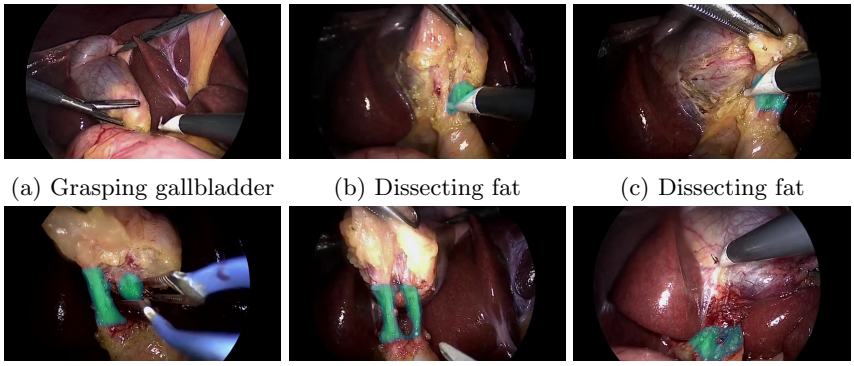
		Val			Test		
Method	SS	F1	Precision	Recall	F1	Precision	Recall
Segmentation	×	.597	.599	.594	.615	.626	.606
Segmentation	✓	.640	.640	.641	.616	.616	.617
Heatmap	×	.716	.721	.711	.694	.703	.685
Heatmap	✓	**.811**	**.833**	**.790**	**.749**	**.750**	**.749**

non-foreground pixels. Despite this, segmentation IoU (and overall performance) remains worse due to its much lower recall.

Table 2 shows metrics for frame-level presence detection, where artery and duct detections must exceed an IoU threshold 0.5 to count as true positives in a given frame. This means that low IoU detections count as false positives. Such statistics are conservative, as a lower IoU overlap may nonetheless be fairly accurate given the ambiguity of ground truth annotation extent (see Figure 1). Nevertheless, results show a similar pattern to the pixel-level performance metrics, with our proposed heatmap method outperforming segmentation, and self-supervision improving models' performance. Notably, the increased pixel-level precision of segmentation methods does not translate to structure detection, where our heatmap method performs better by every metric.

3.3 Qualitative performance across surgery

Figure 3 shows example frames and model outputs from an 11 minute excerpt of a laparoscopic cholecystectomy video in the test set. The full example video is included in Supplementary Material. Performance is generally strong. The model typically does not suffer false detections before structures are visible (3a), although it can be fooled by similar shapes near a tool tip, particularly if such shapes are visible near the gallbladder. Even when the structures are heavily coated by fat, the model tends to recognise them at least partially (3b, 3c). Manipulation does not usually prevent detection (3d).



(a) Grasping gallbladder (b) Dissecting fat (c) Dissecting fat

(d) Coagulation (e) Critical View of Safety (f) After division

Fig. 3: Example frames taken across an 11 minute excerpt from a test set video, using our self-supervised heatmap model. Model detections in blue-green.

Structures often remain detectable after division (3e). Although we were impressed by the model's generalisation, this might be undesirable in a practical implementation. This would be fixable by using surgical phase recognition [16], which could deactivate detection after division of structures.

3.4 Surgeon preference

Table 3 shows how surgeons ranked the different methods in frames taken from the test set. Example frames were intentionally chosen to show differences between the methods, as model outputs were often similar between methods. Set 1 was randomly chosen from frames with greatest differences between segmentation and heatmap outputs ($|IoU_{seg} - IoU_{heat}|$). Set 2 was randomly chosen from frames with greatest differences between supervised and self-supervised heatmap outputs ($|IoU_{sup} - IoU_{ss}|$). Each set used frames from eight different videos.

Surgeons preferred our heatmap method to segmentation, and preferred heatmaps with self-supervision to vanilla heatmaps. These preferences failed to show significance in Set 1, but did show statistical significance in Set 2 ($p < 0.05$). We discuss this further in Section 4.1. Free text feedback was generally positive where provided ("all pretty good"), although one participant did note that in one frame common bile duct was detected.

4 Discussion and conclusion

4.1 Heatmaps improve accuracy, but can impair visualisation

Our heatmap models are more accurate than segmentation, as shown in low-level pixel metrics such as IoU and higher-level presence detection such as F1 score. This is supported by blind ratings from surgeons, where they favoured our

Table 3: Surgeon preferences (blind) by surgeon and model. Average ranking (1-3, lower is better) for each method and set, with post hoc Wilcoxon signed-rank p value, bolded for significance ($p < 0.05$) after multiple comparison adjustment. p_{heat} is for difference between heatmap and segmentation, p_{SS} is for difference between self-supervised and supervised heatmaps.

Set	Participant	Segm	Heat	Heat+SS	p_{heat}	p_{SS}
1	1	2.50	2.50	1.75	.340	.425
	2	2.13	2.00	2.00		
	3	2.50	1.88	2.13		
	Avg	2.22	2.06	1.97		
2	1	2.50	2.25	1.50	**5.57e-3**	**0.0221**
	2	2.63	1.75	1.88		
	3	2.63	2.00	1.75		
	Avg	2.38	1.97	1.63		

heatmaps over segmentations in Set 2. Counterintuitively, surgeons did not show a statistically significant preference in Set 1 – despite this set being selected for maximum differences between segmentation and heatmap outputs.

We believe this discrepancy was due to visualisation preferences, based on free text feedback ("I like narrow overlays not zones"). Our heatmap models, by design, tend to detect larger areas than the segmentation models. When we selected Set 1 to maximise differences between segmentation and heatmaps, this selected several frames where the difference is due to the heatmap highlighting a larger area (see Figure 2, row 1). Conversely, Set 2 was chosen to maximise differences between supervised and self-supervised models, and did not show this effect to the same extent. This emphasises the importance of visualisation, and suggests an important direction for future work.

4.2 Self-supervision particularly helps in difficult videos

Self-supervision improves accuracy in general, but is particularly beneficial for a few difficult cases. This can be seen in Figure 2. In rows 1, 3 and 4, self-supervision slightly improved the accuracy, but the overall detection was not changed significantly, and hence the IoU remains similar. In row 2, however, the accuracy improvement was much larger as the supervised model entirely misses the cystic artery, whereas the self-supervised model detected it. This finding is borne out by per-video IoU: for most videos the IoU difference between methods is on the order of 0-5pp, but for three videos it is 10pp or greater.

4.3 Conclusion

Our work is among the first to detect the critical structures during laparoscopic cholecystectomy. When trying to detect structures with ambiguous extent and

challenging annotations, a novel heatmap-based approach based on label relaxation significantly outperformed a segmentation baseline. Self-supervision provided further improvement by using unlabelled data for additional training. Our method was validated on held-out test data and surgeon evaluations supported these findings. We hope to develop the method further by using a greater variety of anatomic classes [8, 12], such as considering the cystic artery and duct separately, and possibly annotating the common bile duct. Modelling temporal consistency across frames might also be beneficial [2]. Finally, another important advance would be to further validate the method in a large dataset covering the full range of variability. Automatic detection of critical structures in surgery has tremendous potential to improve surgical safety, training and workflow and ultimately patient outcomes. Our work will contribute towards this goal.

Acknowledgements This work was supported by the Wellcome/EPSRC Centre for Interventional and Surgical Sciences (WEISS) at UCL (203145Z/16/Z), EPSRC (EP/P012841/1, EP/P027938/1, EP/R004080/1) and the H2020 FET (GA 863146). Danail Stoyanov is supported by a Royal Academy of Engineering Chair in Emerging Technologies (CiET18196) and an EPSRC Early Career Research Fellowship (EP/P012841/1).

References

1. Chen, L.C., Lopes, R.G., Cheng, B., et al.: Naive-student: Leveraging semi-supervised learning in video sequences for urban scene segmentation. In: European Conference on Computer Vision. pp. 695–714. Springer (2020)
2. Colleoni, E., Moccia, S., Du, X., De Momi, E., Stoyanov, D.: Deep learning based robotic tool detection and articulation estimation with spatio-temporal layers. IEEE Robotics and Automation Letters **4**(3), 2714–2721 (2019)
3. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The PASCAL visual object classes (VOC) challenge. International journal of computer vision **88**(2), 303–338 (2010)
4. Fuentes-Hurtado, F., Kadkhodamohammadi, A., Flouty, E., Barbarisi, S., Luengo, I., Stoyanov, D.: EasyLabels: weak labels for scene segmentation in laparoscopic videos. International journal of computer assisted radiology and surgery **14**(7), 1247–1257 (2019)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
6. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3431–3440 (2015)
7. Mascagni, P., Fiorillo, C., Urade, T., Emre, T., Yu, T., Wakabayashi, T., Felli, E., Perretta, S., Swanstrom, L., Mutter, D., et al.: Formalizing video documentation of the Critical View of Safety in laparoscopic cholecystectomy: a step towards artificial intelligence assistance to improve surgical safety. Surgical endoscopy pp. 1–6 (2019)
8. Mascagni, P., Vardazaryan, A., Alapatt, D., Urade, T., Emre, T., Fiorillo, C., Pessaux, P., Mutter, D., Marescaux, J., Costamagna, G., et al.: Artificial intelligence for surgical safety: automatic assessment of the Critical View of Safety in laparoscopic cholecystectomy using deep learning. Annals of Surgery (2021)

9. Namazi, B., Iyengar, N., Hasan, S., Balachandra, S., Madani, A., Hashimoto, D., Alseidi, A.A., Fleshman, J.W., Sankaranarayanan, G.: AI for automated detection of the establishment of Critical View of Safety in laparoscopic cholecystectomy videos. Journal of the American College of Surgeons **231**(4), e48 (2020)
10. Pucher, P.H., Brunt, L.M., Fanelli, R.D., Asbun, H.J., Aggarwal, R.: SAGES expert Delphi consensus: critical factors for safe surgical practice in laparoscopic cholecystectomy. Surgical endoscopy **29**(11), 3074–3085 (2015)
11. Ross, T., Zimmerer, D., Vemuri, A., Isensee, F., Wiesenfarth, M., Bodenstedt, S., Both, F., Kessler, P., Wagner, M., Müller, B., et al.: Exploiting the potential of unlabeled endoscopic video data with self-supervised learning. International journal of computer assisted radiology and surgery **13**(6), 925–933 (2018)
12. Tokuyasu, T., Iwashita, Y., Matsunobu, Y., Kamiyama, T., Ishikake, M., Sakaguchi, S., Ebe, K., Tada, K., Endo, Y., Etoh, T., et al.: Development of an artificial intelligence system using deep learning to indicate anatomical landmarks during laparoscopic cholecystectomy. Surgical endoscopy pp. 1–8 (2020)
13. Yalniz, I.Z., Jégou, H., Chen, K., Paluri, M., Mahajan, D.: Billion-scale semi-supervised learning for image classification. arXiv preprint arXiv:1905.00546 (2019)
14. Yuan, Y., Xie, J., Chen, X., Wang, J.: Segfix: Model-agnostic boundary refinement for segmentation. In: European Conference on Computer Vision. pp. 489–506. Springer (2020)
15. Zhu, Y., Sapra, K., Reda, F.A., et al.: Improving semantic segmentation via video propagation and label relaxation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8856–8865 (2019)
16. Zisimopoulos, O., Flouty, E., Luengo, I., Giataganas, P., Nehme, J., Chow, A., Stoyanov, D.: Deepphase: surgical phase recognition in cataracts videos. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 265–272. Springer (2018)