Graph Convolutional Nets for Tool Presence Detection in Surgical Videos

Sheng Wang, Zheng Xu, Chaochao Yan, and Junzhou Huang(✉)

University of Texas at Arlington, Arlington, TX 76019, USA
jzhuang@uta.edu

Abstract. Surgical tool presence detection is one of the key problems in automatic surgical video content analysis. Solving this problem benefits many applications such as the evaluation of surgical instrument usage and automatic surgical report generation. Given the fact that each video is only sparsely labeled at the frame level, meaning that only a small portion of video frames will be properly labeled, existing approaches only model this problem as an image (frame) classification problem without considering temporal information in surgical videos. In this paper, we propose a deep neural network model utilizing both spatial and temporal information from surgical videos for surgical tool presence detection. The proposed model uses Graph Convolutional Networks (GCNs) along the temporal dimension to learn better features by considering the relationship between continuous video frames. To the best of our knowledge, this is the first work taking videos as input to solve the surgical tool presence detection problem. Our experiments demonstrate the employment of temporal information offers a significant improvement to this problem, and the proposed approach achieves better performance than all state-of-the-art methods.

Keywords: Surgical video analysis · Graph convolution networks · Surgical tool detection

1 Introduction

Automatic content analysis of surgical videos recorded by an endoscopic camera in minimally invasive surgery is significant for many functions in the operating room of the future [3], such as analysis of the operation steps, review of the techniques employed, evaluation of instrument usage, and automatic surgical report generation [14]. Among all the tasks of surgical video content analysis, one crucial problem is surgical tool presence detection, to detect which surgical tools are being used at a certain time during surgery. The problem is different from surgical tool detection [16] or object detection [15,19] since it does not require the awareness of the location of surgical tools or general objects. However, the

This work was partially supported by US National Science Foundation IIS-1718853 and the NSF CAREER grant IIS-1553687.

A. C. S. Chung et al. (Eds.): IPMI 2019, LNCS 11492, pp. 467–478, 2019.
https://doi.org/10.1007/978-3-030-20351-1_36

problem is challenging due to several reasons: First, multiple surgical tools could be used at the same time. Second, different tools could have partial presence and occlusion which makes it even harder to detect. Third, since the frequencies of different surgical tools being used vary a lot, the data could be very imbalanced among certain surgical tools [17].

Existing approaches and models solve this problem by engaging multi-label image classification: sampling every frame with ground truth as an image dataset, learning features from each still image and then perform classification [2,8,10,16–18]. There are two ways of feature extraction. One is to use manually hand-crafted features or pre-designed features, e.g., SIFT features. The other is to use deep neural networks such as convolution neural networks (CNNs) to extract high-level features. After applying deep neural networks, the classification accuracy generally improves. However, one key piece that is still missing from the current methods is the information along the temporal dimension, which is the nature of videos. As shown in Fig. 1, almost all surgical tool detection datasets are labeled sparsely, i.e. the tools being used are not labeled for every frame. Only a very **tiny portion** (usually only a few percentages) of video frames are manually labeled. The insufficient label information leads to a huge challenge for the research of machine learning based surgical tool presence detection. To address this problem intuitively, the temporal information from neighbor frames could help the presence detection and should provide better performance than utilizing only the labeled image. For instance, one tool might be occluded at a certain frame and it can be very difficult to recognize it from the complex background by one single image. However, when using a continuous sequence of frames, even slight movement of the surgical tool could be noticed and help the tool get detected correctly.

To utilize the temporal information of the surgical videos for detection, it is not easy to apply current methods straightforwardly. Since almost all current surgical tool detection datasets are sparsely labeled at the frame level, using fixed length frames around the labeled image as a video could either introduce noise or lack enough temporal information. It might not offer enough temporal information when the video length is too small, while it might introduce noise when the video length is too large. Besides, if we use continuous frames around the labeled image as a video, the length of videos in this problem will not be long enough or the variation of the frame contents will not be large enough to learn long-range temporal dependency with Recurrent Neural Networks (RNNs) such as Long Short-Term Memory (LSTM) [7,21,23] for general video understanding.

In this paper, we propose a novel deep neural network model named Surgical Tool Graph Convolutional Networks (STGCN) combining the power of both Convolutional Neural Networks (CNNs) and Graph Convolutional Networks (GCNs) [12]. We model the problem as a video classification problem by using the sparsely labeled frame and the neighbor frames around it. STGCN uses DenseNet [9] as our backbone to learn the spatial features from the input images and extracts the features directly from the videos with inflated 3D DenseNet. Then it applies GCNs along the temporal dimension to learn better feature with

consideration of the relationships among continuous frames. In our experiments, we demonstrate temporal information can always improve the performance by a significant amount in the detection task.

To fully demonstrate the superiority of our model, we evaluate our model on two most recently developed datasets: M2cai-tool and Cholec80 [17]. On M2cai-tool, STGCN beats the first place method of the data challenge[1] by more than 28% in mean average precision (mAP) and surpasses the previous best performance in literature as we know of. On Cholec80 dataset, the proposed STGCN improves the best performance by 10% in terms of mAP.

The contributions of this paper are summarized below:

- To the best of our knowledge, this is the first work to utilize the temporal information for surgical tool presence detection problem.
- We propose to apply GCNs to better model the temporal information from surgical videos.
- The proposed STGCN achieves state-of-the-art results with a significant gain on both M2cai-tool and Cholec80 datasets.

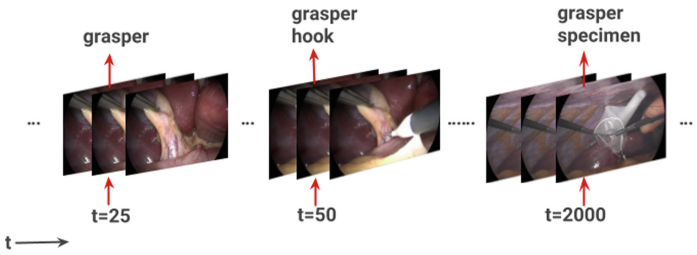

Fig. 1. Sparsely labeled surgical tool detection dataset. In this dataset, the tools being used in one image is labeled every 25 frames. Existing methods only use the labeled images for model training. In this paper, we propose to use both the labeled frame and the unlabeled frames around it as a video for model training.

2 Related Work

Surgical Tool Detection. By introducing deep neural networks to extract high-level image feature for surgical tool detection, many approaches have been developed on larger-scale datasets [2,8,10,17,18] and the overall accuracies have been improved. EndoNet [17] first proposed to use CNNs to train a tool detection model on labeled images. Since the M2cai-tool challenge, an increasing number of methods have been developed to solve this problem with M2cai-tool dataset. The winner of M2cai-tool challenge [18] modeled the problem as

[1] M2CAI Surgical Tool Presence Detection Challenge 2016: http://camma.u-strasbg.fr/m2cai2016/.

a multi-label image classification problem and used VGGNet and InceptionNet. The authors ensembled the results of these two deep models as the final result. After that, two methods have been proposed to further improve the detection performance by labeling extra localization information of surgical tools to the original dataset [2,10]. AGNet [8] proposed a model with two parts: one attention model as a global network to detect the areas with high possibilities to contain the surgical tools, and one local model to detect the tools from selected areas. AGNet has achieved the best performance. However, given the fact that each video is only sparsely labeled at the frame level, all these existing methods modeled surgical tool presence detection problem as an image classification problem without taking the advantage of the temporal information from unlabeled neighboring frames around the labeled frame.

Video Understanding and Surgical Video Understanding. Meanwhile, many researchers focus on video inference for the better ability of computer video understanding. A great number of cutting edge approaches have been proposed to improve the video understanding performance, and several complex datasets have been built to promote related research [1,6].

Recent video understanding work focuses on modeling long term temporal information with Recurrent Neural Networks. There has also been some surgical video understanding work on the surgical phase recognition using RNNs [11,22]. Different from the tool presence detection problem, surgical phase recognition demands to model long term temporal information on a whole surgical video, while the short video among the single labeled surgical frame does not need long term temporal modeling. Thus, RNNs based methods do not serve as a good fit in our problem.

Graph Convolutional Networks. Until recent years, very little attention has been devoted to the generalization of neural network models to more general structure such as graphs or networks [4,13]. The deep models handling the graph-like structure are named Graph Convolutional Networks (GCNs).

Our work is motivated by recent work on human recognition [20] using GCN as one crucial part of their proposed deep neural network model. In this work, the authors built a graph containing nodes corresponding to different object proposals aggregated over video frames. Different from this work, we model the feature extracted from each frame as a node and build the graph as the relationship within the continuous frames of a video segment to learn better feature with temporal information.

3 Methodology

3.1 Problem Definition

Image Classification. Existing methods for surgical tool detection models the problem as an multi-label image classification problem. Given the image x_t at frame t, models are trained to get the prediction for the input image $F(x_t)$ close to its groundtruth y_t.

Video Classification. In this paper, we propose to use not only the labeled image but also the neighbor images as a video segment for model training and evaluation. Thus, the problem becomes that given a video segment corresponding to the t frame $[x_{t-l}, ..., x_t, ..., x_{t+l}]$, where l is the number of frames before and after the labeled frame image we take into consideration, models are trained to get the prediction for the input video $F(x_{t-l}, ..., x_t, ..., x_{t+l})$ close to its groundtruth y_t.

3.2 Model Overview

As shown in Fig. 2, the proposed STGCN contains several components. To get the features from the input video, we use an inflated 3D DenseNet-121 [1,9] to get the representation of each frame in the video. We take the representation of each frame as a node and build a similarity graph on these nodes. By applying GCNs on the constructed graph, the GCNs will adaptively generate the features considering the relationships among the nodes in the graph, i.e., the temporal relationship in continuous frames. After that, we use pooling over all the nodes corresponding to continuous frames. We note the pooling layer as temporal pooling since what it does is applying the pooling on the temporal dimension. The details of each component will be discussed in the following sections.

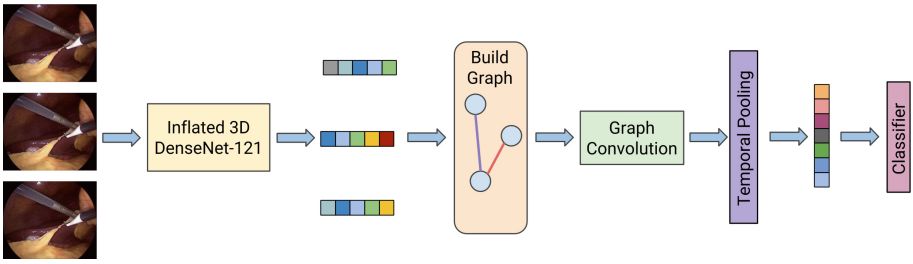

Fig. 2. The overview of the proposed STGCN.

3.3 Inflated 3D DenseNet

Different from most deep convolutional neural networks, DenseNet [9] connects all the convolutional layers in pairs when their spatial output sizes are the same. The output of each feature maps also serves as the input of all following convolutional layers. The idea is similar to Residual Networks. However, it can reuse all the features in the network. This sort of network almost exhaustively maximizes the network capacity to squeeze its spatial feature extraction and prediction power. Also, the network can alleviate the vanishing-gradient problem, strengthen feature propagation and substantially reduce the number of the parameters in the network.

In our proposed model, we use DenseNet to learn and extract spatial features for each frame in the input video. To adapt DenseNet for video input, the original DenseNet needs to be inflated to 3D ConvNet (I3D) [1,9]. That is, to support the input video of length t, a 3D kernel with $t \times k \times k$ dimensions can be inflated from a 2D $k \times k$ kernel by copying the weight t times and rescaling by $1/t$. In our implementation, we use 11 as the number of frames. The growth rate is 32 as the default number for DenseNet-121.

3.4 Graph Convolutional Networks

We apply GCNs [4] in the proposed framework to better capture the temporal relationship along the continuous frames.

Similarity Graph Building. For a video input $X = [x_{t-l}, ..., x_t, ..., x_{t+l}]$ with length N, where x_t is with the dimension of d, containing the labeled surgical tools while others not. We use the output of the fully-connected layer right after the fourth dense block from our inflated DenseNet-121 model to get the feature representations noted as $[f(x_{t-l}), ..., f(x_{t-1}), f(x_t), f(x_{t+1}), ..., f(x_{t+l})]$. We regard the representation for each frame as one vertex (node) v_k of a graph, and use the similarity S_{ij} between each pair of nodes (v_i, v_j) as the corresponding edge of the graph. Thus, the graph could reflect the temporal relationship of the continuous frames.

There are quite a few different methods to build the similarity graph. In the proposed STGCN, we use the **cosine similarity** to build the graph as

$$S_{ij} = \frac{f(x_i) \cdot f(x_j)}{\|f(x_i)\| \, \|f(x_j)\|}, \tag{1}$$

and we can get the similarity graph G after normalizing each row of S as

$$G_{ij} = \frac{e^{S_{ij}}}{\Sigma_{j=1}^{N} e^{S_{ij}}}. \tag{2}$$

Graph Convolutional Layer. After building the similarity graph, the graph convolutional layer could be represented as

$$Z = GXW, \tag{3}$$

where W is the weight mapping feature of each node to another dimension. The graph convolutional layer could not only map the feature as a general convolutional layer, but also take the graph information (temporal relationship among the frames in the input video) into consideration. In the surgical tool detection problem, graph convolutional layer could learn features while adaptively reference the relationship among the frames to generate the correct prediction.

The graph convolutional layers could be stacked as a deep GCNs or in general CNNs by

$$X^{(l)} = GX^{(l-1)}W^{(l-1)}, \tag{4}$$

where $X^{(l-1)}$ is the feature map as the input to current graph convolutional layer, $W^{(l-1)}$ is the weight. $X^{(l)}$ is the output of current layer as well as the input of next layer.

In our proposed model, we use a residual variation of the graph convolutional layer as

$$X^{(l)} = \sigma\left(GX^{(l-1)}W^{(l-1)}\right) + X^{(l-1)}, \tag{5}$$

where $\sigma(\cdot)$ is the activation function after the graph convolutional layer and we add $X^{(l-1)}$ to the output of the layer as a residual component.

3.5 Temporal Pooling

The feature after the last graph convolutional layer contains N features for the N frames. Then we add a temporal pooling layer to combine all the N features from N frames in the video. Temporal pooling layer has no difference than general pooling layer that it aggregates the features along the temporal dimension. It should not be a crucial factor in the performance of the proposed model since the features for the pooling layer has utilized the temporal information with GCNs. However, we still try different pooling strategies in STGCN to seek potential improvement. There are a lot of methods for pooling such as l_p pooling, average pooling, max pooling, and max-min pooling [5]. In later ablation experiments, we will show the performance of different pooling methods on Cholec80 dataset.

Given a sequence of N d-dimensional dense features after GCNs as $x^{(i)}$, where i is from 1 to N, temporal pooling pools the features along the time dimension. Assume the N-dimensional feature after temporal pooling as $\tilde{x}$, for **max temporal pooling**, $\tilde{x}_k = max(x_k^{(i)})$ where i from 1 to N, for **average temporal pooling**, $\tilde{x}_k = \frac{1}{N}\sum_{i=1}^{N} x_k^{(i)}$ and for l_p **temporal pooling**, $\tilde{x}_k = \sqrt[p]{\sum_{i=1}^{N}\left(x_k^{(i)}\right)^p}$ where k is from i to d for all temporal pooling methods. For **max-min pooling**, we apply a simple version of max-min pooling, which could be computed as:

$$\tilde{x}_k = max(x_k^{(i)}) + \alpha min(x_k^{(i)}), \tag{6}$$

where α is a hyperparameter balancing the weights of max pooling and min pooling.

4 Experiments

4.1 Implementation Details

DenseNet. We use DenseNet-121 pretrained from ImageNet to continue training on surgical tool detection datasets for a multi-label image classification. Then we inflate the trained DenseNet to 3D DenseNet. To avoid using temporal information in the inflated DenseNet, we keep all the dimension of kernels in either dense blocks or other convolutional/pooling layers as 1. Thus, all the temporal

information is used in the GCNs part of the proposed model. We fix the length of the video segment around each labeled image to 11 to train the GCNs and following classifier. The DenseNet is trained with Adam optimizer with learning rate 0.0001 for 200 epochs. The learning rate will be decayed if the training loss does not decrease after three continuous training epochs.

GCNs. After extracting the feature presentation for each frame from Inflated DenseNet-121, we input the features along with the similarity graph into the GCNs. The feature we get from the inflated DenseNet-121 has the dimension of 1024. In our GCNs, we use one graph convolutional layer which maps the input feature from 1024 dimensions to 1024 dimensions. Then the temporal pooling layer is added to pool the features along the temporal dimension. After that is followed by a layer maps 1024 dimensions feature to the number of surgical tools for classification. In GCNs, both batch normalization and dropout are added after the graph convolutional layer. Batch normalization is also added before the graph convolutional layer. We train the GCNs with Adam optimizer with learning rate 0.0001 for 300 epochs. The dropout rate is set as 0.75 in our training. The same learning rate decay strategy is used as the one in training DenseNet. For max-min pooling, we fix the hyperparameter α to 0.75.

4.2 Data Description

M2cai-Tool Dataset [17]**.** This dataset from M2CAI surgical tool presence detection challenge contains 15 videos of laparoscopic cholecystectomy procedures from the University Hospital of Strasbourg/IRCAD (Strasbourg, France). The dataset is split into two parts: the training subset (containing 10 videos) and the testing subset (5 videos) by the challenge organizers. The videos are recorded at 25 fps and labeled at 1 fps (one labeled frame in every 25 frames). There are 23287 training samples and 12541 testing samples. The evaluation process only considers the labeled frames in testing dataset.

In this dataset, there are seven kinds of surgical tools in total as shown in Fig. 3: grasper, hook, clipper, bipolar, irrigator, scissors, and specimen bag.

Cholec80 Dataset. The Cholec80 dataset is larger than M2cai-tool dataset. It contains 40 videos (86304 labeled frames) for training and 40 videos (98194 labeled frames) for testing. The Cholec80 is also from the University Hospital of Strasbourg/IRCAD and has the same recording rate, labeling rate, and tool set as M2cai-tool dataset.

Validation Sets. For both M2cai-tool and Cholec80 datasets, we split 10% samples from training sets as validation sets. We tune our hyperparameters on the validation sets.

Fig. 3. The surgical tools used in M2cai-tool and Cholec80 datasets. Both of the datasets have the same seven surgical tools.

4.3 Evaluation Metric

We use the mean average precision (mAP) among the average precision (AP) on each of the seven surgical tools, which is the same as the challenge evaluation metric. To ensure a fair comparison with all the methods during and after the challenge, we exactly follow every detail of data usage and evaluation protocol used in M2CAI challenge.

4.4 Experimental Results

M2cai-Tool Dataset. In this experiment, we choose the winner's and the 3rd place's methods from the challenge, as well as three approaches after the challenge as comparison methods. Among the challenge methods, EndoNet [17] first proposed using CNN as a baseline model. The winner of the challenge [18] introduced an ensemble model of VGGNet and Inception Net. However, the highest mAP is a little above 60%. For the methods after the challenge, both Jin *et al.* [10] and Choi *et al.* [2] added location information of the tools by adding surgical tools bounding box to the dataset. These two approaches improved the mAP by 10%. AGNet [8] proposed to use an attention model to increase the detection performance. AGNet trained two cascaded deep convolutional neural networks: the first one as a global model to locate the area which has higher responses by the attention based classification network, and then the second one as a local model to classify the cropped areas with higher attention. Before our method, AGNet has the best mAP among all the approaches. We compare all these methods with our results of STGCN results. We include three variations of the proposed STGCN as side ablation experiments. STGCN (DenseNet) is the model we train and test on the labeled images without using any temporal information. STGCN (3D DenseNet + LSTM) contains the inflated 3D DenseNet as the backbone, and add an LSTM layer after it to extract the temporal information from continuous frames in the video. The difference between STGCN (3D DenseNet + GCNs) and STGCN (3D DenseNet + LSTM) is that STGCN (3D DenseNet + GCNs) uses GCNs to exploit the temporal information.

As shown in Table 1, the STGCN (DenseNet) model has achieved better performance than all existing methods. Compared to AGNet, STGCN (DenseNet) has not used any attention strategy to boost the performance to have around 2% better mAP than AGNet. By adding temporal information, the STGCN (3D DenseNet + LSTM) and the proposed STGCN (3D DenseNet) both improves our image classification model STGCN (DenseNet). With GCNs, it

Table 1. The results on M2cai-tool dataset.

Methods	Mean AP
STGCN (3D DenseNet + GCNs)	**90.24**
STGCN (3D DenseNet + LSTM)	89.03
STGCN (DenseNet)	88.27
AGNet [8]	86.8
Choi et al. [2]	72.3
Jin et al. [10]	71.8
Sheng et al. [18]	63.8
Twinanda et al. [17]	52.5

could have 1% better mAP than LSTM. Our results demonstrate that temporal information is effectively helpful for surgical tool presence detection, and GCNs is better than LSTM in this problem.

Cholec80 Dataset. We compare the proposed STGCN result with the two baseline methods ToolNet and EndoNet on this dataset in [17]. We also try the four different temporal pooling methods: l_2 pooling (STGCN(l_2)), average pooling (STGCN(avg)), max pooling (STGCN(max)), and max-min pooling (STGCN) on this dataset. Results are shown in Table 2. On this larger dataset, the proposed STGCN has better performance than the baseline methods ToolNet and EndoNet modeling the problem as a multi-label image classification problem. By utilizing the temporal information, the proposed STGCN has improved the performance around 10% in mAP.

Among all the results with different temporal pooling strategies, max-min pooling has better performance. However, the improvement is so small that it could be caused by randomness during model training. The slight difference among the four pooling methods offers support to our analysis that the graph convolutional layer has utilized the temporal information so how to aggregate the information along the temporal dimension is not sensitive, which could be convenient for model designing.

Table 2. The results on Cholec80 dataset.

	ToolNet [17]	EndoNet [17]	STGCN (l_2)	STGCN (avg)	STGCN (max)	STGCN
mAP	80.9	81.0	90.05	90.11	90.08	**90.13**

By comparing the results of the proposed STGCN with the existing methods on both M2cai-tool and Cholec80 datasets, it demonstrates that there is always significant improvement by utilizing the extra temporal information by modeling the surgical tool presence detection as a video classification problem.

Besides, with the power of GCNs, STGCN has better accuracy even compared with existing leading methods using multiple CNNs [8] or labeling additional localization ground truth [10].

5 Conclusion

Surgical tool presence detection is an essential problem for automatic surgical video analysis. To use the temporal information from the video data, we propose a novel model named STGCN which applies graph convolutional learning on continuous video frames to better use the temporal information. STGCN can directly take a video (a sequence of image frames) as input, extract both spatial and temporal features of the input and get excellent surgical tool detection precision. To the best of our knowledge, this is the first model which can take video sequences as inputs for surgical tool presence detection. On both of the two datasets to evaluate our model, STGCN has the best mean average precision. Comparing with the models that only use spatial features, we demonstrate that with GCNs, the temporal information is effective to improve surgical tool presence detection performance.

References

1. Carreira, J., Zisserman, A.: Quo vadis, action recognition? A new model and the kinetics dataset. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4724–4733. IEEE (2017)
2. Choi, B., Jo, K., Choi, S., Choi, J.: Surgical-tools detection based on convolutional neural network in laparoscopic robot-assisted surgery. In: 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 1756–1759. IEEE (2017)
3. Cleary, K., Chung, H.Y., Mun, S.K.: OR 2020 workshop overview: operating room of the future. In: International Congress Series, vol. 1268, pp. 847–852. Elsevier (2004)
4. Defferrard, M., Bresson, X., Vandergheynst, P.: Convolutional neural networks on graphs with fast localized spectral filtering. In: Advances in Neural Information Processing Systems, pp. 3844–3852 (2016)
5. Durand, T., Mordan, T., Thome, N., Cord, M.: Wildcat: weakly supervised learning of deep convnets for image classification, pointwise localization and segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017), vol. 2 (2017)
6. Gu, C., et al.: Ava: a video dataset of spatio-temporally localized atomic visual actions. arXiv preprint arXiv:1705.08421 (2017)
7. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
8. Hu, X., Yu, L., Chen, H., Qin, J., Heng, P.-A.: AGNet: attention-guided network for surgical tool presence detection. In: Cardoso, M.J., et al. (eds.) DLMIA/ML-CDS -2017. LNCS, vol. 10553, pp. 186–194. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-67558-9_22

9. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: CVPR, vol. 1, p. 3 (2017)
10. Jin, A., et al.: Tool detection and operative skill assessment in surgical videos using region-based convolutional neural networks. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 691–699. IEEE (2018)
11. Jin, Y., et al.: SV-RCNet: workflow recognition from surgical videos using recurrent convolutional network. IEEE Trans. Med. Imaging **37**(5), 1114–1126 (2018)
12. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2016)
13. Li, R., Wang, S., Zhu, F., Huang, J.: Adaptive graph convolutional neural networks. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)
14. Loukas, C.: Video content analysis of surgical procedures. Surg. Endosc. **32**(2), 553–568 (2018)
15. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems, pp. 91–99 (2015)
16. Sznitman, R., Becker, C., Fua, P.: Fast part-based classification for instrument detection in minimally invasive surgery. In: Golland, P., Hata, N., Barillot, C., Hornegger, J., Howe, R. (eds.) MICCAI 2014. LNCS, vol. 8674, pp. 692–699. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10470-6_86
17. Twinanda, A.P., Shehata, S., Mutter, D., Marescaux, J., De Mathelin, M., Padoy, N.: EndoNet: a deep architecture for recognition tasks on laparoscopic videos. IEEE Trans. Med. Imaging **36**(1), 86–97 (2017)
18. Wang, S., Raju, A., Huang, J.: Deep learning based multi-label classification for surgical tool presence detection in laparoscopic videos. In: 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017), pp. 620–623. IEEE (2017)
19. Wang, S., Yao, J., Xu, Z., Huang, J.: Subtype cell detection with an accelerated deep convolution neural network. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 640–648. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46723-8_74
20. Wang, X., Gupta, A.: Videos as space-time region graphs. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11209, pp. 413–431. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01228-1_25
21. Xu, Z., Wang, S., Zhu, F., Huang, J.: Seq2seq fingerprint: an unsupervised deep molecular embedding for drug discovery. In: Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, pp. 285–294. ACM (2017)
22. Yengera, G., Mutter, D., Marescaux, J., Padoy, N.: Less is more: surgical phase recognition with less annotations through self-supervised pre-training of CNN-LSTM networks. arXiv preprint arXiv:1805.08569 (2018)
23. Zhang, X., Wang, S., Zhu, F., Xu, Z., Wang, Y., Huang, J.: Seq3seq fingerprint: towards end-to-end semi-supervised deep drug discovery. In: Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, pp. 404–413. ACM (2018)