UNIVERSITY OF SCIENCE
FACULTY OF INFORMATION TECHNOLOGY
ADVANCED PROGRAM IN COMPUTER SCIENCE

HUỲNH LÂM HẢI ĐĂNG - HỒ THỊ NGỌC PHƯỢNG

ENHANCING VIDEO SUMMARIZATION WITH CONTEXT AWARENESS

BACHELOR OF SCIENCE IN COMPUTER SCIENCE

HO CHI MINH CITY, 2023

UNIVERSITY OF SCIENCE
FACULTY OF INFORMATION TECHNOLOGY
ADVANCED PROGRAM IN COMPUTER SCIENCE

HUỲNH LÂM HẢI ĐĂNG - 19125003

HỒ THỊ NGỌC PHƯỢNG - 19125014

ENHANCING VIDEO SUMMARIZATION WITH CONTEXT AWARENESS

BACHELOR OF SCIENCE IN COMPUTER SCIENCE

THESIS ADVISORS:
ASSOC.PROF. TRẦN MINH TRIẾT - PROF. LÊ TRUNG NGHĨA

HO CHI MINH CITY, 2023

COMMENT OF THESIS'S ADVISORS

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

COMMENTS OF THESIS'S REVIEWER

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

ACKNOWLEDGEMENT

Authors

Huỳnh Lâm Hải Đăng & Hồ Thị Ngọc Phượng

THESIS PROPOSAL

Thesis title: ENHANCING VIDEO SUMMARIZATION WITH CONTEXT AWARENESS

Advisors: Assoc.Prof. Trần Minh Triết, Dr. Lê Trung Nghĩa

Duration: January 1st, 2023 to August 14th, 2023

Students: Huỳnh Lâm Hải Đăng (19125003) - Hồ Thị Ngọc Phượng (19125014)

Theme of Thesis: theoretical research, proposed improvements.

Content:

We aim to propose a novel approach for improving video summarization quality by integrating context awareness. We also aim to propose an evaluation metric that better suits the practical use of problem in real life.

The details include:

- Literature Review and Proposal Writing
  - Conduct a comprehensive literature review on video summarization, identifying the current state-of-the-art techniques and their limitations, as well as opportunities for improvement.
  - Analyze the importance of context in video summarization and compare existing methods and tools for context extraction in videos, in terms of performance and applicability for video summarization.
  - Develop a research proposal, including research questions, hypothesis, and methodology, based on the findings from the literature review.

- Dataset Collection
  - Collect datasets suitable for training and testing.
  - Analyze the current evaluation metrics for video summarization and identify their flaws.
  - Define relevant performance metrics for evaluating the effectiveness of the context awareness in improving the quality of video summarization.
- Model Development
  - Develop baseline model for the sake of benchmarking.
  - Develop different models to prove the proposed hypothesis.
  - Train and optimize the model using the collected datasets.
- Comparison with Existing Video Summarization Techniques
  - Conduct experiments on proposed enhancements with a thoroughly designed ablation study.
  - Analyze the strengths and weaknesses of the proposed approach.
  - Conduct surveys based on the proposed evaluation metric.
- Demo Application Development
  - Develop a demo application that can demonstrate the functionality and usability of the proposed framework for video summarization.

- Thesis Writing and Submission
  - Write up the thesis, including an introduction, literature review, methodology, results, discussion, and conclusion.
  - Submit the thesis for review and evaluation by the thesis committee.

Implementation plan:

- Literature Review and Proposal Writing: 01-01-2023 to 31-01-2023
- Dataset Collection: 01-02-2023 to 15-02-2023
- Saliency Detection Model Development: 16-02-2023 to 15-03-2023
- Video Summarization Model Development: 16-03-2023 to 15-04-2023
- Integration of Saliency Detection into Video Summarization: 16-04-2023 to 15-05-2023
- Comparison with Existing Video Summarization Techniques: 16-05-2023 to 31-05-2023
- Demo Application Development: 01-06-2023 to 30-06-2023
- Thesis Writing and Submission: 01-07-2023 to 31-07-2023

	December 26th 2022
Advisors	**Authors**
Assoc. Prof. Trần Minh Triết	Huỳnh Lâm Hải Đăng
Dr. Lê Trung Nghĩa	Hồ Thị Ngọc Phượng

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Since the development of **computational imaging**, the number of medical images have increased dramatically, occupying a lot of time of pathologists to provide further assessments. Therefore, **computer-aided diagnosis (CADx)** systems have been built to automatically transform these data to useful information, shorten the amount of analyzing time spent by professionals.

Abnormal detection and **segmentation** are considered as two of the main tasks to build up a CADx system. While the former focuses on designing and implementing a method that has an ability to detect different diseases symptoms, abnormal signals and anatomical landmarks, the latter, for example in the microscopic tissue images, identifies regions belonging the cell nucleus and extracts nuclear morphometric, pleomorphism or appearance features including average size or density.

Inspired by recent advances in deep learning, such as **Faster R-CNN** and the **U-Net** architecture, the authors propose new enhancements to tackle problems appearing in medical image analysis, especially abnormal detection and nuclear segmentation.

In more details, new novel **data augmentation** has been designed and adapted to corresponding domains. Moreover, in abnormal detection, the combination of **Residual Neural Network** and **Faster R-CNN** are applied to classify endoscopic images, while the **rotation equivariance** and **residual blocks** have been cooperated to the U-Net architecture to improve performance of nuclear segmentation task.

Through conducted experiments, ablation studies, as well as the results from challenges, the merit of each enhancement has been demonstrated, showing the effectiveness of each proposed method. In Medico: The 2018 Multimedia for Medicine Task of MediaEval 2018, our detection methods achieve the best performance not only in term of the accuracy but also in the inference time, which are 94.24% and 99.33% in term of Matthew Correlation Coefficient.

A subset of proposed enhancements in nuclear segmentation also ranks $11^{th}/32$ with 65.57% on AJI in the MoNuSeg Challenge 2018. These results show the great potential of deep learning applications in medical image processing.

CHAPTER 1

EXPERIMENTAL RESULTS OF ABNORMALITIES FINDINGS AND LANDMARK DETECTION FOR ENDOSCOPIC IMAGES

In this chapter, we describe the dataset, task descriptions as well as the evaluation metrics used in both following challenges: The 2018 Multimedia for Medicine Task and and The Biomedia ACM MM Grand Challenge 2019. We not only report our official results provided by organizers but also discuss about the performance of each component in our algorithm by trying different experiment configurations and analyze the corresponding results.

1.1 Dataset and evaluation metrics

1.1.1 Dataset

Dataset description

According to organizers of The 2018 Multimedia for Medicine Task, the offical dataset contains a multi-class set of frames and videos for at least 4 different diseases, at least 4 different landmarks and at least 2 different findings. Each class will consist of at least 1000 frames and at least 2 short videos. The training set will be released with ground truth. All the frames in training set will be labelled with corresponding class-label. Up to 10% of training set frames will have a detailed ground truth masks showing the exact location of disease or finding within the frame. The ground truth is collected with the help of GI endoscopists from our partner hospitals. The pre-extracted visual features such as global image features, namely: JCD, Tamura, ColorLayout, EdgeHistogram, AutoColorCorrelogram and PHOG are also included in this dataset.

Classes

Totally, there are 16 classes of abnormal symptoms and landmark of human's gastrointestinal track (GI track). The visualisation of landmark's corresponding position in the GI track are shown in Figure 1.1. As given in the figure, the *esophagitis, normal z-line, normal pylorus, ulcerative colitis, colon, normal cecum, retroflex-stomach, retroflex-rectum and stool (plenty/inclusion)* are classes which illustrate the landmark in human's GI Track. Noticeably, the *esophagitis* and *normal z-line* share a same anatomical position. However, the class *esophagitis* (marked with a red bounding box) refers to an abnormal situation and *normal z-line* (marked with a green bounding box) appears in normal people's GI track. Similarly, the *ulcerative colitis* (marked with a red bounding and *colon* (marked with a green bounding box) refer to an abnormal and a normal situation respectively.

Besides, the *dyed-lifted-polyps, dyed-resection-margins, polyps* and *instruments* are abnormal symptoms that can be appeared at any position in the human's GI track. In addition, the *polyp* is a living entity in the GI track, doctors can inject some chemical component in order to demolish polyps. This chemical changes polyp's color to white, which became the *dyed-lifted polyps*. After that, doctors will use some special *instruments* to remove the polyp completely, the remains are *dyed-resection-margins*.

Multi-classes classification

When we conducting our experiments on this dataset, we find out that both the development set and the test test contain a number of images can be classified into several classes simultaneously. After noticing the task organizers for this problem, they define a priority list according to the important level of that finding during the endoscopy screening process. The list is given bellow in descending

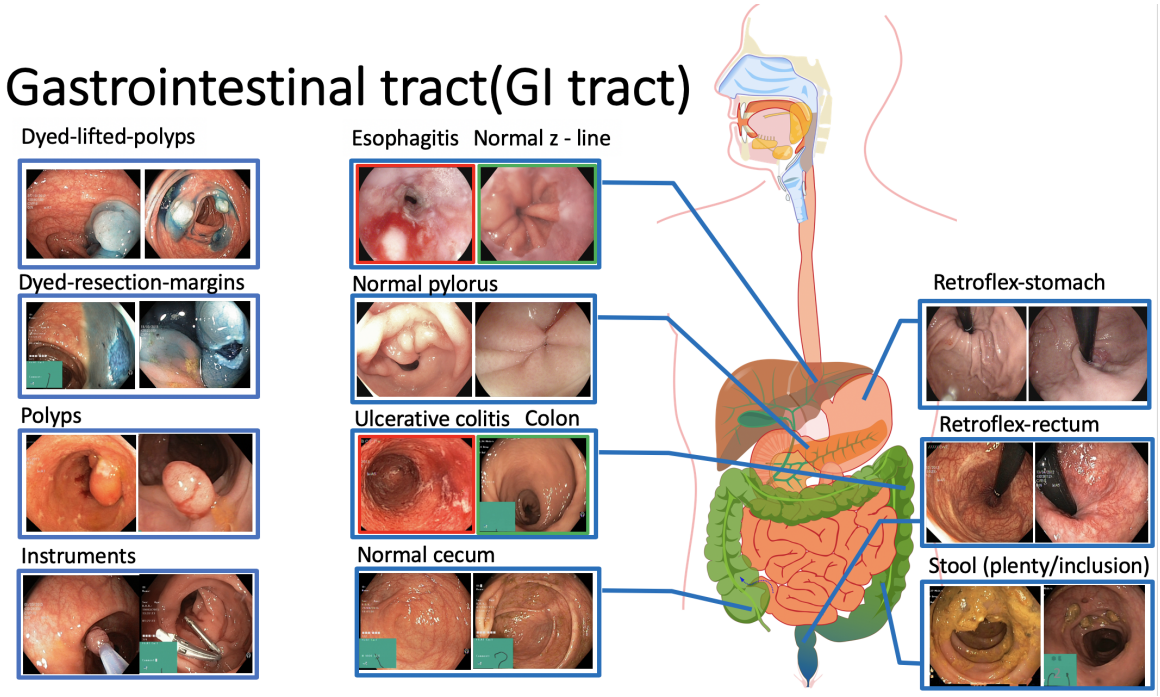


Figure 1.1: Classes of the Medico: The 2018 Multimedia for Medicine Task official dataset and their corresponding order in the Gastrointestinal track.

important level (most important to less important):

1. out-of-patient
2. instruments
3. dyed-lifted-polyps
4. dyed-resection-margins
5. polyps
6. esophagitis
7. ulcerative-colitis
8. retroflex-rectum
9. retroflex-stomach
10. normal-cecum
11. normal-pylorus
12. normal-z-line
13. stool-plenty
14. stool-inclusions
15. colon-clear
16. blurry-nothing

Development Set and Test Set

In this challenge, task organizers divide this 16-classes dataset into two subset: *the development set* consists of 5293 images and the *test set* consits of 8740

images. Only the image's label of the *development set* are visible to participants.

1.1.2 Evaluation metrics

Regarding to the evaluation process, task organizers propose several metrics for each sub-task. Totally, there are four sub-tasks can be conducted on this dataset, including *Classification of diseases and findings*, *fast and efficient classification*, *basic reporting* and *advance reporting*. In our work, we evaluate our approaches on the first and the second sub-task. The remain sub-tasks are not conducted by our team or others and they are left for future development.

- **Classification of diseases and findings:** The evaluation metric is the multiclass version of the Mathews Correlation Coefficient (MCC) [**?**], which captures the quality of classification as reflected in the correlation between the ground truth and the classifier's predictions. The MCC can be calculated directly from the confusion matrix using the formula:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{1.1}$$

  where TP stands for the number of true positives, TN is the number of true negatives, FN is the number of false negatives and FP is the number of true negatives.

- **Fast and efficient classification:** Participants of this sub-task are asked to run the code on a standard PC and provide information about hardware used. The time from input to output will be measured and weighed by the accuracy of the output. Participants are also asked to record the amount of training data used. The amount of training data will also be weighted by the accuracy of the output.

- **Reporting (exploratory/optional):** The metric will reflect the com-

pleteness and the correctness of the summary. This sub-task allows organizers to focus on understanding the performance of the automatic systems in practice (from the point of view of the medical expert). However, this sub-task is currently an optional sub-task.

- **Advanced reporting (exploratory/optional):** In order to evaluate participant's result on this sub-task, two medical partners will participate in the evaluation process in terms of how useful it is for them and if it satisfies existing demands for documentation of endoscopic procedures. Similar to the above sub-task, this sub-task is also an optional sub-task.

1.2 Experimental setup

1.2.1 Re-labeling Medico Development Dataset

After training with the development set of Medico 2018 challenge, we find some issues related to some training samples in the given dataset. Firstly, with a help of a medical student, we found out that there are inappropriate labels according between different classes. Secondly, there are several samples in the development set that conflict with the priority list that the organizers proposed. Therefore, in order to help our model to learn with the least confusing, we apply the new labels and make them as the modified version of the original dataset. We also make that modified version of the given dataset available to be contributed for the task organizers in future challenges.

1.2.2 Abnormalities localization with Faster R-CNN

We inherited the work from Chen et al. [**?**] which is a Tensorflow implementation of faster RCNN detection framework. Due to the small size of abnormalities in endoscopic image dataset, we use several anchors size at $4, 8, 16, 32$ and

corresponding aspect ratios at $0.5, 1, 2$. ResNet 101 pre-trained model [1] is used as Faster R-CNN backbone. The concepts we used to train including *instruments, polyps, dyed-lifted-polyps* and *dyed-resection-margins*. Original and our augmented dataset are trained simultaneously for 70000 iterations. It usually takes 5-6 hours training on GeForce GTX 1080 GPU.

In test-time inference, each image is fed through the module and we only keep bounding boxes that have confident score over 0.95.

1.2.3 ResNet Classifier and Multi-tasks Classifier

Classifiers used in our work are implemented with PyTorch - a Python machine learning library. Original and our augmented dataset are trained simultaneously for 200 epochs with Adam optimizer [**?**]. The learning rate is set at 10^{-3} and decay by a factor of 0.1 every 10 epochs. When training the model, we freeze every layer except the last convolution and fully connected layer. In training phase, each image is resized into a batch of 64 images with size 224×224 are fed into the network at each step. We also apply random horizontal flip and random crop augmentation during this phase.

1.3 Experimental results

1.3.1 Abnormalities localization with Faster R-CNN results

According to our experiments, applying object detection model such as Faster R-CNN shows significant improvements in two situation:

- In the first situation, abnormal symptoms appeared in small regions of an image, that can be easily mis-classified by the image classification module. This situation usually happens with the *instruments* class. Three examples

[1] `github.com/tensorflow/models/tree/master/research/slimÕá#pre-trained-models`

are given in Figure 1.2 where *instruments* appeared in really small region at the corner of an image.

- In the second situation, multiple abnormal symptoms seem to appeared together in the same image. Figure 1.3 illustrated this situation where *dyed-lifted-polyps* and *dyed-resection-margins* symptoms appeared together. By having information about every symptom and their position inside an image, it is easier for the system to fit with the priority list of this challenge. Besides, it also an useful source of information for future endoscopy diagnosis system that can localize the position of abnormal symptoms for doctors.

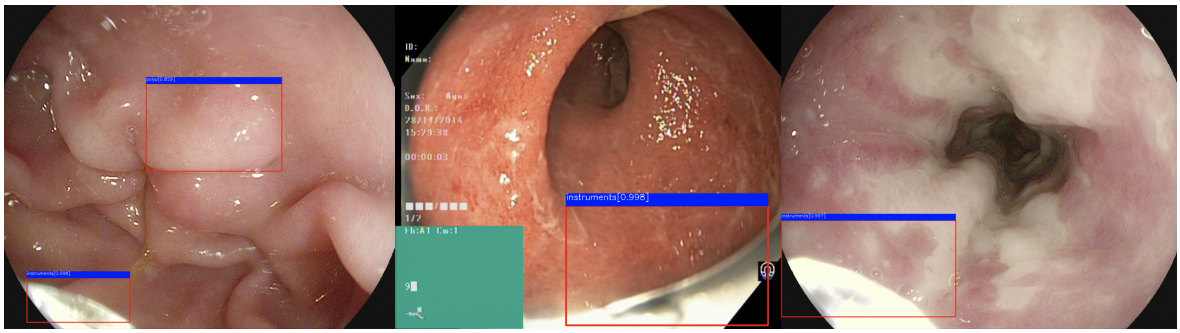


Figure 1.2: *Instrument* objects detected in endoscopic images. They are used to be mis-detected by image classification model which usually focus on global visual features.

1.3.2 Medico: The 2018 Multimedia for Medicine Task - Official results

Two out of three our improvements are used in the Medico: The 2018 Multimedia for Medicine Task challenge including applying object detection model on abnormal findings and dataset augmentation. Table 1.2 shows the official

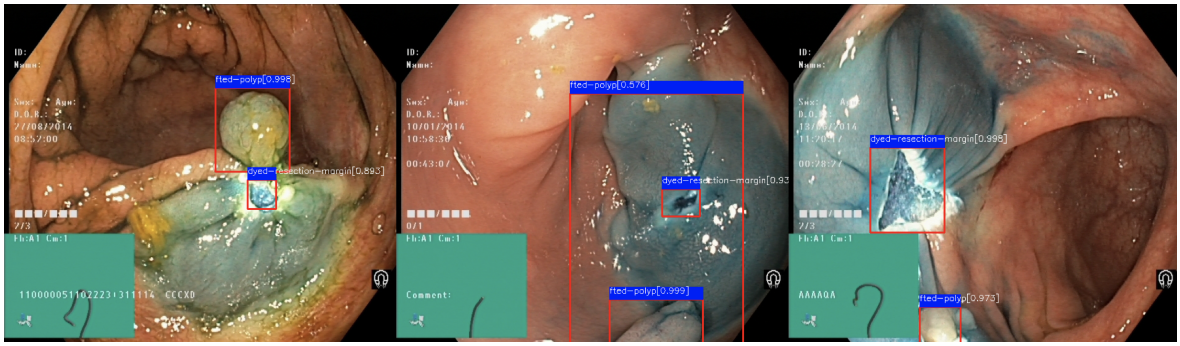


Figure 1.3: *Dyed-lifted-polyps* and *dyed-lifted-margins* symptoms detection.

results of the international challenge. We received the first place winner among 10 teams around the world with the MCC score at 94.24% and 99.33% in term of the accuracy.

There is a trade-off between speed and accuracy when comparing the result of *Run01* and *Run02*. In *Run02*, we reduce a large number of images passing through Faster R-CNN for the sake of time, so its performance seems to be relatively worse than *Run01*'s.

As we mentioned earlier, data pre-processing takes an important role in building a deep-neural network model. Through our experiments, in the case of less training data, the augmented dataset helps us improve the performance of deep-neural network model. *Run03* and *Run05* show impressive results comparing to the first two runs. This implies that training on our re-labeled development set provides better models.

On the other hand, using the Residual neural network cannot classify efficiently the two classes *esophagitis* and *normal-z-line*. The same problem also occurs between the *dyed-resection-margins* and *dyed-lifted-polyps* classes. It can be observed in the confusion matrices of the two pairs (Figure 1.5). Therefore, these are the two main reasons which mainly bring negative impact to our results. In Figure 1.4, selected samples from *esophagitis* and *normal-z-line* are given in or-

der to illustrate the fine-grain situation of these classes that is challenging even for human without special knowledge to distinguish them correctly.

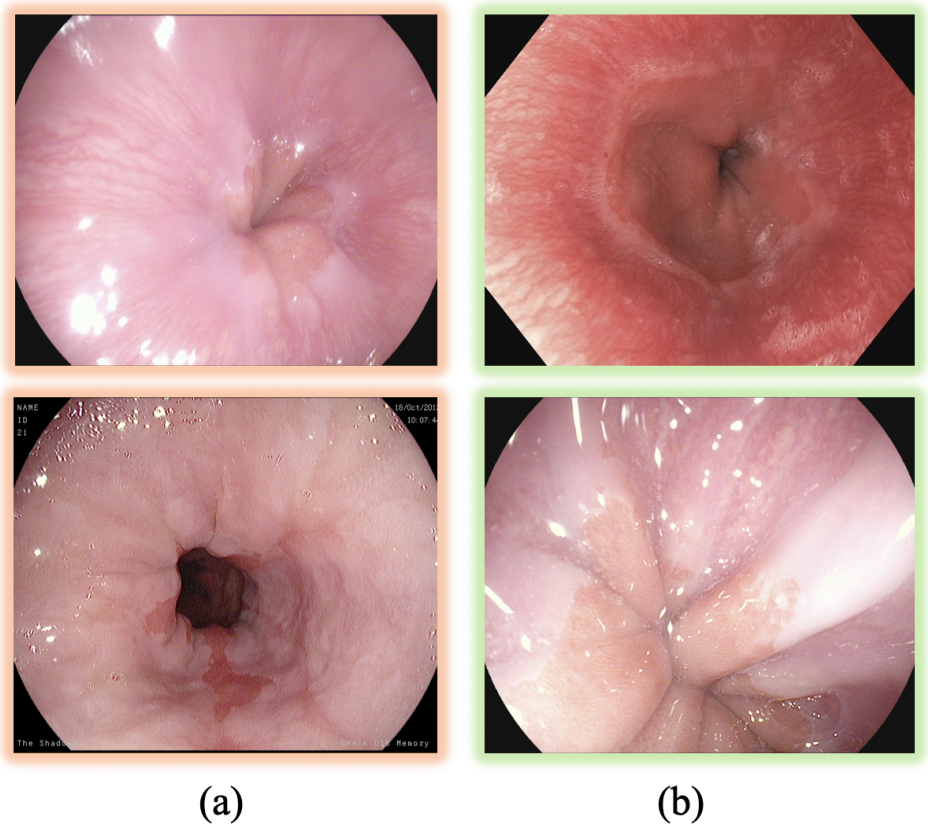


Figure 1.4: Example of confusing cases taken from the Development Set (ground-truth provided). *esophagitis* samples are marked with red box and *normal-z-line* samples with green box. Two samples from the first row are harder to be distinguished than those in the second row

Additionally, as we mentioned in section 3, the configuration of Run05 intuitively prefers *esophagitis* to *normal-z-line*, which may leads to an increasing of the false-positive cases in the result.

By comparison to the others, *Run04* has the lowest precision since it uses 75% of training data. Decreasing the amount of training samples of course affects the

performance in deep-learning models. Nevertheless, the result is still acceptable when it decreases only a few percentages and its configuration is as same as *Run03*. This is an evidence that we are even able to reduce up to 50% of data when the less training time is preferred over the accuracy.

Last but not least, we performed training and testing our model on a Tesla K80 GPU. The average inference time for all images in the test set we measured for Run01 is around 150ms per image and around 43ms for others.

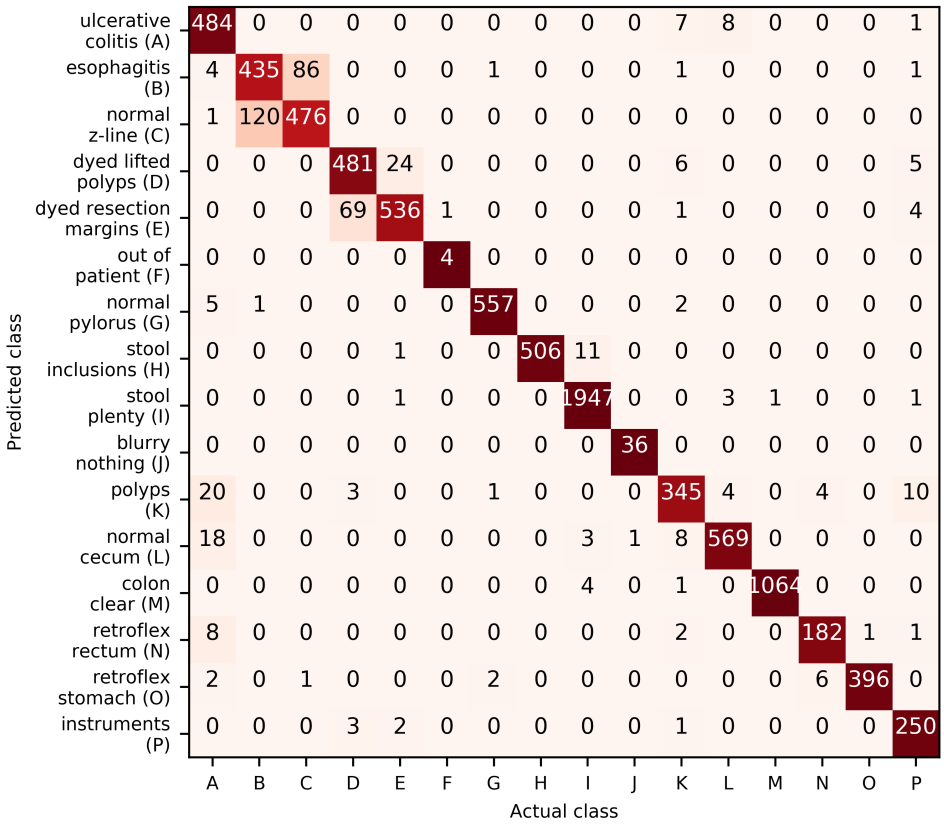


Figure 1.5: The confusion matrix of our best run - *Run03*.

Table 1.2: Medico: The 2018 Multimedia for Medicine Task challenge result

Rank	Authors	ACC	MCC	Affiliation
1	**Hoang et al.**	**0.9933**	**0.9424**	**University of Science, VNU-HCM Eurecom, France University of Information Technology, VNU-HCM**
2	Thambawita et. al.	0.9421	0.9932	Simula Research Laboratory, Norway Oslo Metropolitan University University of Oslo
3	S. Hicks et. al.	0.9350	0.9920	Simula Research Laboratory University of Oslo Simula Metropolitan Center for Digital Engineering
4	R. J. Borgli et al.	0.9310	-	Simula Research Laboratory, Norway
5	T. H. Ko et. al.	-	0.9471	University of Hong Kong, HKSAR, China Hong Kong Baptist University, HKSAR, China
6	Kirkerød et al.	0.9110	0.9890	Oslo Metropolitan University University of Oslo
7	D. Dias et al.	-	0.9880	University of Campinas, Brazil
8	M. Taschwer et al.	0.8942	0.9876	Klagenfurt University (AAU), Austria Florida Atlantic University (FAU), USA
9	Z. Khan, A. Tahir	0.7560	0.9790	National University of Computer and Emerging Sciences, Karachi Campus, Pakistan
10	M. Steiner et. al.	0.5473	0.9469	Alpen-Adria-Universität Klagenfurt, Austria SimulaMet, Norway University of Oslo, Norway

Table 1.1: The official evaluation result of HCMUS team for both sub-tasks (provided by the organizers) and speed (fps) on Tesla K80 GPU

RunID	PREC	REC	ACC	F1	MCC	RK	FPS
Run01	94.245	94.245	99.281	94.245	93.861	93.590	6.589
Run02	93.959	93.959	99.245	93.959	93.556	93.273	**23.191**
Run03	**94.600**	**94.600**	**99.325**	**94.600**	**94.240**	**93.987**	23.148
Run04	93.043	93.043	99.130	93.043	92.579	92.257	22.654
Run05	94.508	94.508	99.314	94.508	94.142	93.884	21.413

In this chapter, we describe the results of our proposed approaches.

Through experiments on endoscopic image dataset, we have shown that simple samples augmentation mechanism can bring impressive result. Besides, utilizing the advantages of several deep learning models we can increase the overall performance of our system instead of using each of them separately.

CHAPTER 2

EXPERIMENTAL RESULTS FOR NUCLEAR SEGMENTATION PROBLEM

In this chapter, we describe the Hematoxylin and Eosin (H&E) stained histopathology images dataset taken from Multi-organ Nuclei Segmentation Challenge 2018 as well as the evaluation metrics we used to evaluate our proposed methods. We also discuss more about the results acquired in each step from our algorithm, which shows the merit of each component. Moreover, the ablation study are conducted to show the improvement of our proposed enhancements.

2.1 Data set and evaluation metrics

2.1.1 Hematoxylin and Eosin (H&E) stained histopathology images dataset

2.1.1.1 Overview

As mentioned in [**?**], a lot of information about the health of the tissue to a pathologist could be revealed through histologic structure of a tissue which contributes to more general features such as shape, size, color, and crowding of glands, as well as various nuclei in epithelium and stroma. However, there are many challenges to process a whole-slide images of various type of organs as well as level of disease. "The combination of hematoxylin and eosin, or H&E, is a ubiquitous, general,and inexpensive staining (dyeing) scheme. Hematoxylin renders nuclei dark blueish purple and epithelium light purple, while eosin renders stroma pink. Together, H&E enhance the contrast between nuclei, epithelium, and stroma for examination under a microscope." This dyeing scheme makes the WSIs process-able with low cost, enables the primary diagnosis. Figure 2.1 visualizes some examples from different organs. As mentioned, nuclear regions

are enhanced by hematoxylin while pink ones are results from eosin. The contrast between nuclear and background are better after going through this dyeing scheme.

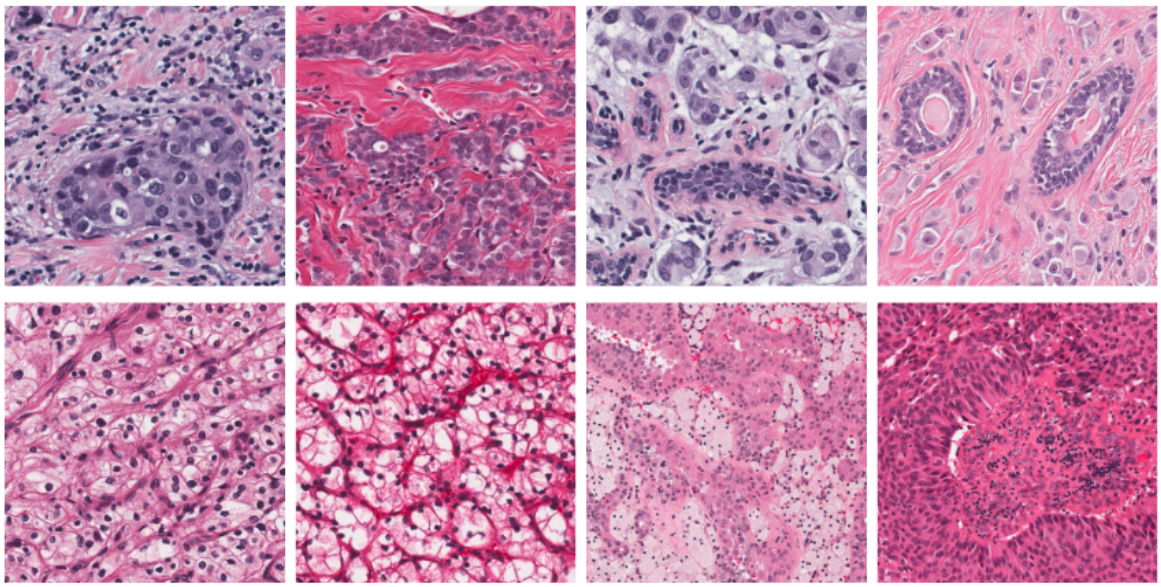


Figure 2.1: Examples of Hematoxylin and Eosin (H&E) stained histopathology images dataset. The first row contains images from breast while the last row are from kidney.

2.1.1.2 Dataset construction

We evaluated our enhanced U-Net architecture and proposed training data augmentation approach on the data set curated by [**?**], which consists of 30 histopathology images with accompanying full segmentation masks. The images are 1000×1000 pixels in size and were extracted from WSIs from unique TCGA samples, from a variety of different organs. To comprise the training set, four images were randomly selected from breast, liver, kidney and prostate samples. The remaining 14 images were from 7 different organs and were evenly split into validation and testing sets.

2.1.1.3 Three-class ground-truth generation

As mentioned in chapter **??**, we formulate nuclear segmentation as a pixel labeling problem with three potential labels, namely, nuclear interior, nuclear boundary, and background. However, there are a gap between the annotation scheme and the ground-truth we need to formalized to train our proposed network. In more details, the label provided in the dataset containing index ($id_{i,j}$) which indicates which nuclear the pixel (i, j) belongs to. $id_{i,j} = 0$ is the background region. We define the nuclear boundary pixel (i, j) by considering ones having $id_{i,j} > 0$. If there are other pixel (u, v) in a small windows which center at (i, j) that $id_{u,v} > 0$ and $id_{u,v} \neq id_{i,j}$, then we consider pixel (i, j) as nuclear boundary. For other cases, if $id_{i,j} = 0$, pixel (i, j) is background, otherwises, it is nuclear interior. We choose windows size equals to 3. Figure 2.2 visualizes our defined ground-truth used to train our proposed neural network.

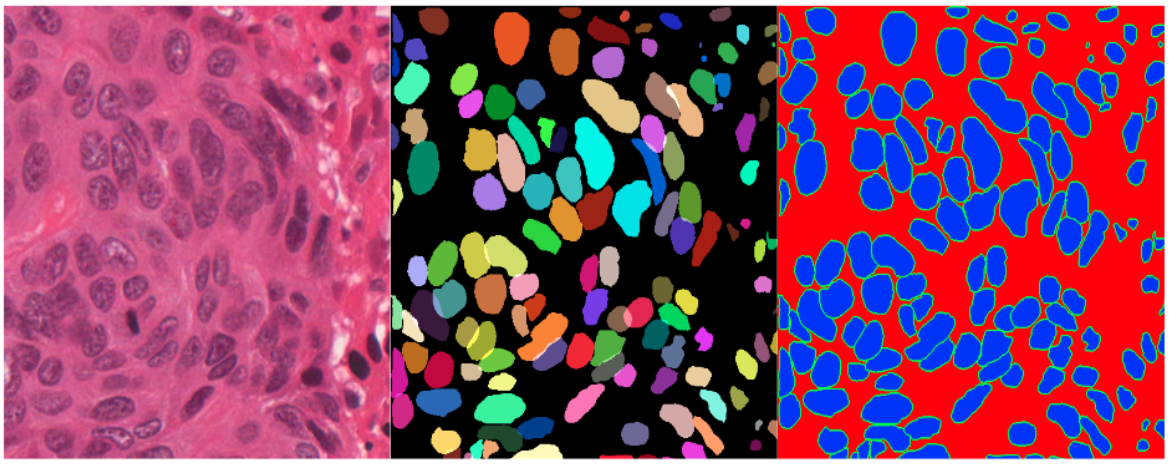


Figure 2.2: Proposed three-class ground-truth. The first image in input image. The middle is annotation from [**?**]. Each color indicates different nuclear regions. The last image is the generated three-class ground-truth. Red region is background, blue is nuclear interior while green is nuclear boundary. This results are further used to train our proposed neural network.

2.1.2 Evaluation metrics

For evaluation, we focused on the AJI metric, proposed in [**?**], which balances detection accuracy with the accuracy of the delineated boundaries of nuclei, though we also considered the F1-score & Dice coefficient to shed more light upon the performance of the various architectures.

2.2 Experimental setup

Our U-Net architectures operate on patches of size 128×128 (due to size limitations of the GPU), so to generate a set of training patches, we extracted random patches from each image during training. Before extracting the patch, the original image was rotated and scaled by random amounts. The network was trained to minimize the cross entropy loss plus the generalized Dice loss [**?**], which helps specifically to learn sharper boundaries by addressing the class imbalance problem of boundary pixels. We used the Python library NiftyNet [**?**] for an implementation of the Dice loss. We adopt the weight initialization proposed in [**?**]. In our experiments, we found that our enhanced U-Net works better without dropout layers, so we removed them, since they only increased the number of training steps required to converge. We used the Adam [**?**] optimizer for learning. The learning rate was set constant as $5e-5$ during the training process because of the adaptive property of Adam optimization [**?**]. Our batch size was restricted to be four, due to memory limitations on the GPU. We trained each network for roughly 300,000 steps. We used the validation set to fine-tune parameters.

During test-time inference, to generate a complete segmentation map, since the images are larger than 128×128 pixels, we perform inference on overlapping patches, with an overlap of 62 pixels, and then merge the results. We use the reflection transformation to pad patches on the boundaries of the original image.

2.3 Experimental results

2.3.1 Synthetic dataset

To aid in training, for each image, we generate 25 new synthetic images and their corresponding masks by our proposed method, yielding a total of 750 new synthetic images. Figure 2.3 visualizes our new dataset. As can be seen in figure 2.3, there are not too much different between the original image and the synthetic one, which can improve the performance of our proposed U-Net architectures. The details are given more in section 2.3.6.

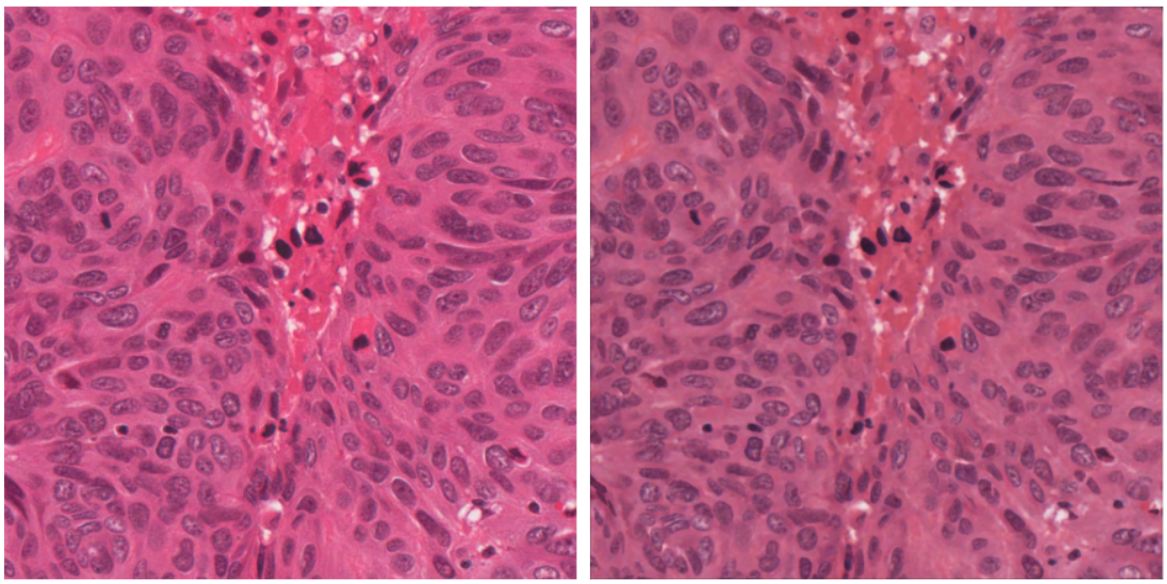


Figure 2.3: The proposed synthetic data. Left image is the original image from [**?**] dataset, the right one is our generated image.

2.3.2 Stain normalization

As mentioned in section **??**, to reduce the uninformative and possibly confusing color variation inherent to H&E stained images, we use the structure-preserving color normalization method proposed in [**?**]. Figure 2.4 visualizes some results

from this step. The color variance in the H&E stained histopathology images could be reduced by converting all the color space of images in dataset into same color space of a chosen one. We choose color space of image *TCGA-18-5592-01Z-00-DX1* as the target space.

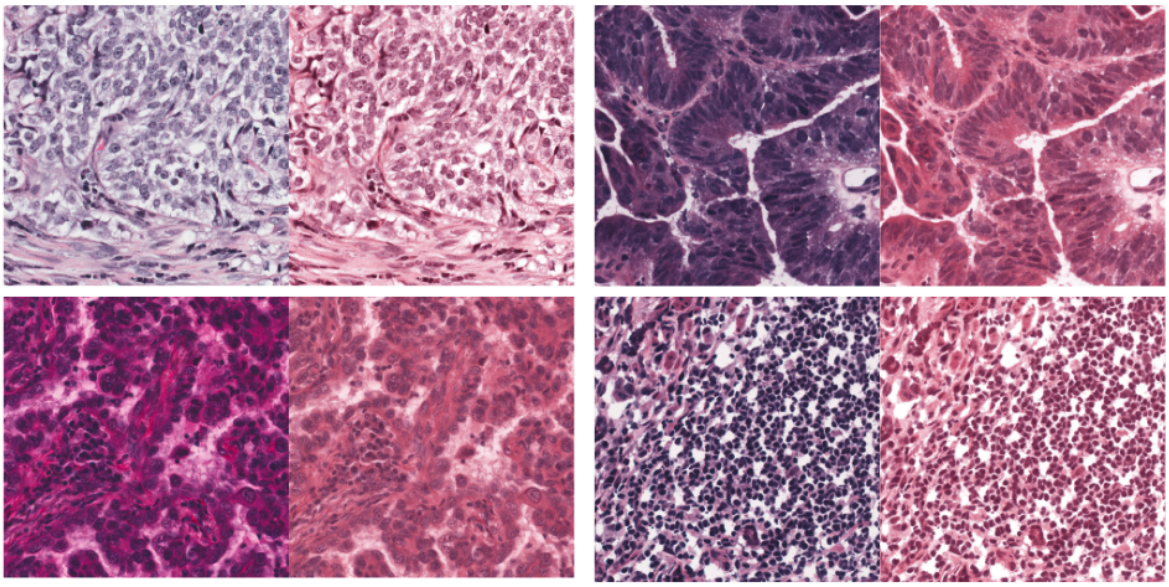


Figure 2.4: Stain normalization results. For each pair of image, the left one is the original image while the right one is the normalized image.

2.3.3 Post-processing results

The results from our proposed U-Net only contain 3 class namely, nuclear interior, nuclear boundary, and background. The post-processing step is needed to assign the identify to each region in the image. Though we formalize our problem as mentioned three-class nuclear segmentation, there are much challenges to completely separate the overlapping nuclear regions. Figure 2.5 visualize some of this mentioned case. With the help of post-processing step, we can alleviate this problem by applying proposed combination of morphology operations. The indexes are generated and visualized as different color in figure 2.5.

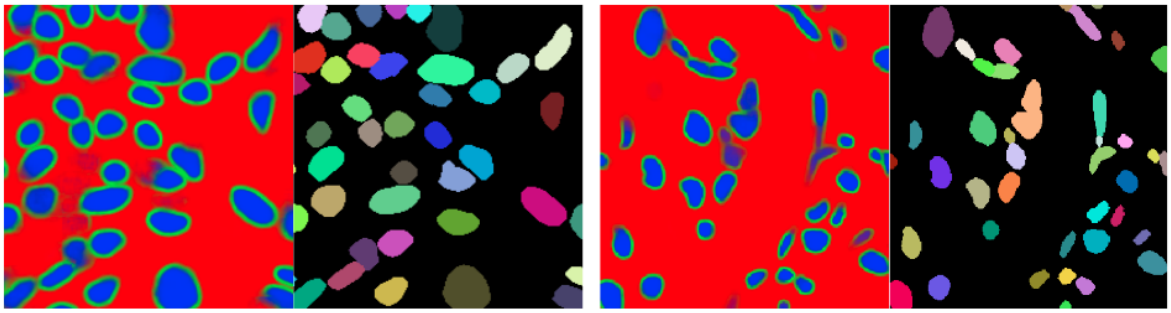


Figure 2.5: The U-Net results and their corresponding post-processing results.

2.3.4 Fully-enhanced network

Fig. 2.6 shows the results of our fully-enhanced U-Net, trained with augmented images, on several example histopathology images from the held-out test set, and the first row of Table 2.2 shows its performance according to the aforementioned metrics. As can be seen in Fig. 2.6, the boundary between overlapping nuclei can be reasonably separated by our method. Moreover, our method is able to generalize well to other types of organs, even those for which it was not trained. The first two lines in Fig. 2.6 are from stomach and colon tissue, respectively, which tissue types were not in the training set, yet our method still produces strong results. On the entire test set, our method achieved an AJI of 0.629. For comparison, the method proposed in [**?**] achieved an AJI of 0.508 on the same data set, although this comparison is not definitive, since 7 fewer images from the data set were used for training in their experiments.

Processing an entire image of size 1000×1000 on a single TITAN V 12GB GPU with our enhanced U-Net architecture takes only about 17 seconds. The subsequent post-processing incurs an additional 2 seconds on a CPU to create the final result.

2.3.5 MoNuSeg Challenge result

Two of three our enhancements, including new generated synthetic data and additive residual blocks along the long skip connection of U-Net architecture, are used to participate in MoNuSeg Challenge 2018. There are some differences of the solution we used to participate in this challenge. In the details, we extract patches of size 256×256 from augmented, pre-processed images for training data. For validation, we use a subset of the original images, one of each organ, leaving the other 23 images for training. Rotation and scaling are used when extracting these patches. Other configurations are kept as description in section 2.2. Table 2.1 shows our result in the competition. With two proposed enhancements, we rank 11^{th} among 32 participants around the world, with 65.65% as our performance.

2.3.6 Ablation study

We further evaluated the effect of each proposed component to the performance of our method through an ablation study. Example resulting segmentations from each of the following experiments are shown in Fig. 2.7 and the performance according to the aforementioned metrics are given in Table 2.2.

Residual blocks on long-skip connections

In this experiment, we removed all residual blocks on the long-skip connections of the U-Net architecture. To make a fair comparison, we compensated by adding more residual blocks in the encoder and decoder components to maintain the number of trainable parameters. All of other hyper-parameters were kept constant. The first and third line in Table 2.2 show the comparison between the two different models. Without residual blocks on the long-skip connection, the AJI value decreased by 1.4%, and the Dice score and F1-score decreased by 1.3% and

Table 2.1: MoNuSeg Challenge 2018 result.

Rank	Team	AJI	Affiliation	Country
1	CUHK& IMSIGHT	0.6907	The Chinese University of Hong Kong; Imsight Technology	China
2	BUPT.J.LI	0.6868	Beijing University of Post and Telecommunication	China
3	pku.hzq	0.6852	Peking University	China
4	Yunzhi	0.6788	University of Oklahoma	USA
5	Navid Alemi	0.6779	University of Warwick	UK
6	xuhuaren	0.6642	Shanghai Jiao Tong University National University of Defense Technology	China
...	...	...	...	...
11	**CMU-UIUC-HCMUS**	**0.6557**	**Carnegie Mellon University; University of Illinois at Urbana-Champaign; University of Science, VNU-HCM**	**USA Vietnam**
12	Graham&Vu	0.6532	University of Warwick, Sejong University	UK
13	Unblockabulls	0.6514	American Express, India	India
14	Tencent AI Lab	0.6459	University of California, Berkeley; The Hong Kong University of Science and Technology	USA
...	...	...	...	...
29	VISILAB	0.4441	Visilab Research Group, University of Castilla-La Mancha, Ciudad Real, Spain	Spain
30	Sabarinathan	0.4437	Cognizant Technology Solutions	India
31	Silvers	0.278	Xiamen University	China
32	TJ	0.1301	Tongji University	China

Table 2.2: Quantitative comparison of each proposed component of the U-Net architecture and training procedure. -D.A: without our proposed synthetic data; -R.S: without residual blocks on the long-skip connections; -G.E.: the U-Net without group-equivariant operations.

Method	AJI	F1-score	Dice's coef.	#Params
Ours	**0.6291**	0.8469	**0.7980**	102M
Ours - D.A	0.6019	0.8006	0.7796	102M
Ours - R.S	0.6151	0.8349	0.7846	101M
Ours - G.E.	0.6125	**0.8490**	0.7893	101M

1.2%, respectively. This strengthens the credibility of our hypothesis that residual blocks on the long-skip connections help the network extract richer low-level features and thereby aid the network in delineating nuclear boundaries of touching nuclei. Fig. 2.7c and 2.7d further visualize the results of the two architectures. Some nuclei in Fig. 2.7d evidence a difficulty in separating their overlap, while our architecture with long-skip residual blocks successfully separates them.

New synthetic data

To see the effect of the new synthetic data we proposed, we withheld these synthetic images from the training set and trained a separate network on this reduced training set. The architecture of the network and all of other hyper-parameters were kept the same. As can be seen in Table 2.2, training with these new synthetic images can improve the model's performance by approximately 2.7% in AJI, showing the most profound impact upon performance compared to the other two contributions. This suggests that this method of data augmentation further offsets the problem of sample scarcity for histopathological analysis, even beyond standard augmentation techniques. This is significant since labeling this type of data is extremely labor intensive.

Group equivariance

To see the effect of group-equivariant operations, we replaced these operations with standard operations for a CNN and retrained the network. Following the work in [**?**], to preserve the same number of trainable parameters, we doubled the number of filters in each convolution layer in the ordinary network, while keeping the same architecture otherwise. This new model was trained with a similar set of hyper-parameters, except that we increased the learning rate to $1e-4$ and we were able to use a batch size of eight, since the network did not require as much storage on the GPU as the group-equivariant version. Since there were no group-equivariant operations, we added dropout to the model to help with regularization. As shown in Table 2.2, both the AJI and Dice's coefficient drop by a significant margin, roughly 1.7% and 0.9% respectively, without encoded group equivariance. This implies that integrating group-equivariant convolution and operations into the current U-Net architecture can indeed enable the network to learn better parameters that generalize well to simple transformations, namely, translations and rotations. Although, since it is only equivariant to rotations of 90 degrees, the network can still benefit from data augmentation of rotations of finer, arbitrary angles.

In this chapter, we describe the results of our proposed method. Through carefully designed experiments as well as ablation study, we have shown the value of several enhancements to the standard U-Net architecture, namely, encoding rotation and translation equivariance and adding additional residual blocks, and our novel data augmentation method for automated nuclear segmentation in histology images.

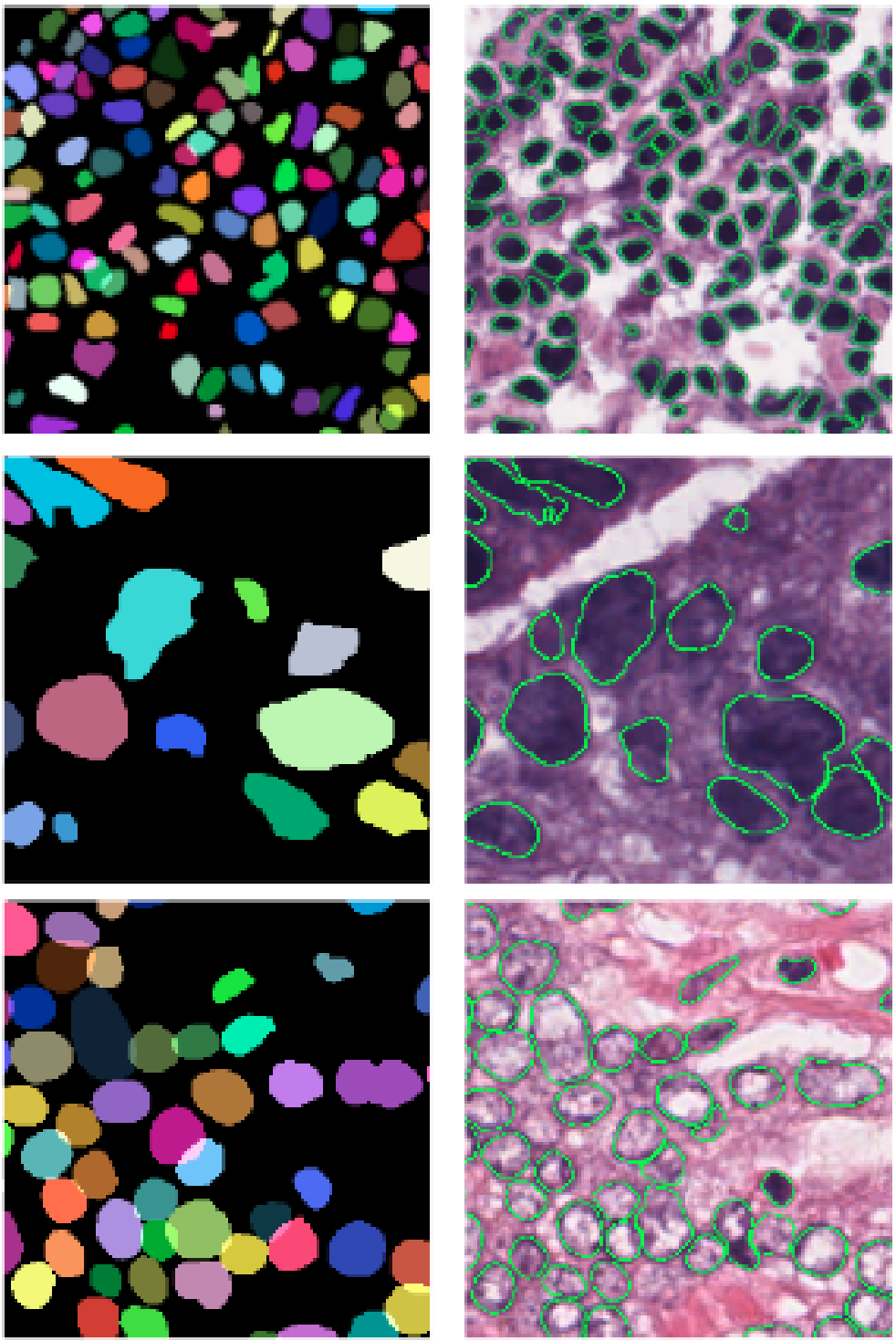


Figure 2.6: Segmentation results of our fully-enhanced U-Net on example test histopathology images. *Left.* Ground-truth annotations. *Right.* Our results. The estimated nuclear boundary is visualized in green. Patches are cropped for better visualization. Best view in color.

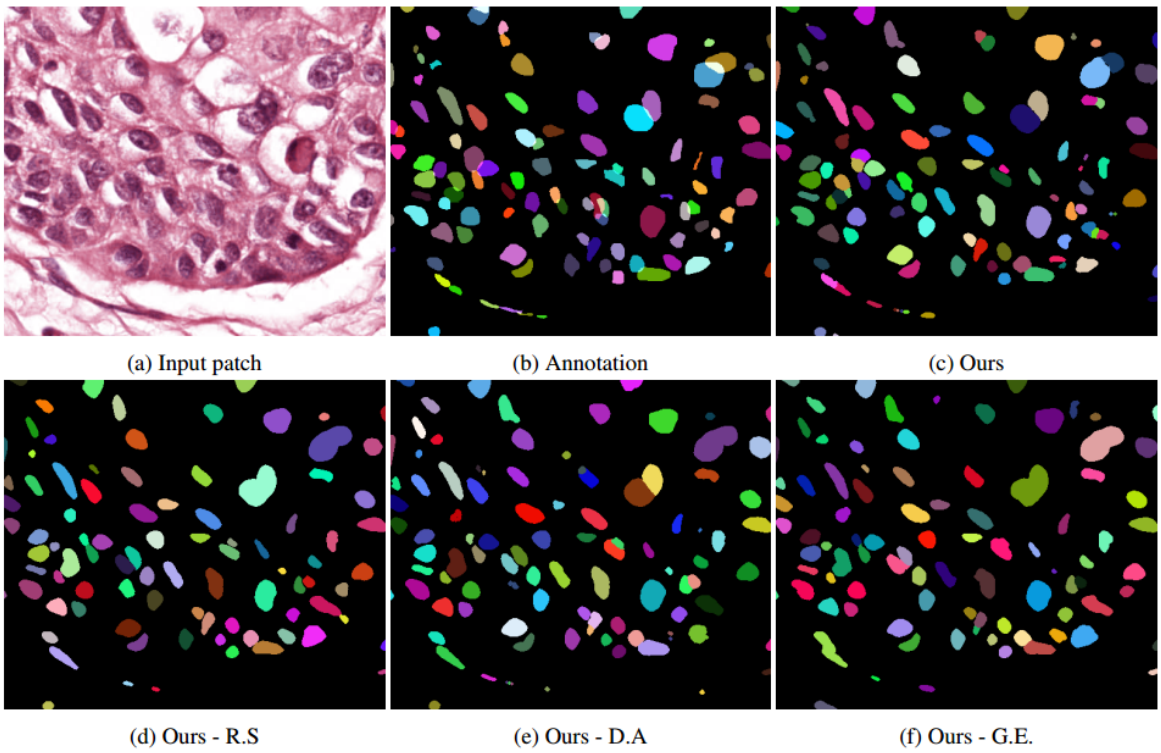


Figure 2.7: The visualization of example results of different methods in the ablation study. The segmented region for each distinct detected nucleus is shown by a unique color. The background is visualized in black. Patches are cropped for better visualization. Best view in color. -D.A: without our proposed synthetic data; -R.S: without residual blocks on the long-skip connections. -G.E.: the U-Net without group-equivariant operations.

APPENDIX

An application of Residual Network and Faster - RCNN for Medico: Multimedia Task at MediaEval 2018

Trung-Hieu Hoang[1], Hai-Dang Nguyen[2], Thanh-An Nguyen[1],
Vinh-Tiep Nguyen[3], Minh-Triet Tran[1],
[1] Faculty of Information Technology, University of Science, VNU-HCM, Vietnam [2] Eurecom, France
[3] University of Information Technology, VNU-HCM, Vietnam
{hthieu,ntan}@selab.hcmus.edu.vn,nguyenhd@eurecom.fr,tiepnv@uit.edu.vn,tmtriet@fit.hcmus.edu.vn

ABSTRACT

The Medico: Multimedia Task focuses on developing an efficient framework for predicting and classifying abnormalities in endoscopic images of gastrointestinal (GI) tract. We present the HCMUS Team's approach, which employs a combination of Residual Neural Network and Faster R - CNN model to classify endoscopic images. We submit multiple runs with different modifications of the parameters in our combined model. Our methods show potential results through experiments.

1 INTRODUCTION

Medico: Multimedia Task at MediaEval 2018 challenge [4] aims to bring new achievements in computer vision, image processing and machine learning to the next level of computer and multimedia assisted diagnosis. The goal of the challenge is to predict abnormalities and diseases in an efficient way with as less training data as possible [5]. The task organizers also provide a priority list for the classes in other to accommodate with the single-class classification challenge. Thus, this leads to some modifications of our model, which are meticulously described in section 3.

In our approach, we introduce a stacked model consisting of two deep networks, a Residual Neural Network (Resnet) [2] followed by a Faster Region-based Convolutional Neural Network (Faster R-CNN) [7]. Since Resnet mostly focuses on deep global features of image, it fails to classify images that symptoms of abnormal diseases or instruments appear as small objects on diversity backgrounds. Therefore, this is the reason of using Faster R-CNN to re-classify the images of some classes that Resnet usually mis-classify.

2 RELATED WORK

In the field of medical image processing, deep neural networks have been used in order to solve several problems related to endoscopic images of the gastrointestinal (GI) tract. Particularly, to localize and identify polyps within real-time constraint, deep CNNs has recently shown an impressive potential when achieving up to 96.4% accuracy - published in 2018 by Urban G et al. [9]. Another interesting article of Satoki Shichijo et al. [8] also applies multiple deep CNNs to diagnose Helicobacter pylori gastritis based on endoscopic images. Further, gastrointestinal bleeding detection using deep CNNs on endoscopic images has been successfully done and published by Xiao Jia et al. [3].

MediaEval'18, 29-31 October 2018, Sophia Antipolis, France

3 APPROACH

3.1 Dataset Preparation

3.1.1 Disease region localization. In order for the Faster R-CNN model to be trained, objects in the image have to be tagged with bounding boxes and passed to the model as input. We annotate the signal of disease in all images of the following classes: *dyed-resection-margins, dyed-lifted-polyps, instruments* and *polyps*.

3.1.2 Re-labeling Medico development dataset. After training with the development set, we find some training samples with inappropriate labels according to the priority list. Therefore, in order for our model to learn with the least confusing, we apply the new labels, predicted by the trained model, to these images.

3.1.3 Instruments dataset augmentation. Instruments - the second highest priority class has only 36 images with the limitation of background context in the development set. In order to maintain the balancing between all of the classes and also improve the diversity of the *instruments* images, we generate more images for the *instruments* based on the current given development set by placing the instruments on the foreground of other diseases backgrounds.

Among the 36 *instruments* images, we carefully select 24 of them and crop the instruments along their edges. Then, we randomly select 20% of the images from *dyed-lifted-polyps, dyed-resection-margins*, *ulcerative-colitis* classes, and use them as the background of the cropped instruments. By applying this method, we are able to generate more than 800 images for the *instruments* class.

3.2 Method

3.2.1 Fine-tuning deep neural network for medical images. In our approach, both Residual Network with 101 layers and Faster R-CNN [1] (both pre-trained on ImageNet) are fine-tuned by using our modified development dataset. In term of using convolution neural network for medical images, knowledge transferring from natural images to medical images is possible, even though there is a large difference between the source and target databases. It is especially useful in the case of small dataset of images provided [6]. Our experiment results also support this idea. Fine-tuning on the ImageNet pre-trained model significantly improves the efficient of classification model.

3.2.2 First run. Residual network with 101 layers model are fine-tuned on the original development set provided by the task organizers along with our instruments increased dataset. After passed through Resnet101, output images classified as special classes become the input of Faster R-CNN network, which is trained for detecting instruments in images.

- First case: Images predicted as *instruments* by Resnet101 are double-checked. In case instruments are not detected by Faster R-CNN in those images, they are re-labeled as the class of their second highest score proposed by Resnet101.
- Second case: Images predicted as *dyed-lifted-polyps, dyed-resection-margins, ulcerative colitis* by Resnet101 are fed forward through Faster R-CNN network to detect instruments. They are classified as *instruments* if detected or keep the original prediction otherwise.

3.2.3 Second run. Feeding forward a large number of images in the three classes through Faster R-CNN causes a bottle-neck of inference time, as Faster R-CNN has high time complexity. Therefore, in this second run, we limited the images passed through Faster R-CNN by only performing the first case of the first run.

3.2.4 Third run. The configuration of the third run is as same as the second run. Instead of using the original training set mentioned in the first run, we train our model on the re-labeled development set combined with the augmented instrument set.

3.2.5 Forth run. In this run, we reduce the number of images used for training by selecting randomly 75% images of each class in the same training set as the third run. Other processing steps are also configured in the same way.

3.2.6 Fifth run. Throughout our experiments, *normal-z-line* and *esophagitis* are the top most confusing classes not only for Resnet101 but also for human to distinguish them. In the priority list, *esophagitis* has a higher rank than *normal-z-line's*. Thus, after several times evaluating our model on the development dataset, we propose a condition for these two classes when they are predicted by Resnet101. As Resnet101 provides a probability distribution over the 16 classes for each image, whenever the *normal-z-line* appears to be the highest class, we add a small bias 0.3 to the probability of the *esophagitis*. Hence, the model is more likely to emit the *esophagitis* class. This intuitively means that our model prefers *esophagitis* to *normal-z-line* when it is confused between these classes.

4 RESULTS

Table 1: Official evaluation result for both sub-tasks (provided by the organizers) and speed (fps) on Tesla K80 GPU

RunID	PREC	REC	ACC	F1	MCC	RK	FPS
Run01	94.245	94.245	99.281	94.245	93.861	93.590	6.589
Run02	93.959	93.959	99.245	93.959	93.556	93.273	**23.191**
Run03	**94.600**	**94.600**	**99.325**	**94.600**	**94.240**	**93.987**	23.148
Run04	93.043	93.043	99.130	93.043	92.579	92.257	22.654
Run05	94.508	94.508	99.314	94.508	94.142	93.884	21.413

There is a trade-off between speed and accuracy when comparing the result of Run01 and Run02. In Run02, we reduce a large number of images passing through Faster R-CNN for the sake of time, so its performance seems to be relatively worse than Run01's.

As we mentioned earlier in section 3, data pre-processing takes an important role in building a deep-neural network model. Through our experiments, in the case of less training data, the augmented dataset helps us improve the performance of deep-neural network model. Run03 and Run05 show impressive results comparing to

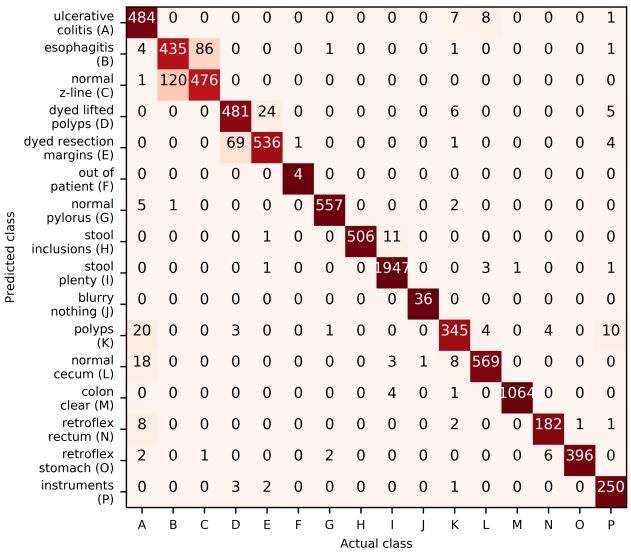


Figure 1: Confusion matrix of our best run - *Run03*

the first two runs. This implies that training on our re-labeled development set provides better models.

On the other hand, using the Residual neural network cannot classify efficiently the two classes *esophagitis* and *normal-z-line*. The same problem also occurs between the *dyed-resection-margins* and *dyed-lifted-polyps* classes. It can be observed in the confusion matrices of the two pairs (Figure 1). Therefore, these are the two main reasons which mainly bring negative impact to our results.

Additionally, as we mentioned in section 3, the configuration of Run05 intuitively prefers *esophagitis* to *normal-z-line*, which may leads to an increasing of the false-positive cases in the result.

By comparison to the others, Run04 has the lowest precision since it uses 75% of training data. Decreasing the amount of training samples of course affects the performance in deep-learning models. Nevertheless, the result is still acceptable when it decreases only a few percentages and its configuration is as same as Run03. This is an evidence that we are even able to reduce up to 50% of data when the less training time is preferred over the accuracy.

5 CONCLUSION AND FUTURE WORKS

Medico image classification is a challenging problem because of the fine-grained images, less training data and require high accuracy. In our current approach, we focus on training a combination of Residual Neural Network and Faster R-CNN with different modifications of the training set. Additionally, object detection method is applied to detect small symptoms of diseases, which are useful evidences for the classification task. Accuracy and inference time that we reach is acceptable and appropriate for real-time constraint. However, for future works, we need a more robust approach to exploit the distinction between easy-confused classes, e.g, *esophgitis* and *normal-z-line*, or *dyed-lifted-polyps* and *dyed-resection-margins*.

REFERENCES

[1] Xinlei Chen and Abhinav Gupta. 2017. An Implementation of Faster RCNN with Study for Region Sampling. *arXiv preprint arXiv:1702.02138* (2017).

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Jun 2016). https://doi.org/10.1109/cvpr.2016.90

[3] Xiao Jia and Max Q.-H. Meng. 2016. A deep convolutional neural network for bleeding detection in Wireless Capsule Endoscopy images. *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (2016). https://doi.org/10.1109/embc.2016.7590783

[4] PËŽal Halvorsen Thomas de Lange Kristin Ranheim Randel Duc-Tien Dang-Nguyen Mathias Lux Olga Ostroukhova Konstantin Pogorelov, Michael Riegler. 2018. Medico Multimedia Task at MediaEval 2018. *Media Eval' 2018* (2018).

[5] Konstantin Pogorelov, Kristin Ranheim Randel, Carsten Griwodz, Sigrun Losada Eskeland, Thomas de Lange, Dag Johansen, Concetto Spampinato, Duc-Tien Dang-Nguyen, Mathias Lux, Peter Thelin Schmidt, Michael Riegler, and Pål Halvorsen. 2017. KVASIR: A Multi-Class Image Dataset for Computer Aided Gastrointestinal Disease Detection. In *Proceedings of the 8th ACM on Multimedia Systems Conference (MMSys'17)*. ACM, New York, NY, USA, 164–169. https://doi.org/10.1145/3083187.3083212

[6] Adnan Qayyum, Syed Anwar, Muhammad Majid, Muhammad Awais, and Majdi Alnowami. 2017. Medical Image Analysis using Convolutional Neural Networks: A Review. 42 (09 2017).

[7] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett (Eds.). Curran Associates, Inc., 91–99.

[8] Aoyama Kazuharu Nishikawa Yoshitaka Miura Motoi Shinagawa Takahide Takiyama Hirotoshi Tanimoto Tetsuya Ishihara Soichiro Matsuo Keigo Tada Tomohiro Shichijo Satoki, Nomura Shuhei. 2017. Application of Convolutional Neural Networks in the Diagnosis of Helicobacter pylori Infection Based on Endoscopic Images. *EBioMedicine* 25 (01 Nov 2017), 106–111. https://doi.org/10.1016/j.ebiom.2017.10.014

[9] Gregor Urban, Priyam Tripathi, Talal Alkayali, Mohit Mittal, Farid Jalali, William Karnes, and Pierre Baldi. 2018. Deep Learning Localizes and Identifies Polyps in Real Time With 96% Accuracy in Screening Colonoscopy. *Gastroenterology* 155, 4 (2018). https://doi.org/10.1053/j.gastro.2018.06.037

Enhanced Rotation-Equivariant U-Net for Nuclear Segmentation

Benjamin Chidester[1,*], That-Vinh Ton[2,3,*], Minh-Triet Tran[2], Jian Ma[1], Minh N. Do[3]
[1]Carnegie Mellon University [2]Univ. of Science, VNU-HCM
[3]Univ. of Illinois at Urbana-Champaign
{bchidest, jianma}@cs.cmu.edu ttvinh@selab.hcmus.edu.vn
minhdo@illinois.edu tmtriet@fit.hcmus.edu.vn

Abstract

Despite recent advances in deep learning, the crucial task of nuclear segmentation for computational pathology remains challenging. Recently, deep learning, and specifically U-Nets, have shown significant improvements for this task, but there is still room for improvement by further enhancing the design and training of U-Nets for nuclear segmentation. Specifically, we consider enforcing rotation equivariance in the network, the placement of residual blocks, and applying novel data augmentation designed specifically for histopathology images, and show the relative improvement and merit of each. Incorporating all of these enhancements in the design and training of a U-Net yields significantly improved segmentation results while still maintaining a speed of inference that is sufficient for real-world applications, in particular, analyzing whole-slide images (WSIs). Code for our enhanced U-Net is available at https://github.com/thatvinhton/G-U-Net.

1. Introduction

The recent surge in interest in deep learning coupled with increasing availability of large-scale histopatholical image data sets, such as The Cancer Genome Atlas [17], has resulted in significant advances in computational histological analysis [13, 19]. A crucial step in such analysis pipelines is accurate and efficient segmentation of cell nuclei [5, 24]. With the aid of large-scale training data, deep-learning-based methods for automatically segmenting nuclei have surpassed traditional approaches, such as watershed [23] and thresholding [18] algorithms, though, despite these improvements, this step remains challenging and continues to be an active area of research [9, 26]. Changes in nuclear morphology are well-studied indicators of diseases, such as cancer, which motivates the continued development of effective methods of automated segmentation.

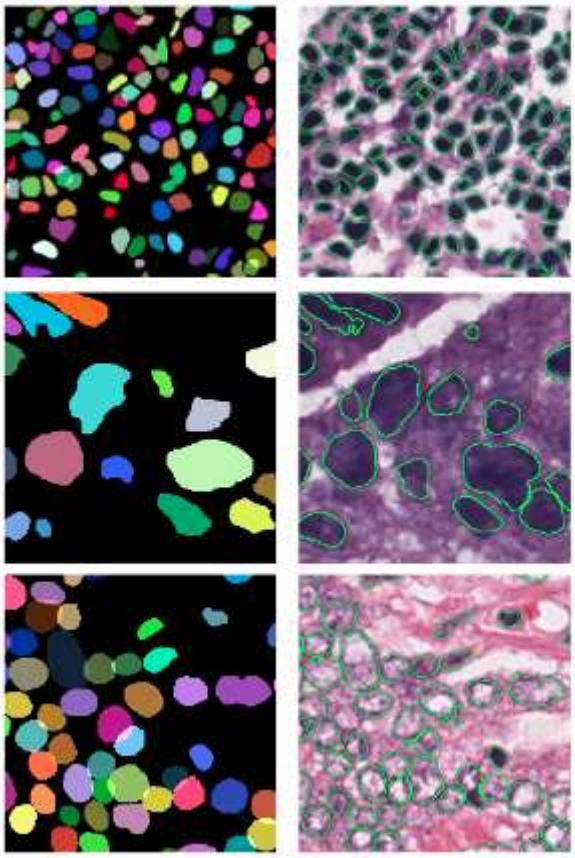


Figure 1: Segmentation results of our fully-enhanced U-Net on example test histopathology images. *Left.* Ground-truth annotations. *Right.* Our results. The estimated nuclear boundary is visualized in green. Patches are cropped for better visualization. Best view in color.

Initial approaches using deep learning operated by scanning the image patch-by-patch and generating a label (e.g., nucleus or non-nucleus) for each pixel centered in each patch [10, 25, 13]. Subsequent morphological processing

*Authors contributed equally.

could be applied to help smooth and refine the generated label masks and ensure that the segmented regions are contiguous; an example is the work of [25] that uses learned nuclear shape priors to encourage consistent inference. More recently, the U-Net architecture was proposed [19], which operates on the entire image and jointly infers the label at each pixel simultaneously, leading to more spatially coherent segmentation. U-Nets have been shown to achieve improved accuracy on several bioimage segmentation tasks, even when the data set is relatively small [19].

In this work, we bring together several recent developments in bioimage analysis to enhance the current methodology for deep-learning-based nuclear segmentation and, through a thorough ablation study, show the relative improvement and merit of each. We show that combining these enhancements together can achieve significantly improved scores for several important metrics at a speed that is sufficient for real-world applications, such as analyzing whole-slide images (WSIs).

The first enhancement to the U-Net we consider is to encode equivariance to groups, specifically rotation and translation, following the work of group-equivariant CNNs (G-CNNs) [3], thereby obviating the need to learn such equivariance through extensive and time-consuming data augmentation. This enhancement helps the learned network to better generalize to such variation in unseen data. G-CNNs have recently been shown to demonstrate top performance on segmenting regions of interest in histology images [22] and biological structures in other types of bioimages [1], but have not yet been applied to the task of nuclear segmentation.

Secondly, we enhance the long-skip connections in the U-Net from the downsampling to upsamling arms of the "U" with residual blocks, an insight that has shown improved performance for other applications [2]. The motivating hypothesis of this modification is that providing richer low-level features from the downsampling arm, learned through the residual blocks, to the upsampling arm will aid in producing detailed boundaries, especially between touching nuclei.

Lastly, we propose a novel means of data augmentation specifically designed for histological images to aid in training. Although the rate of generation of H&E image data is increasing, fully-labeled training data is still scarce. Therefore, augmentation of training data is still crucial for learning robust models. In addition to standard augmentation techniques of elastic deformations, blurring, and additive noise, we generate synthetic images by slightly translating and deforming the nuclei and filling any empty pixels by inpainting [27]. This method was inspired by recent work in video object segmentation [11], which demonstrated significant improvements in performance. Others works have also proposed means of generating synthetic, realistic histological images [16], though ours is much simpler to implement.

We demonstrate the merit of each of these considerations for designing and training U-Nets for nuclear segmentation on a data set of several histology images from TCGA samples for various types of tissue [13]. Several demonstrative results of our fully-enhanced U-Net are shown in Fig. 1. We use the aggregated Jaccard index (AJI), Dice coefficient, and F1-score as metrics for evaluation. We show that the combination of these improvements yields significant gains for segmentation.

2. Method

The standard U-Net architecture consists of two arms, one for downsampling the feature maps to a lower-dimensional space and one for upsampling the feature maps back to full resolution. Each downsampling layer consists of convolution, non-linear activation, pooling, and batch normalization. Each upsampling layer consists of similar operations except that pooling is replaced by upsampling. Additionally, residual blocks [7], which are used to increase the depth of the network, can be added. In [6], some experiments evaluated the performance of different types of residual blocks. We inherit from their work the architecture producing the best reported performance. Lastly, to help with interpolating the higher-resolution feature maps, features from the downsampling arm are conveyed to the upsampling layers by long-skip connections. From this baseline architecture, we incorporate two enhancements, namely group-equivariant operations to encode equivariance to rotation and translation, and residual blocks along the long-skip connections. We describe these enhancements below.

As in other deep-learning-based approaches [13], we formulate nuclear segmentation as a pixel labeling problem with three potential labels, namely, nuclear interior, nuclear boundary, and background. The network is designed and trained to produce a probability map for each label. Creating a label especially for the boundary results in a larger contribution to the objective for boundary pixels and thereby encourages the network to produce a more accurate boundary, which is generally harder to infer by post-processing than pixels away from nuclear boundaries. Our post-processing method of morphological operations, described below, helps to refine the output of the U-Net by smoothing edges and ensuring contiguous segmentation boundaries.

Input data is first stain-normalized before being fed to the U-Net. To help train the model, we employ several means of data augmentation, including our proposed histology-specific method, which we describe below.

2.1. Encoding group-equivariance

As noted in [3], it is helpful to think of the input image to a neural network as a function $f\colon \mathbb{Z}^2 \to \mathbb{R}^K$ that maps 2D

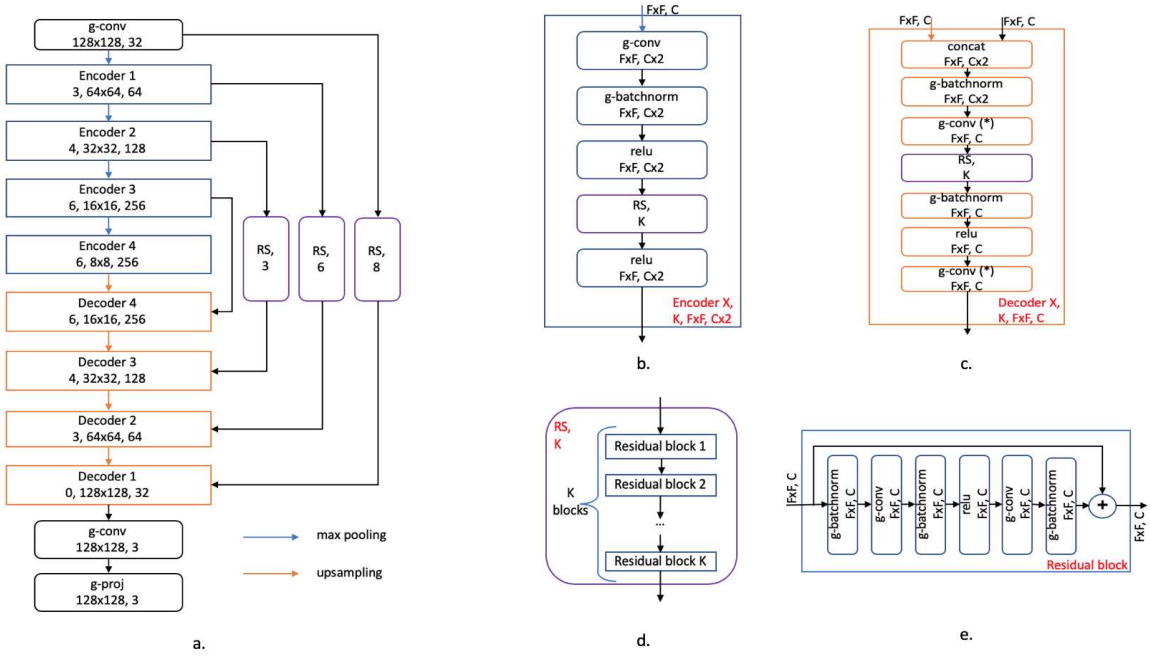


Figure 2: Our proposed rotation and translation-equivariant U-Net architecture. (a) Highest-level view, showing sequential encoder and decoder blocks. Our residual blocks along the long-skip connections of low-level layers are also shown. (b) Details of encoder blocks. (c) Details of decoder blocks. (d) Details of RS blocks, with several residual blocks in series, which are found on long-skip connections and within decoder and encoder blocks. (e) Details of residual blocks. (*) Because of limited space, we omit the batch normalization layers and ReLu activation layers after g-conv blocks.

space to pixel intensities. The insight of [3] for CNNs was to generalize equivariance to translation, which is inherent to standard convolution, to other transformations by defining convolution for *groups* in general, of which $\mathbb{Z}^2$ with translations is a specific example. An important group for histology images is the $p4$ group, which consists of translations and rotations about the origin by 90 degrees of elements in $\mathbb{Z}^2$. A group-equivariant neural network can be created by the composition of group-equivariant convolution with several other group operations, given below, which preserve equivariance to such transformations throughout the network. The placement of these operations in the network can be seen in Fig. 2.

Group-equivariant convolution

Group-equivariant convolution is the generalization of convolution to functions on groups, the set $\mathbb{Z}^2$ with translation, on which convolution is normally defined, being a specific type of group. For a group G, the convolution of a filter $\psi\colon G \to \mathbb{R}^K$ with a feature map $f\colon G \to \mathbb{R}^K$ is defined to be the sum, over all elements in G, of their inner product:

$$(f * \psi)(g) = \sum_k \sum_{h \in G} f_k(h)\psi_k(g^{-1}h). \qquad (1)$$

Here, the action of element g on $h \in G$ is expressed by gh, and $g^{-1}h$ is the action of the inverse of g. For example, if the group is the translation of elements $x \in \mathbb{Z}^2$, then $gx = x + g$ and $g^{-1}x = x - g$ and we would have standard convolution. Since the output function is a function of G, which indexes not only pixel locations, but also rotations, this information can be preserved throughout the network thereby preserving equivariance to such transformations.

Since the input images to the network are functions on $\mathbb{Z}^2$, the output of the first group-equivariant convolutional layer is a special case, given by

$$(f * \psi)(g) = \sum_k \sum_{z \in \mathbb{Z}^2} f_k(z)\psi_k(g^{-1}z). \qquad (2)$$

Group-equivariant upsampling

For the upsampling arm of the U-Net, before each layer, we first upsample the feature map from the layer below by 2 using nearest-neighbor interpolation, which preserves equivariance to translations and rotations of 90 degress. This method of upsampling is the same as *deconvolution* or *transpose* convolution with a 2×2 filter of all ones [15] and helps to keep the number of trainable filters in the network manageable.

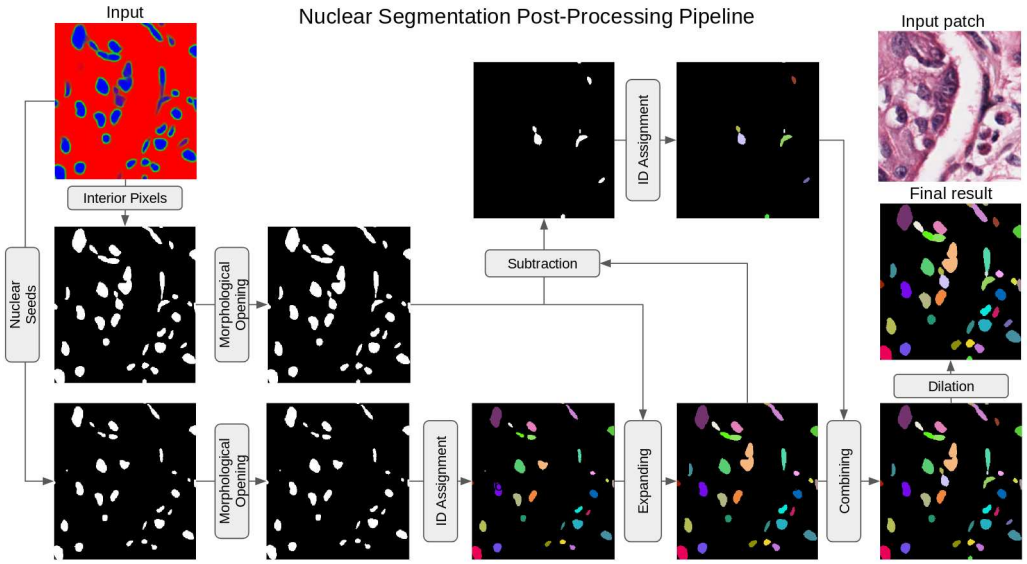


Figure 3: Our proposed post-processing method, consisting of a series of several morphological operations. Each step is visualized by its result.

Group-equivariant pooling and residual blocks

As noted in [3], for the $p4$ group, max-pooling with a stride of 2 preserves the equivariance properties of the network, and this is the pooling operation we use in the downsampling arm of the U-Net. Also noted in [3], since residual blocks are simply the addition of two group-equivariant feature maps, the output will also be group-equivariant.

Group projection

To create the final segmentation image, we must transform the domain of the feature maps in the U-Net from G back to $\mathbb{Z}^2$. To do so, we average the feature map for each filter over rotations. This is called the group *projection* layer, as in other works [14].

2.2. Long-skip connections with residual blocks

In the typical U-Net [19] architecture, the number of convolution blocks at low-level layers is small, which limits the effective local field-of-view and thereby decreases the quality of features which the network can use to delineate the output boundary. To provide richer low-level features to the final layers of the upsampling arm of the U-Net, we enhanced the baseline U-Net by adding residual blocks on long-skip connections. Our long-skip connections enhanced with residual blocks are visualized in Fig. 2a, and a detailed view of the long-skip connections and residual blocks are shown in Fig. 2d and Fig. 2e, respectively.

2.3. Morphological post-processing

Even though the U-Net deep learning architecture produces a full segmentation mask, unlike patch-based deep learning methods, post-processing, specifically morphological operations, are still essential to yielding contiguous regions and accurate nuclear boundaries. We designed a post-processing pipeline, shown in Fig. 3, to accomplish this. It consists of the following steps. A mask of confident interior pixels is created by identifying pixels for which the inside probability is greater than other labels, followed by morphological opening. A map of nuclear seeds is created by thresholding ($thres = 0.85$) the probability of the interior class of these pixels, followed by opening. Each seed is assigned a unique index, which is propagated to all connected interior pixels. Regions not covered will be assigned new index and combined with the previous result. Then, we apply binary dilation to create the final result. The parameters, such as window size, for these various morphological operations were optimized on the validation set in our experiments.

2.4. Data augmentation by nuclear deformation

To aid in training a robust network that minimizes overfitting, we implemented a novel approach designed specific for histology image analysis to generate augmented training images. Given the ground-truth annotations, each nucleus in an image is extracted and then deformed slightly by affine, spline, and elastic transformations as well as small random translation, yielding a new orientation, position,

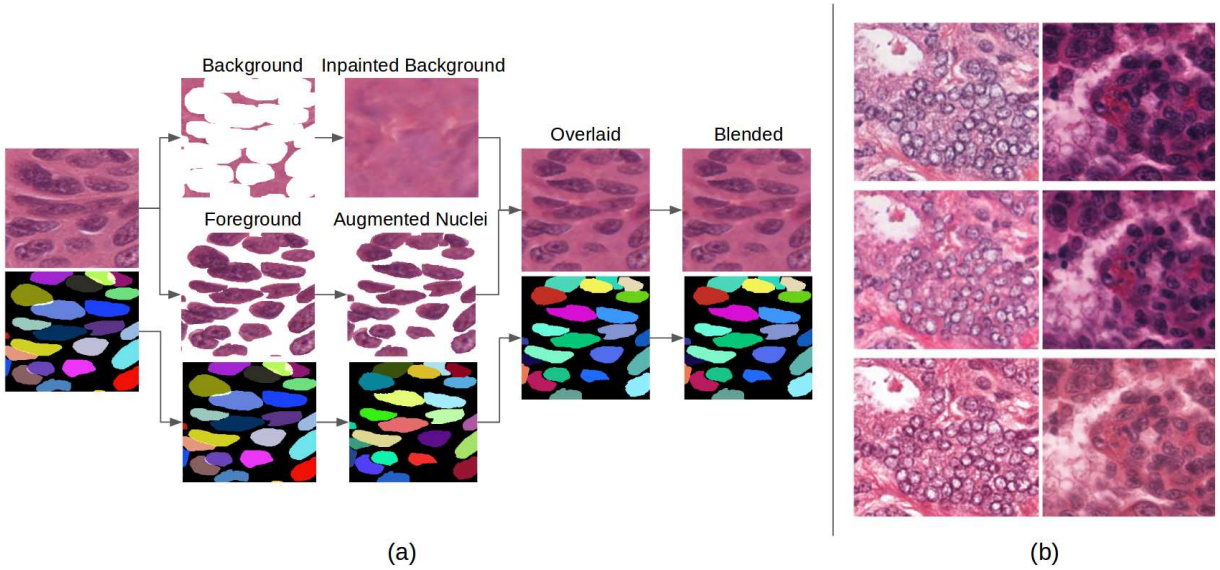


Figure 4: (a) Our proposed process generating new synthetic images with their corresponding annotations. (b) Data augmentation and color normalization examples. The first row shows example training images from [13]. The second row shows synthetic images generated by our proposed data-augmentation method. The last row shows the original training images normalized by the method proposed in [21].

and shape. Extracting the nucleus from the image produces empty holes in the background, which are filled by the image inpainting method proposed in [27]. Some elastic transformations and slight Gaussian blurring are applied at this point, along with some added speckle noise. The transformed nuclei are then overlaid back onto the reconstructed background image. Finally Gaussian blurring is applied to help smooth any harsh or unnatural edges between the augmented nuclei and the background. Fig. 4 diagrams this process and shows examples of resulting augmented training images.

2.5. Stain normalization

To reduce the uninformative and possibly confusing color variation inherent to H&E stained images, we use the structure-preserving color normalization method proposed in [21]. One image from the training data set is used as the target and all other images, including the new augmented data, are converted to its color space. We use $\lambda = 0.1$ as recommended in [21]. The last row in Fig. 4b shows example normalized images.

3. Experiments and Results

3.1. Data set and evaluation metrics

We evaluated our enhanced U-Net architecture and proposed training data augmentation approach on the data set curated by [13], which consists of 30 histopathology images with accompanying full segmentation masks. The images are 1000×1000 pixels in size and were extracted from WSIs from unique TCGA samples, from a variety of different organs. To comprise the training set, four images were randomly selected from breast, liver, kidney and prostate samples. The remaining 14 images were from 7 different organs and were evenly split into validation and testing sets. To aid in training, for each image, we generate 25 new synthetic images and their corresponding masks by our proposed method, yielding a total of 750 new synthetic images.

For evaluation, we focused on the AJI metric, proposed in [13], which balances detection accuracy with the accuracy of the delineated boundaries of nuclei, though we also considered the F1-score & Dice coefficient to shed more light upon the performance of the various architectures.

3.2. Experimental setup

Our U-Net architectures operate on patches of size 128×128 (due to size limitations of the GPU), so to generate a set of training patches, we extracted random patches from each image during training. Before extracting the patch, the original image was rotated and scaled by random amounts. The network was trained to minimize the cross entropy loss plus the generalized Dice loss [20], which helps specifically to learn sharper boundaries by addressing the class imbalance problem of boundary pixels. We used the Python library NiftyNet [4] for an implementation of the Dice loss. We adopt the weight initialization proposed in [8]. In our ex-

periments, we found that our enhanced U-Net works better without dropout layers, so we removed them, since they only increased the number of training steps required to converge. We used the Adam [12] optimizer for learning. The learning rate was set constant as $5e-5$ during the training process because of the adaptive property of Adam optimization [12]. Our batch size was restricted to be four, due to memory limitations on the GPU. We trained each network for roughly 300,000 steps. We used the validation set to fine-tune parameters.

During test-time inference, to generate a complete segmentation map, since the images are larger than 128×128 pixels, we perform inference on overlapping patches, with an overlap of 62 pixels, and then merge the results. We use the reflection transformation to pad patches on the boundaries of the original image.

3.3. Experimental results

Fig. 1 shows the results of our fully-enhanced U-Net, trained with augmented images, on several example histopathology images from the held-out test set, and the first row of Table 1 shows its performance according to the aforementioned metrics. As can be seen in Fig. 1, the boundary between overlapping nuclei can be reasonably separated by our method. Moreover, our method is able to generalize well to other types of organs, even those for which it was not trained. The first two lines in Fig. 1 are from stomach and colon tissue, respectively, which tissue types were not in the training set, yet our method still produces strong results. On the entire test set, our method achieved an AJI of 0.629. For comparison, the method proposed in [13] achieved an AJI of 0.508 on the same data set, although this comparison is not definitive, since 7 fewer images from the data set were used for training in their experiments.

Processing an entire image of size 1000×1000 on a single TITAN V 12GB GPU with our enhanced U-Net architecture takes only about 17 seconds. The subsequent post-processing incurs an additional 2 seconds on a CPU to create the final result.

3.4. Ablation study

We further evaluated the effect of each proposed component to the performance of our method through an ablation study. Example resulting segmentations from each of the following experiments are shown in Fig. 5 and the performance according to the aforementioned metrics are given in Table 1.

Residual blocks on long-skip connections

In this experiment, we removed all residual blocks on the long-skip connections of the U-Net architecture. To make a fair comparison, we compensated by adding more residual blocks in the encoder and decoder components to maintain the number of trainable parameters. All of other hyper-parameters were kept constant. The first and third line in Table 1 show the comparison between the two different models. Without residual blocks on the long-skip connection, the AJI value decreased by 1.4%, and the Dice score and F1-score decreased by 1.3% and 1.2%, respectively. This strengthens the credibility of our hypothesis that residual blocks on the long-skip connections help the network extract richer low-level features and thereby aid the network in delineating nuclear boundaries of touching nuclei. Fig. 5c and 5d further visualize the results of the two architectures. Some nuclei in Fig. 5d evidence a difficulty in separating their overlap, while our architecture with long-skip residual blocks successfully separates them.

Method	AJI	F1-score	Dice's coef.	#Params
Ours	**0.6291**	0.8469	**0.7980**	102M
Ours - D.A	0.6019	0.8006	0.7796	102M
Ours - R.S	0.6151	0.8349	0.7846	101M
Ours - G.E.	0.6125	**0.8490**	0.7893	101M

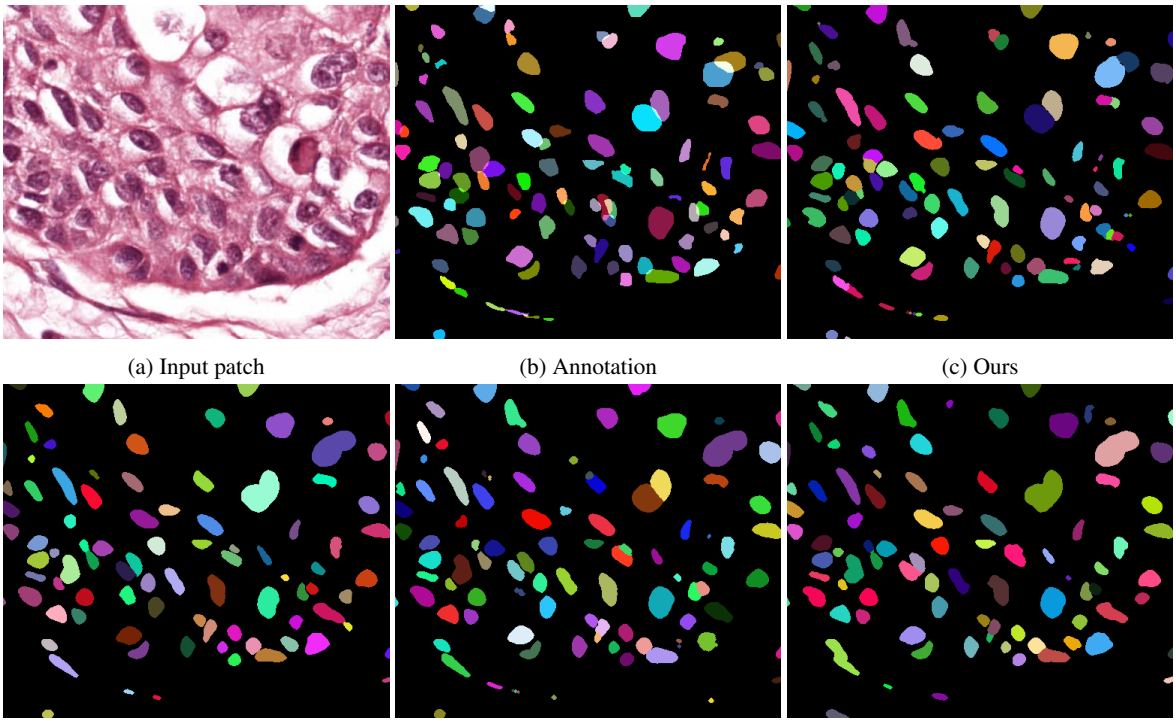
Table 1: Quantitative comparison of each proposed component of the U-Net architecture and training procedure. -D.A: without our proposed synthetic data; -R.S: without residual blocks on the long-skip connections; -G.E.: the U-Net without group-equivariant operations.

New synthetic data

To see the effect of the new synthetic data we proposed, we withheld these synthetic images from the training set and trained a separate network on this reduced training set. The architecture of the network and all of other hyper-parameters were kept the same. As can be seen in Table 1, training with these new synthetic images can improve the model's performance by approximately 2.7% in AJI, showing the most profound impact upon performance compared to the other two contributions. This suggests that this method of data augmentation further offsets the problem of sample scarcity for histopathological analysis, even beyond standard augmentation techniques. This is significant since labeling this type of data is extremely labor intensive.

Group equivariance

To see the effect of group-equivariant operations, we replaced these operations with standard operations for a CNN and retrained the network. Following the work in [3], to preserve the same number of trainable parameters, we doubled the number of filters in each convolution layer in the ordinary network, while keeping the same architecture oth-



(a) Input patch (b) Annotation (c) Ours

(d) Ours - R.S (e) Ours - D.A (f) Ours - G.E.

Figure 5: The visualization of example results of different methods in the ablation study. The segmented region for each distinct detected nucleus is shown by a unique color. The background is visualized in black. Patches are cropped for better visualization. Best view in color. -D.A: without our proposed synthetic data; -R.S: without residual blocks on the long-skip connections. -G.E.: the U-Net without group-equivariant operations.

erwise. This new model was trained with a similar set of hyper-parameters, except that we increased the learning rate to $1e-4$ and we were able to use a batch size of eight, since the network did not require as much storage on the GPU as the group-equivariant version. Since there were no group-equivariant operations, we added dropout to the model to help with regularization. As shown in Table 1, both the AJI and Dice's coefficient drop by a significant margin, roughly 1.7% and 0.9% respectively, without encoded group equivariance. This implies that integrating group-equivariant convolution and operations into the current U-Net architecture can indeed enable the network to learn better parameters that generalize well to simple transformations, namely, translations and rotations. Although, since it is only equivariant to rotations of 90 degrees, the network can still benefit from data augmentation of rotations of finer, arbitrary angles.

4. Conclusion

In this work, we have shown the value of several enhancements to the standard U-Net architecture, namely, encoding rotation and translation equivariance and adding additional residual blocks, and our novel data augmentation method for automated nuclear segmentation in histology images. This work can be considered as a parallel work to the many other current developments in deep-learning-based nuclear segmentation, which also could be incorporated to further improve performance. By contributing to improved performance for this crucial step for many computational pathology pipelines, without adding significant computational overhead, we believe these enhancements will help to enable future discoveries in pathology, with the hope of increasing the clinical impact of computational pathology.

References

[1] E. J. Bekkers, M. W. Lafarge, M. Veta, K. A. Eppenhof, J. P. Pluim, and R. Duits. Roto-translation covariant convolutional networks for medical image analysis. In *MICCAI*, pages 440–448. Springer, 2018.

[2] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *arXiv:1606.00915*, 2016.

[3] T. Cohen and M. Welling. Group equivariant convolutional networks. In *ICML*, pages 2990–2999, 2016.

[4] E. Gibson, W. Li, C. Sudre, L. Fidon, D. I. Shakir, G. Wang, Z. Eaton-Rosen, R. Gray, T. Doel, Y. Hu, T. Whyntie, P. Nachev, M. Modat, D. C. Barratt, S. Ourselin, M. J. Cardoso, and T. Vercauteren. Niftynet: a deep-learning platform for medical imaging. *Computer Methods and Programs in Biomedicine*, 2018.

[5] M. N. Gurcan, L. Boucheron, A. Can, A. Madabhushi, N. Rajpoot, and B. Yener. Histopathological image analysis: A review. *IEEE reviews in biomedical engineering*, 2:147, 2009.

[6] D. Han, J. Kim, and J. Kim. Deep pyramidal residual networks. *CoRR*, 2016.

[7] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CoRR*, 2015.

[8] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *CoRR*, 2015.

[9] H. Irshad, A. Veillard, L. Roux, and D. Racoceanu. Methods for nuclei detection, segmentation, and classification in digital histopathology: a review – current status and future potential. *IEEE reviews in biomedical engineering*, 7:97–114, 2014.

[10] A. Janowczyk and A. Madabhushi. Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. *Journal of pathology informatics*, 7, 2016.

[11] A. Khoreva, R. Benenson, E. Ilg, T. Brox, and B. Schiele. Lucid data dreaming for object tracking. *CoRR*, 2017.

[12] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, 2014.

[13] N. Kumar, R. Verma, S. Sharma, S. Bhargava, A. Vahadane, and A. Sethi. A dataset and a technique for generalized nuclear segmentation for computational pathology. *IEEE transactions on medical imaging*, 36(7):1550–1560, July 2017.

[14] J. Linmans, J. Winkens, B. S. Veeling, T. S. Cohen, and M. Welling. Sample efficient semantic segmentation using rotation equivariant convolutional networks. *arXiv:1807.00583*, 2018.

[15] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015.

[16] F. Mahmood, D. Borders, R. Chen, G. N. McKay, K. J. Salimian, A. Baras, and N. J. Durr. Deep adversarial training for multi-organ nuclei segmentation in histopathology images. *arXiv:1810.00236*, 2018.

[17] T. C. G. A. Network et al. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61, 2012.

[18] N. Otsu. A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, 9(1):62–66, 1979.

[19] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, 2015.

[20] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. J. Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pages 240–248. Springer, 2017.

[21] A. Vahadane, T. Peng, S. Albarqouni, M. Baust, K. Steiger, A. M. Schlitter, A. Sethi, I. Esposito, and N. Navab. Structure-preserved color normalization for histological images. In *ISBI*, pages 1012–1015, April 2015.

[22] B. S. Veeling, J. Linmans, J. Winkens, T. Cohen, and M. Welling. Rotation equivariant cnns for digital pathology. In *MICCAI*, pages 210–218. Springer, 2018.

[23] M. Veta, A. Huisman, M. A. Viergever, P. J. van Diest, and J. P. Pluim. Marker-controlled watershed segmentation of nuclei in h&e stained breast cancer biopsy images. In *ISBI*, pages 618–621. IEEE, 2011.

[24] M. Veta, J. P. Pluim, P. J. Van Diest, and M. A. Viergever. Breast cancer histopathology image analysis: A review. *IEEE transactions on biomedical engineering*, 61(5):1400–1411, 2014.

[25] F. Xing, Y. Xie, and L. Yang. An automatic learning-based framework for robust nucleus segmentation. *IEEE transactions on medical imaging*, 35(2):550–566, 2016.

[26] F. Xing and L. Yang. Robust nucleus/cell detection and segmentation in digital pathology and microscopy images: a comprehensive review. *IEEE reviews in biomedical engineering*, 9:234–263, 2016.

[27] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang. Generative image inpainting with contextual attention. *arXiv:1801.07892*, 2018.

A Multi-Organ Nucleus Segmentation Challenge

Neeraj Kumar, Ruchika Verma, Deepak Anand, Yanning Zhou, Omer Fahri Onder, Efstratios Tsougenis, Hao Chen, Pheng-Ann Heng, Jiahui Li, Zhiqiang Hu, Yunzhi Wang, Navid Alemi Koohbanani, Mostafa Jahanifar, Neda Zamani Tajeddin, Ali Gooya, Nasir Rajpoot, Xuhua Ren, Sihang Zhou, Cheng-Kun Yang, Chi-Hung Weng, Wei-Hsiang Yu, Chao-Yuan Yeh, Shuang Yang, Shuoyu Xu, Pak Hei Yeung, Peng Sun, Amirreza Mahbod, Gerald Schaefer, Isabella Ellinger, Rupert Ecker, Orjan Smedby, Chunliang Wang, Benjamin Chidester, That-Vinh Ton, Minh-Triet Tran, Jian Ma, Minh N. Do, Akshaykumar Gunda, Raviteja Chunduri, Corey Hu, Xiaoyang Zhou, Yuhsuan Wu, Liu Hao, Yunzhe Zhang, Zitao Zeng, Weihao Xie, Dariush Lotfi, Reza Safdari, Antanas Kascenas, Alison O'Neil, Dennis Eschweiler, Johannes Stegmaier, Yanping Cui, Baocai Yin, Kailin Chen, Xinmei Tian, Philipp Gruening, Elad Arbel, Itay Remer, Amir Ben-Dor, Ekaterina Sirazitdinova, Matthias Kohl, Stefan Braunewell, Yuexiang Li, Xinpeng Xie, Linlin Shen, Jun Ma, Krishanu Das Baksi, Mohammad Azam Khan, Jaegul Choo, Adrin Colomer, Valery Naranjo, Linmin Pei, Khan M. Iftekharuddin, Okyaz Eminaga, Mirabela Rusu, Kaushiki Roy, Debotosh Bhattacharjee, Anibal Pedraza, Maria Gloria Bueno, Sabarinathan D, Saravanan R,
Praveen K, Zihan Wu, Johannes Bernhard, Rebecca Stone, David Hogg, Guanyu Cai, Xiaojie Liu, Yuqin Wang, and Amit Sethi

***Abstract*—Generalized nucleus segmentation techniques can contribute greatly to reducing the time to develop and validate visual biomarkers for new digital pathology datasets. We summarize the results of MONuSeg 2018 Challenge whose objective was to develop techniques that generalize to new datasets and organs for segmenting nuclei in digital pathology. The challenge was a satellite event of the MICCAI 2018 conference. Contestants were given a training set with 30 images from seven organs with annotations of 21,623 individual nuclei. A test dataset with 14 images taken from seven organs, including two organs that did not appear in the training set was released without annotations. Entries were evaluated based on average aggregated Jaccard index (AJI) on the test set to prioritize accurate instance segmentation as opposed to merely semantic segmentation. More than half the teams that completed the challenge outperformed a previous baseline [1]. Among the trends observed that contributed to increased accuracy were the use of color normalization as well as heavy data augmentation. Additionally, fully convolutional networks inspired by variants of U-Net [2], FCN [3], and mask R-CNN [4] were popularly used, typically based on ResNet [5] or VGG [6] base architectures. Watershed segmentation on predicted semantic segmentation maps seeded by predicted nuclear centers was a popular post-processing strategy. Using the techniques described by the contestants, we hope that the computational pathology community will find it much easier to design studies based on nuclear morphometrics.**

***Index Terms*—Multi-organ, nucleus segmentation, digital pathology.**

I. Introduction

Examination of H&E stained tissue under a microscope remains the mainstay of pathology. The popularity of H&E is due to its low cost and ability to reveal tissue structure and nuclear morphology, which is sufficient for primary diagnosis of several diseases including many cancers. Nuclear shapes and spatial arrangements often form the basis of the examination of H&E stained tissue. For example, grading of various types of cancer and risk stratification of patients is usually done by examining different types of nuclei on a tissue slide [7]. Nuclear morphometric features and appearance including the color of their surrounding cytoplasm also helps in identifying various types of cells such as epithelial (glandular), stromal, or inflammatory, which in turn give an idea of the glandular structure and disease presentation at low power [8]–[11]. Segmentation of nuclei accurately in H&E images therefore has high utility in digital pathology.

Nucleus segmentation algorithms that work well on one dataset can perform poorly on a different dataset. There is far too much variation in the appearance of nuclei and their surroundings by organs, disease conditions, and even digital scanner brands or histology technicians. Examples of such variations are shown in Figure 1, along with the problems of some common segmentation algorithms such as Otsu thresholding [12], marker controlled watershed segmentation [13]–[15] or open-source packages like Fiji [16] and Cell Profiler [17]. Segmentation based on machine learning should be able to do a better job, but that makes designing and refining nucleus segmentation algorithms for a new study a tedious task because annotations of thousands of nuclei are needed to train such segmentation models on datasets of interest. Nucleus segmentation that generalize to new datasets and organs that were not seen during training can reduce this effort substantially and contribute to rapid experimentation with new phenotypical (visual) biomarkers.

Until recently, one of the major challenges in training generalized nucleus segmentation models has been the unavailability of large multi-organ datasets with annotated nuclei. In 2017 Kumar *et al*.released a dataset [1] with more than 21,000 hand-

Authors were at various institutes. Emails for correspondence: neeraj.kumar.iitg@gmail.com, ruchika@case.edu, deepakanand@iitb.ac.in, and asethi@iitb.ac.in

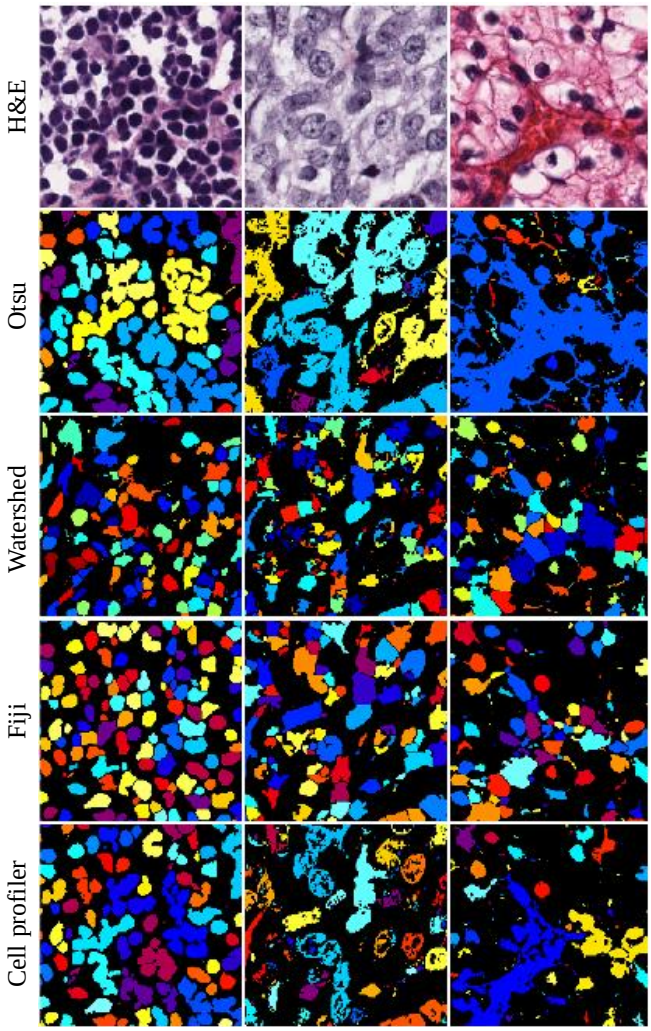


Fig. 1: Challenges in nucleus segmentation: Original H&E stained tissue images show crowded and chromatin-sparse nuclei with color variation across tissue slides. Otsu thresholding [12] and Cell Profiler [17] leads to merged nuclei (under-segmentation). Marker controlled watershed segmentation [13] and Fiji [16] leads to fragmented nuclei (over-segmentation).

annotated nuclei in H&E stained tissue images acquired at the commonly used 40x magnification, sourced from seven organs and multiple hospitals in The Cancer Genome Atlas (TCGA) [18]. Kumar *et al.*also introduced a metric called Aggregated Jaccard Index (AJI) that is more appropriate to evaluate algorithms for this instance segmentation problem as opposed to other popular metrics such as Dice coefficient, which are more suited for semantic segmentation problems. This is because nucleus segmentation algorithms should not only tell the difference between nuclear and non-nuclear pixels, but they should also be able to tell pixels belonging to two nuclei apart that touch or overlap with each other. Additionally, they had released a trained model that performed reasonably well on unseen organs from the test subset of images.

We organized the Multi-Organ nucleus segmentation Challenge (MONuSeg) Challenge at MICCAI 2018 to build upon Kumar *et al.*'s work by enlarging the dataset and by encouraging others to introduce new techniques for generalized nucleus segmentation. The participation was wide and several of participants outperformed a previous benchmark [1] by a significant margin. In this paper we describe in detail the objectives of the competition, the released dataset, and the emerging trends of techniques that performed well on the challenge task. We hope that using the algorithms described on the challenge webpage [19] will be of use to the computational pathology research community.

The rest of the paper is organized as follows. We describe the prior work on nucleus segmentation and dataset creation in Section II. We describe the dataset and competition rules in Section III. We present an organized summary of the techniques used by the challenge participants in Section IV. Finally, we discuss emerging trends in nucleus segmentation techniques in Section V.

II. Background and Prior Work

In this section we describe the importance of H&E stained images in histopathology and provide details of some previous notable techniques and datasets for nucleus segmentation from H&E stained images.

A. Hematoxylin and eosin (H&E) stained images

Pathologists usually observe tissue slides under a microscope at specific resolution (ranging between 5x and 40x with a 10x eyepiece) to report their diagnoses including tumor grade, extent of spread, surgical margin, etc. Their assessment is primarily based on the appearance, size, shape, color and crowding of glands as well as various nuclei in epithelium and stroma. Stains are used to enhance the contrast between these tissue components to help a pathologist looking for specific nuclei and gland features. The combination of hematoxylin and eosin (H&E) is a frequently-used, universal, and inexpensive staining scheme for general contrast enhancement of histologic structures of a tissue. Hematoxylin renders the nuclei dark blueish purple and the epithelium light purple, while eosin renders the stroma pink. Compared to the general use of H&E, immunohistochemical staining is more specialized as it targets proteins specific to certain disease states for visual identification.

With the advent of high resolution cameras mounted on microscopes, and more importantly, digital whole slide scanners, it is now possible to capture and store the whole slide images (WSIs) of the tissue sections for computer assisted diagnosis (CAD). However, the development of CAD systems requires automated extraction of rich information encoded in the pixels of WSIs. In recent years computer based assessment of tissue images has been used for tumor molecular subtype detection [20], mortality prediction [8], and treatment effectiveness prediction [11], [21]. Notably, nucleus detection and segmentation is often a first step for several such CAD systems that rely on nuclear morphometrics for disease state stratification and predictive modelling. In this challenge we focused on crowdsourcing *techniques* for nucleus segmentation in H&E stained images captured at a popular resolution of 40x.

B. Nucleus segmentation techniques

Prior to the advent of deep learning, approaches to segment nuclei relied on watershed segmentation, morphological operations – such as erosion, dilation, opening and closing – color-based thresholding, and variants of active contours [13], [14], [22]–[25]. These techniques were often complemented with a collection of pre-processing methods, such as contrast enhancement and deblurring to improve the 'image quality. Additionally, several post-processing techniques, such as hole filling, noise removal, graph-cuts, etc., were also used to refine the outputs of the segmentation algorithms. However, these approaches do not generalize well across a wide spectrum of tissue images due to reasons such as (a) variations in nuclei morphologies of various organs and tissue types, (b) inter- and intra-nuclei color variations in crowded and chromatin-sparse nuclei, and (c) diversity in the quality of tissue images owing to the differences in image acquisition equipment and slide preparation protocols across hospitals and clinics.

Use of machine learning provides an opportunity to automatically discount the aforementioned sources of variations in pixel values and concentrate on relative differences between nuclear and non-nuclear pixels as well as overall shapes for better generalized segmentation. There have been tremendous advances in the recent years to develop learning-based nucleus segmentation methods to advance the state-of-the-art. The use of learning based approaches started with the extraction of hand-crafted local features based on color and spatial filtering that were fed to traditional learning-based models such as random forests, support vector machines, etc. to segment nuclei and non-nuclei regions [26]–[29]. The selection of features is dependent on domain knowledge and trial-and-error for improving nucleus segmentation performance, and yet it is difficult to detect all nuclei with diverse appearances and crowding patterns.

To circumvent the constraints of hand-crafted features representation learning algorithms, popularly known as deep learning techniques, have recently emerged. These methods – specifically the ones using convolutional neural networks (CNNs) – have outperformed previous techniques in nucleus detection and segmentation tasks by significant margins [1], [30]–[34]. To use deep learning the problem is often cast as one of semantic segmentation wherein an a two-class probability map for nuclear and non-nuclear regions is to be computed. After semantic segmentation, sophisticated post-processing methods – such as graph partitioning [30], or the computation of distance transform of the nuclear map followed by H-minima transform and region growing [31] – are often used to obtain final nuclei shapes with the desired separation of touching and overlapping nuclei. Semantic segmentation of third class of pixels – those on the nuclear boundaries including that between two touching nuclei – has also been proposed to exclusively refine the separation between the segmented touching and overlapping nuclei [1]. More recently, nucleus segmentation problem has been formulated as a regression task to predict a distance map with respect to centroids or boundaries of nuclei using fully convolutional networks (FCNs) to achieve both segmentation and computational performance gains over previous deep learning based approaches [34]. More comprehensive reviews of state-of-the-art nucleus segmentation algorithms can be found in [35] and [36].

One of the major barriers in out of the box (without re-training) application state-of-the art deep learning based nucleus segmentation algorithms was the lack of publicly available source codes and trained models by previously published techniques until Kumar *et al*. [1] and Naylor *et al*. [34] released their source codes. The other major barrier was the lack of publicly available annotated datasets for benchmarking, which we address next.

C. Nucleus segmentation datasets

The success of machine learning and development of state-of-the art deep learning algorithms in computer vision can be attributed to the healthy competition enabled by publicly available datasets such as ImageNet [37] and CIFAR [38] for object recognition in images, and UCF for action recognition in videos [39]. Unfortunately, we do not see similar progress in digital pathology image analysis as there is dearth of annotated datasets for solving various tasks of pathologist's interest. This is because annotating pathology images requires expert knowledge and manually annotating tissue structures (such as identifying nuclear boundaries) is quite tedious. However, there have been a few recent efforts dedicated to the release of hand-annotated H&E stained tissue slide images for nucleus segmentation as summarized in Table I. These datasets can also be downloaded from the challenge webpage [19]. Please note that we have not included datasets where the nuclei were annotated for detection alone in Table I because these can not be used for the segmentation task. We also excluded datasets annotated for other specific objectives such as gland segmentation, mitosis detection, epithelial segmentation, and tumor type classification, as opposed to generalized nucleus segmentation.

Most of the datasets listed in Table I focus on a specific organ with the exception of Kumar *et al*. [1] and Wienert *et al*. [25].

III. Dataset and Competition Rules

The objective of MONuSeg 2018 was to encourage the development of learning based generalized nucleus segmentation techniques that work right out of the box (without re-training) on a diverse set of H&E stained tissue images. The images therefore spanned a range of patients, organs, disease states, and sourcing hospitals with potentially different slide preparation and image acquisition methods. Training and test datasets were carefully curated and the competition rules were crafted in accordance with the these objectives.

A. Training dataset

The training data of MONuSeg 2018 was the same as that released previously by Kumar *et al*. [1], which comprised 30 tissues images, each of size 1000×1000, containing $21,623$

[1]Only annotations verified by a pathologist were considered.

TABLE I: Publicly available H&E stained tissue image datasets annotated for nucleus segmentation

Dataset	Image Size	Images	Nuclei	Organs	Annotation type
Kumar *et al.* [1]	1000×1000	30	21,623	Multiple (7)	Individual boundaries
Janowczyk *et al.* [40]	2000×2000	143	12,000	Breast	Foreground Mask
Wienert *et al.* [25]	600×600	36	7,931	Multiple (5)	Individual boundaries
Naylor *et al.* [34]	512×512	50	4,022	Breast	Foreground Mask
Irshad *et al.*[1] [41]	400×400	63	2,532	Kidney	Foreground Mask
Gelasca *et al.* [42]	896×768 (768×512)	50	1,895	Breast	Foreground Mask

TABLE II: MONuSeg 2018 training and test datset composition.

Data subset ↓	Nuclei	Images									
	Total	Total	Breast	Liver	Kidney	Prostate	Bladder	Colon	Stomach	Lung	Brain
Training set	21,623	30	6	6	6	6	2	2	2	–	–
Testing set	10,000	14	2	–	3	2	2	1	–	2	2
Total	**31,623**	**44**	**8**	**6**	**9**	**8**	**4**	**3**	**2**	**2**	**2**

hand-annotated nuclear boundaries. Each 1000×1000 image in this dataset was extracted from a separate whole slide image (WSI) (scanned at 40x) of an individual patient downloaded from TCGA [18]. The dataset represented 7 different organs *viz.*, breast, liver, kidney, prostate, bladder, colon and stomach, and included both benign and diseased tissue samples to ensure diversity of nuclear appearances. Furthermore, the training images came from 18 different hospitals, which introduced another source of appearance variation due to the differences in the staining practices and image acquisition equipments (scanners) across labs. Representative 1000×1000 sub-images from regions dense in nuclei were extracted from patient WSIs to reduce the computational burden of processing WSIs and increase participation. Only one crop per WSI and patient was included in the dataset to ensure diversity. The distribution of training images across organs is shown in Table II while patient and hospital details are available on the challenge webpage [19].

Both epithelial and stromal nuclei were manually annotated in the 1000×1000 sub-images using Aperio ImageScope®. Annotations were performed on a 25" monitor with a 200x zoom such that each image pixel occupied 5×5 screen pixels to ensure clear visibility for annotating nuclear boundaries with a laser mouse. For overlapping nuclei, each multi-nuclear pixel was assigned to the nucleus that appeared to be on the top in the 3-D structure. The annotators were engineering graduates and the quality control was performed by an expert pathologist with years of experience in analyzing tissue sections. The images and XML files containing containing pixel coordinates of the annotated nuclear boundaries were released for public use by Kumar *et al.* [1]. The reasons that make this dataset ideal for training a generalized nucleus segmentation model are as follows:

1) Kumar *et al.* [1] is the largest repository of hand annotated nuclei which aptly represents a miscellany of nuclei shapes, and sizes across multiple organs, disease states and patients. The inclusion of tissue sections from 18 hospitals further augments the richness of this dataset. From Table I, the only multi-organ alternative to Kumar *et al.* [1] is Weinert *et al.* [25]. However, Weinert *et al.* [25] contains tissues from lesser number of organs captured in a single hospital with a single scanner.
2) Kumar *et al.* [1] extracted only one sub-image of 1000×1000 pixels per patient to maximize nuclear appearance variation. Other datasets mentioned in Table I extracted multiple sub-images each patient and are thus limited in representing nuclear appearance diversity. For example, WSIs of only 10 and 11 patients were used in Irshad *et al.* [41] and Naylor *et al.* [34], respectively.
3) Kumar *et al.* [1] provided coordinates of annotated nuclear boundaries in popular *.xml* format instead of foreground masks. This is crucial for learning to separate touching and overlapping nuclei in any automatic nucleus segmentation algorithm. This helped several participants of MONuSeg 2018 whose nucleus segmentation algorithms explicitly learned to recognize nuclear boundaries in addition to the usual foreground (nuclei pixels) and background classes (non-nuclei pixels).
4) Kumar *et al.* [1] publicly released the source code of their generalized nucleus segmentation algorithm to catalyze natural competition among a newer generation of automatic nucleus segmentation algorithms.

B. Testing dataset

A new testing set comprising 14 images, each of size 1000×1000 pixels, spanning 7 organs (*viz.* kidney, lung, colon, breast, bladder, prostate, brain), several disease states (benign and tumors at different stages), and approximately 10,000 annotated nuclei was prepared in the same manner as used for preparing the training data. As shown in Table II, lung and brain tissue images are exclusive to the test set which makes it more challenging. No two images in the training and test set came from the same hospital. More details about the test set are available in the "supplementary material" tab of the challenge webpage [19]. The annotations of the test set were not released to the participants. To formally conclude the challenge, with this paper, we are releasing the test annotations on the challenge webpage [19] to facilitate future research in the development of generalized nucleus segmentation algorithms.

C. Competition metric

Average aggregated Jaccard Index (AJI) was used as the metric to evaluate nucleus segmentation performance of the competing algorithms because of it established advantages

over other segmentation metrics [1], [33], [34]. The value of AJI ranges between 0 to 1, indicating no overlap to perfect segmentation respectively. Computing AJI involves matching every ground truth nuclei to one detected nuclei by maximizing the Jaccard index. The AJI is then equal to the ratio of the sums of the cardinals of intersection and union of these matched ground truth and predicted nuclei. In addition, all detected components that are not matched are added to the denominator. We reproduce Algorithm 1 detailing AJI computation from Kumar *et al.* [1] with permission. The code for computing AJI is available on the challenge webpage [19].

Algorithm 1 Aggregated Jaccard index (AJI)

Input: A set of images with a combined set of annotated nuclei G_i indexed by i, and a segmented set of nuclei S_k indexed by k.
Output: Aggregated Jaccard Index A.

Initialize overall correct and union pixel counts: $C \leftarrow 0; U \leftarrow 0$
  Each ground truth nucleus G_i
  $j \leftarrow \arg\max_k(|G_i \cap S_k|/|G_i \cup S_k|)$
  Update pixel counts: $C \leftarrow C + |G_i \cap S_j|; U \leftarrow U + |G_i \cup S_j|$
  Mark S_j *used*
    Each segmented nucleus S_j
  If S_k is not *used* then $U \leftarrow U + |S_k|$
$A \leftarrow C/U$

Participants were asked to submit 14 segmentation output files (one for each of the 14 test images) to the challenge organizers. For each participant submission, the organizers then computed 14 AJIs (one for each test image) as per Algorithm 1. If a participant did not submit results for a particular testing image then AJI value of zero was assigned for that particular image to that participant. The organizers then computed the average AJI (a-AJI) for each participant by averaging image level AJIs across 14 test images. The participants were then ranked in the descending order of a-AJI to obtain the final leaderboard shown in Table III.

IV. Summary of Segmentation Techniques

In this section we present a summary of the techniques used by the 36 teams who successfully completed the challenge. We describe the trends observed in pre-processing, data augmentation, modeling, task specification, optimization, and post-processing techniques used by the teams. Specific details of all algorithms are provided in respective manuscripts submitted by participants as per challenge policies and are available on the challenge webpage [19] under "manuscripts" tab.

A. Pre-processing and data augmentation

Pre-processing techniques reduce unwanted variations among input images – from both the training and test sets – so that the test data distribution is not very different from the training data distribution, by projecting both to the same low-dimensional manifold. On the other hand, data augmentation techniques increase the training data set size by introducing controlled random variations with the hope of creating a training data distribution that covers most of the test data distribution. There are several ways in which the participants altered the given images and their ground truth masks before passing them to the segmentation learning systems in order to increase test accuracy. We summarize some of the interesting trends observed in this challenge. These results are also summarized in Table III.

1) Color and intensity normalization: Among the data pre-processing techniques, color and intensity transformations were the most common. Approximately half the teams used color normalization techniques that were specifically developed for pathology images to reduce unwanted color variations between training and test data. Structure Preserving Color Normalization (SPCN) by Vahadane *et al.* [43] was used by ten teams due to its demonstrated performance and code availability. Another seven teams used Mecenko *et al.*'s color normalization scheme [44], out of which one used this technique in combination with another technique by Reinhard *et al.* [45]. Two teams used unspecified color normalization techniques.

Pixel intensity and RGB color transformations that are unspecific to pathology were also used by approximately half of the teams. Most popular among this class of techniques were channel-wise mean subtraction, variance normalization (unit variance), and pixel-value range standardization. Six teams also used either contrast enhancement (or histogram equalization), among which CLAHE [46] was the most commonly used technique.

Among the unique techniques, one team used image sharpening to remove unwanted variations between training and test data, one team concatenated HSV and L channels (of L, a*, b* color space) to the RGB channels, and one team used only the blue channel after color normalization of the RGB images.

2) Data augmentation: Among data augmentation techniques, geometric transformations of the image grid were the most common. For example, rigid transformations of the images – such as rotation (especially, by multiples of 90 degrees) and/or flipping – were used by all but four teams to increase the size of the training data. However, as can be seen in Table III, all of the top twelve teams by AJI also augmented the training set using affine transformations, while only five teams below that used this type of augmentation. A closely related non-rigid transformation is elastic deformation, but it was not very popular among the contestants due to the marginal gain it might afford over an affine transform while being more complicated to implement. Another related non-rigid geometric transformation is image scaling, which was used by nine contestants.

Another popular set of augmentation techniques involve changing the pixel values while leaving the geometric structure intact. The most popular among these techniques was the addition of white Gaussian noise, which was used by several of the top performing teams. Another popular technique is color jitter or random HSV shifts, which was used by nine of the top twelve teams. Color jitter is opposite in spirit to color normalization in that it is used to present more color variations of the same input geometric structure to the learning machine with the hope that it will learn to focus on the geometric structure as opposed to the color of nuclei, which may vary between training and test data sets. Random intensity (brightness) shifts were used by fewer participants, as were

blurring by isotropic Gaussian filters of random widths and random image sharpening.

One interesting data augmentation technique that was used by team *CMU-UIUC* involved segmenting the nuclei from one training image, and pasting them on to backgrounds from other images.

B. Specification of the learning task

The challenge of nucleus segmentation can be split into two tasks: distinguishing between nuclear and non-nuclear pixels (semantic segmentation) and separate touching nuclei (instance segmentation). The following were three principal types of outputs that the contestants produced using deep learning to meet these two challenges:

1) **Binary class probability** maps distinguish between pixels that belong to the core of any nucleus versus those do not. The process of not including the outer periphery of the nuclei into the foreground class helps separate touching nuclei. The lost nuclear territory can later be gained back during post-processing.
2) **Ternary class probability** map distinguishes between nuclear core, non-nuclear, and nuclear boundary pixels. Nuclear pixels that are on a shared boundary of two touching nuclei are considered to belong to the third class, which has been shown to be useful in separating touching nuclei [1].
3) **Distance map** estimates how far a nuclear pixel is from the centroid of a nucleus. Such a map can also distinguish between nuclear and non-nuclear pixels by assigning a fixed value to the latter, such as 0. This is a per-pixel regression problem while the previous two are classification problems. A variant of this distance map is to predict the distance from the boundary of the nucleus.

Most teams trained their models to predict variants of one or more of the three types of maps described above. One interesting departure from these three tasks was by Antanas Kascenas who predicted a five-class probability map – one for nuclear pixels, and the other four for their probability of belonging to one of the four Cartesian quadrants of a nucleus in order to separate touching nuclei.

C. Model architectures

All participants used deep convolutional neural networks. Twenty one teams used variants of U-Net [2], of which the original U-Net architecture was used by 11 teams while six teams used based architectures inspired by VGGNet [6], and another 11 teams used architectures inspired by either MRCNN [4], FCN [3], or ResNet [5] with different depths various depths. Eight teams used Mask Region with CNN features (M-RCNN) [4] as the primary models(of which, two also used U-Net), and two used FCN [3] (of which one also used U-Net). Among the remaining, four teams used their own custom models and architectures, and one each used VGG-Net [6], Deep Layer Aggregation [47], PANet [48], and TernausNet [49]. A few teams used multiple architectures for ensembling. Two teams used two architectures each for two different tasks, for example one for semantic segmentation (binary classification between foreground and background pixels) and another for distance map prediction to separate touching nuclei. Notable innovations in model architectures tried by some of the top teams are described in Section IV-G.

D. Model optimization

The choice of loss function depends on the desired output being predicted. Among various choices for the loss function, pixel-wise cross entropy was used by 28 teams for predicting binary or ternary probability maps, and was by far the most popular loss function. Ten teams used Dice loss [50], and two teams used its variant such as smooth Jaccard index loss or IOU (intersection over union) loss [51]. For regression problems, seven teams used a smooth L_1 loss. Five teams used mean square error. In total, 16 teams used more than one loss function.

Most teams trained their models end-to-end, except when an ensemble of more than one model was used, with the exception of team *Yunzhi* that used a cascade of two neural networks trained one after another.

E. Post-processing

For post-processing, watershed segmentation (WS) was used by 17 teams. The most popular way to apply WS was on the nuclear probability pixel map. Additionally, to separate touching nuclei several teams used a neural network to predict the location of a marker for each nucleus, such as by using a nuclear-core probability map, a distance map, or a vector map pointing to the nearest nuclear center.

Cleaning up small or weakly detected nuclei was also a common theme. Non-maxima suppression and h-minima were commonly used along with a threshold to clean up false positive.

F. Training and testing time

Training times ranged from a 2 hours and 17 minutes on using a single Nvidia 1080Ti GPU for team *Junma* to 42 hours for team *Johannes Stegmaier* on a similar hardware. Testing times also had a wide range from 1 second per 1000×1000 image for team *Unblockabulls* on an Amazon Web Services GPU instance powered by an Nvidia K80 GPU to 2 minutes 58 seconds per on an Nvidia Titan X GPU by team *CVBLab*.

G. Description of the top-five techniques

We now describe the top-five techniques in more detail as examples of the innovations and diligence with which the participants tried to get robust generalization. Comparative results of the top-five techniques is shown in Figure 2.

1) CUHK & IMSIGHT: Extensive data augmentation based on random affine transforms, rotations, and color jitters was used. The task was split into that of nucleus and boundary detection. A novel multi-head fully convolutional neural network (MH-FCN) architecture inspired by U-Net [2] and FCN [3] was used. The two prediction tasks were handled by two decoder branches (heads) that shared the same encoder, because the tasks are related to each other.

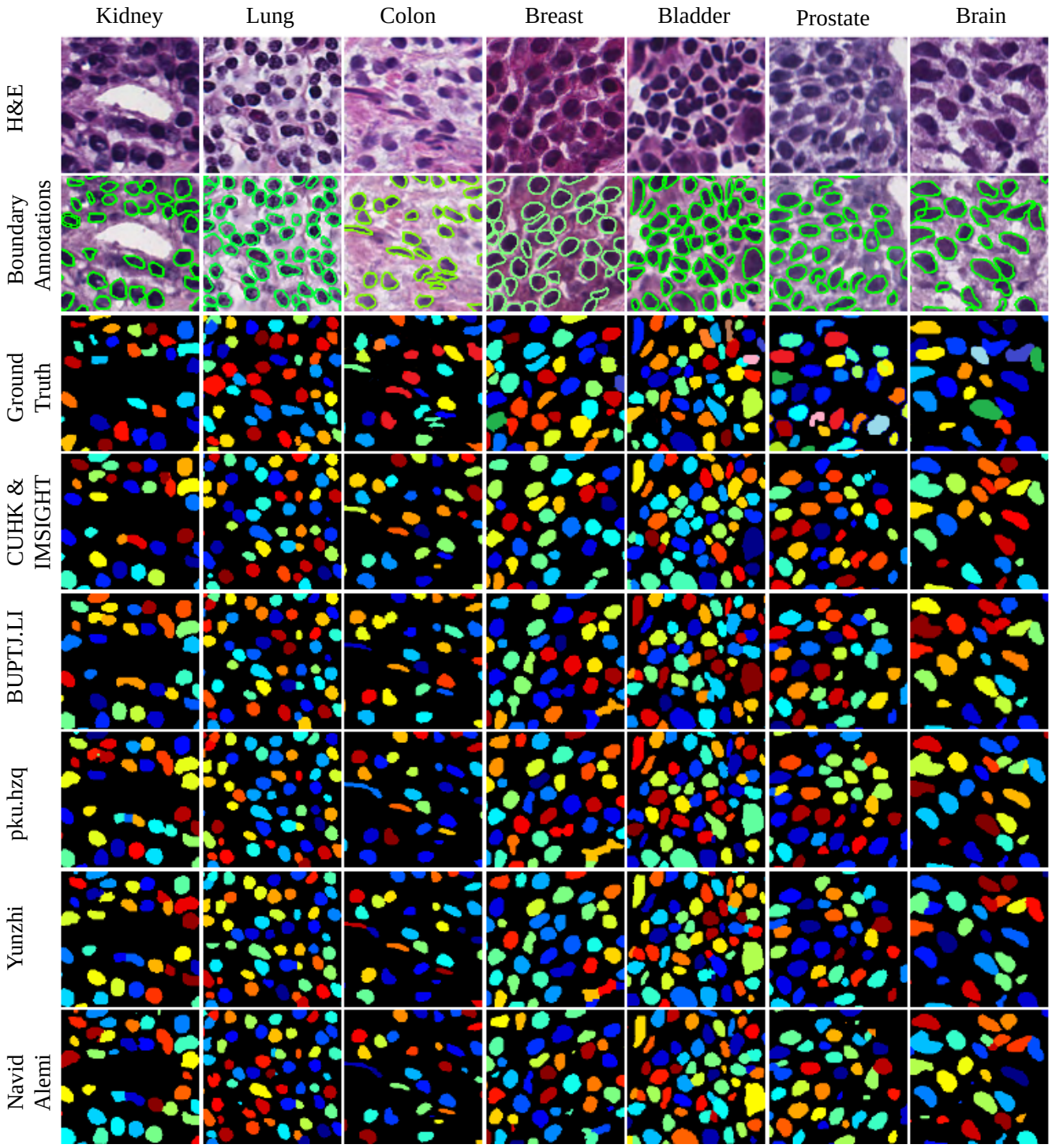


Fig. 2: Test sub-images taken from different organs exemplifying challenges of working with varied nuclear appearances and crowding patterns are shown in columns. Original H&E images, nuclear boundary annotations and segmentation results from the top-five techniques are shown in rows.

TABLE III: Comparison of techniques that completed the MONuSeg challenge

Team Name	AJI	Pre-Proc.				Data Augmentation									Model and Arch.								Loss				Post-Proc.			Additional Notes
		Color Norm.	Unit Var.	Range Stand.	Hist. Eq.	Rotation	Flipping	Affine deform.	Scaling	Elastic deform.	Noise addition	Color jitter	Intensity jitter	Blur/sharpen	U-Net	Mask RCNN	FCN	PAN	ResNet	VGG-Net	DenseNet	Distance Map	Cross Entropy	Dice loss	L1 loss	L2 loss	Watershed seg.	Non-max supp.	Morph. ops.	
CUHK & IMSIGHT	0.691	✓				✓		✓				✓					✓		✓				✓								Macenko CN [44]; multi-headed FCN
BUPT.J.LI	0.687	✓				✓		✓			✓												✓			✓					SPCN [43]; Deep Layer Aggregation; geometric instance vector
pku.hzq	0.685	✓	✓			✓		✓			✓				✓	✓			✓				✓		✓			✓	✓		Macenko CN [44]; ResNet with feature pyramid network
Yunzhi	0.679		✓		✓	✓		✓			✓	✓			✓				✓				✓				✓		✓		Cascaded U-Net with ResNet-like arch.
Navid Alemi	0.678					✓		✓			✓	✓											✓	✓			✓				Concatenated HSV and L channels to RGB multi-headed sphagetti net; smooth Jaccard index for boundary detection; boundary map cleaned by frangi vesselness filter gave markers
xuhuaren	0.664		✓			✓		✓			✓	✓				✓			✓				✓		✓				✓		
aetherAI	0.663					✓		✓			✓	✓						✓	✓		✓		✓		✓						
Shuang Yang	0.662					✓		✓			✓	✓				✓		✓	✓				✓	✓				✓	✓		
Bio-totem & SYSUCC	0.662				✓	✓		✓			✓	✓				✓			✓				✓		✓						Color histogram equilization [52]
Amirreza Mahbod	0.657	✓		✓		✓		✓			✓	✓			✓							✓	✓	✓			✓	✓	✓		Macenko CN [44]; markers are filtered distance maps
CMU-UIUC	0.656	✓						✓		✓	✓		✓	✓	✓								✓	✓					✓		SPCN [43]; nuclei pasted on various BG for data aug.
Graham&Vu	0.653	✓				✓		✓			✓	✓	✓			✓			✓		✓	✓	✓		✓	✓	✓				SPCN [43]; combined detection and distance map pred.
Unblockabulls	0.651	✓				✓	✓		✓						✓								✓				✓				Macenko CN [44], concatenated hematoxylin channel
Tencent AI Lab	0.646	✓				✓	✓		✓		✓		✓	✓		✓			✓				✓		✓						Edge enhancement used in pre-processing to separate nuclei
BioImage	0.638	✓				✓	✓		✓		✓			✓	✓				✓				✓	✓							Four encoders - DPN92, ResNet152, InceptionV3, ResNet 34
Pvmed	0.637	✓	✓												✓				✓										✓		Macenko CN [44]
DeepMD	0.633	✓				✓	✓		✓		✓			✓	✓					✓			✓	✓			✓				SPCN [43]; random sharpening; TernausNet architecture
Antanas Kascenas	0.633			✓		✓	✓		✓	✓					✓					✓		✓	✓								
Johannes Stegmaier	0.623			✓		✓		✓				✓			✓							✓	✓				✓				
Yanping	0.623	✓				✓	✓		✓		✓			✓	✓	✓			✓								✓		✓		SPCN [43]
Philipp Gruening	0.621				✓	✓	✓		✓					✓	✓					✓			✓				✓				Cross entropy and squared cosine losses
Agilent Labs	0.618					✓		✓		✓		✓			✓							✓				✓	✓	✓	✓		
Konica Minolta Lab EU	0.611					✓	✓			✓					✓					✓				✓					✓		
OnePiece	0.606	✓													✓								✓				✓				SPCN [43]
Junma	0.593	✓		✓			✓																	✓			✓				SPCN [43], blue channel extraction
Biosciences R&D, TCS	0.578					✓	✓								✓												✓				Smooth Jaccard index; separate nuclei using convexity
Azam Khan	0.575					✓	✓	✓				✓		✓		✓			✓				✓		✓			✓			
CVBLab	0.574	✓													✓									✓			✓				SPCN [43]
Linmin Pei	0.562	✓				✓	✓								✓											✓	✓				SPCN [43]
DeepMedicine.ai	0.460				✓	✓	✓															✓	✓					✓	✓		CLAHE [46]; high boost filtering to separate touching nuclei
DB-KR-JU	0.455	✓				✓																	✓				✓		✓		Reinhard [45]+Macenko [44] CN; separate nuc. using circularity
VISILAB	0.444	✓				✓	✓													✓			✓				✓				Macenko CN [44]
Sabarinathan	0.444			✓	✓	✓	✓	✓	✓				✓	✓	✓								✓	✓							CLAHE [46]
Silvers	0.278				✓	✓		✓	✓			✓	✓		✓						✓		✓						✓		Combination of DenseNet and U-Net
Rebecca Stone	0.265					✓	✓					✓			✓																Custom arch. for nuclei detection and U-Net for segmentation
TJ	0.130	✓				✓	✓			✓					✓		✓			✓			✓							✓		SPCN [43]; ensemble of multiple architectures

2) BUPTI.J.LI: Images were pre-processed by color normalization. Training data was augmented using random cropping, flipping, rotation, scaling, and noise addition. Using an FCN, which is much faster than MRCNN, not only semantic segmentation (nuclear vs. non-nuclear) was predicted, but whether the pixel was near the geometric center of a nucleus (core) was also predicted. The central region of each nucleus was obtained by eroding the mask using a 5x5 structuring element. Additionally, a vector to the nearest nuclear center was also predicted. Deep layer aggregation [47] was used as the based model. The center vector predictions were clustered and their ensemble with the geometric center predictions was considered to locate nuclear centers.

3) pku.hzq: Extensive data augmentation was used such as flips, rotations, scaling, and noise addition. Then a U-Net [2] was used to predict a ternary class map similar to Kumar *et al*. [1]. Additionally, an MRCNN [4] was used for top-down instance segmentation. Predictions from the two models were combined as an ensemble for both boundary and nucleus prediction. Two U-Nets were used in a cascade. The first one predicted probability of a pixel belonging to a nucleus as well as the vector pointing to the nucleus. The output of the final feature layer of this U-Net was concatenated with the input image and fed to the second U-Net that predicted the nuclear boundaries. A random walker or a watershed segmentation algorithm used the nuclear markers along with the boundary predictions for the final instance segmentation.

4) Yunzhi: For data preparation contrast-limited adaptive histogram equalization (CLAHE) [46] was used. Data augmentation was done using mirror flipping, rotations that were multiples of 90 degrees, color jitter, Gaussian noise addition, and elastic deformation. For each pixel, the probability of it belonging to a nucleus, or a nucleus boundary was predicted. Additionally, a vector to the center of the pixel was also predicted.

5) Navid Alemi: Primarily, marker-controlled watershed segmentation was used. A neural network predicted both foreground (nuclear core) and background (nuclear boundary) markers. The neural network was a multi-scale feature-sharing network. The network used extensive skip connections, and was dubbed SpaghettiNet. For training the marker head prediction, the network uses a combination of weighted Dice and binary cross entropy. For predicting the boundaries, it uses a smooth Jaccard loss. The boundary map was cleaned up using Frangi vesselness filter [53]. The markers were used as foreground seeds for watershed segmentation.

V. Discussion and Conclusion

Some clear trends emerged from analyzing the top few techniques in Table III. As shown previously [1], [54], color normalization plays a role in increasing pixel classification accuracy. Vahadane [43] and Macenko [44] were the most popular color normalization techniques due to the availability of their codes. Most of the top techniques relied on heavy data augmentation including affine transformations and noise addition. ResNet [5] seems to be an architecture of choice for several top performers irrespective of how they formulated the learning task. The residual skip connections in ResNet allow backpropagation of gradient deep into the network without dilution. Most of the highly successful networks stuck to predicting pixel class probabilities or using MRCNN [4] to predict instance maps. Watershed segmentation was among the most heavily utilized post-processing techniques. It was applied to the nuclear probability maps, most often coupled with a marker, where the marker was based on detecting the cores of individual nuclei. Some of the general trends observed corroborated those found in instance segmentation challenges of general photography images such as Common Objects in Context (COCO) Challenge [55].

By examining the sample test results in Figure 2 the performance of the nucleus segmentation techniques seems visually quite satisfactory. While more improvements are possible and welcome, for example by using generative adversarial networks to generalize over organ and stain variations, it seems fully-automated nucleus segmentation techniques, at least for H&E stained images captured at 40x, have reached a usable level of maturity to facilitate future studies based on nuclear morphometric biomarkers. The dataset and the techniques that have emerged as a part of the MONuSeg challenge can also be adapted for multi-scale and multi-stain nucleus detection.

References

[1] N. Kumar, R. Verma, S. Sharma, S. Bhargava, A. Vahadane, and A. Sethi, "A dataset and a technique for generalized nuclear segmentation for computational pathology," *IEEE Transactions on Medical Imaging*, vol. 36, no. 7, pp. 1550–1560, July 2017.

[2] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[3] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[4] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[5] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[6] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[7] S. Naik, S. Doyle, S. Agner, A. Madabhushi, M. Feldman, and J. Tomaszewski, "Automated gland and nuclei segmentation for grading of prostate and breast cancer histopathology," in *2008 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, May 2008, pp. 284–287.

[8] A. H. Beck, A. R. Sangoi, S. Leung, R. J. Marinelli, T. O. Nielsen, M. J. van de Vijver, R. B. West, M. van de Rijn, and D. Koller, "Systematic analysis of breast cancer morphology uncovers stromal features associated with survival," *Science Translational Medicine*, vol. 3, no. 108, pp. 108ra113–108ra113, 2011. [Online]. Available: http://stm.sciencemag.org/content/3/108/108ra113

[9] H. Chang, J. Han, A. Borowsky, L. Loss, J. W. Gray, P. T. Spellman, and B. Parvin, "Invariant delineation of nuclear architecture in glioblastoma multiforme for clinical and molecular association," *IEEE Transactions on Medical Imaging*, vol. 32, no. 4, pp. 670–682, April 2013.

[10] P. Filipczuk, T. Fevens, A. Krzyak, and R. Monczak, "Computer-aided breast cancer diagnosis based on the analysis of cytological images of fine needle biopsies," *IEEE Transactions on Medical Imaging*, vol. 32, no. 12, pp. 2169–2178, Dec 2013.

[11] A. S. et. *al.*, "Computational pathology for predicting prostate cancer recurrence," *Proceedings of AACR* 106^{th} *annual meeting*, August 2015.

[12] J. H. Xue and D. M. Titterington, "t -tests, f -tests and otsu's methods for image thresholding," *IEEE Transactions on Image Processing*, vol. 20, no. 8, pp. 2392–2396, Aug 2011.

[13] X. Yang, H. Li, and X. Zhou, "Nuclei segmentation using marker-controlled watershed, tracking using mean-shift, and kalman filter in time-lapse microscopy," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 53, no. 11, pp. 2405–2414, Nov 2006.

[14] M. Veta, P. J. van Diest, R. Kornegoor, A. Huisman, M. A. Viergever, and J. P. Pluim, "Automatic nuclei segmentation in h&e stained breast cancer histopathology images," *PloS one*, vol. 8, no. 7, p. e70221, 2013.

[15] A. Vahadane and A. Sethi, "Towards generalized nuclear segmentation in histological images," in *Bioinformatics and Bioengineering (BIBE), 2013 IEEE 13th International Conference on*, Nov 2013, pp. 1–4.

[16] J. Schindelin, I. Arganda-Carreras, E. Frise, V. Kaynig, M. Longair, T. Pietzsch, S. Preibisch, C. Rueden, S. Saalfeld, B. Schmid, J.-Y. Tinevez, D. J. White, V. Hartenstein, K. Eliceiri, P. Tomancak, and A. Cardona, "Fiji: an open-source platform for biological-image analysis," *Nature Methods*, vol. 9, no. 7, pp. 676–682, Jul. 2012. [Online]. Available: http://dx.doi.org/10.1038/nmeth.2019

[17] A. E. Carpenter, T. R. Jones, M. R. Lamprecht, C. Clarke, I. H. Kang, O. Friman, D. A. Guertin, J. H. Chang, R. A. Lindquist, J. Moffat, P. Golland, and D. M. Sabatini, "Cellprofiler: image analysis software for identifying and quantifying cell phenotypes," *Genome Biology*, vol. 7, no. 10, p. R100, 2006. [Online]. Available: http://dx.doi.org/10.1186/gb-2006-7-10-r100

[18] "The cancer genome atlas (tcga)," http://cancergenome.nih.gov/.

[19] "Multi-organ nuclei segmentation challenge (MoNuSeg) 2018 [online]," https://monuseg.grand-challenge.org/, accessed 11 Feb. 2019.

[20] R. Verma, N. Kumar, A. Sethi, and P. H. Gann, "Detecting multiple subtypes of breast cancer in a single patient," in *2016 IEEE International Conference on Image Processing (ICIP)*, Sept 2016, pp. 2648–2652.

[21] G. Lee, R. Sparks, S. Ali, N. N. C. Shih, M. D. Feldman, E. Spangler, T. Rebbeck, J. E. Tomaszewski, and A. Madabhushi, "Co-occurring gland angularity in localized subgraphs: predicting biochemical recurrence in intermediate-risk prostate cancer patients." *PloS one*, vol. 9, p. e97954, May 2014.

[22] J. Cheng and J. C. Rajapakse, "Segmentation of clustered nuclei with shape markers and marking function," *IEEE Transactions on Biomedical Engineering*, vol. 56, no. 3, pp. 741–748, March 2009.

[23] S. Ali and A. Madabhushi, "An integrated region-, boundary-, shape-based active contour for multiple object overlap resolution in histological imagery," *IEEE Transactions on Medical Imaging*, vol. 31, no. 7, pp. 1448–1460, July 2012.

[24] Y. Al-Kofahi, W. Lassoued, W. Lee, and B. Roysam, "Improved automatic detection and segmentation of cell nuclei in histopathology images," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 4, pp. 841–852, April 2010.

[25] S. Wienert, D. Heim, K. Saeger, A. Stenzinger, M. Beil, P. Hufnagl, M. Dietel, C. Denkert, and F. Klauschen, "Detection and segmentation of cell nuclei in virtual microscopy images: a minimum-model approach," *Scientific reports*, vol. 2, p. 503, Nov. 2012.

[26] H. Kong, M. Gurcan, and K. Belkacem-Boussaid, "Partitioning histopathological images: An integrated framework for supervised color-texture segmentation and cell splitting," *IEEE Transactions on Medical Imaging*, vol. 30, no. 9, pp. 1661–1677, Sept 2011.

[27] M. E. Plissiti and C. Nikou, "Overlapping cell nuclei segmentation using a spatially adaptive active physical model," *IEEE Transactions on Image Processing*, vol. 21, no. 11, pp. 4568–4580, Nov 2012.

[28] H. Chang, J. Han, A. Borowsky, L. Loss, J. W. Gray, P. T. Spellman, and B. Parvin, "Invariant delineation of nuclear architecture in glioblastoma multiforme for clinical and molecular association," *IEEE Transactions on Medical Imaging*, vol. 32, no. 4, pp. 670–682, April 2013.

[29] M. Zhang, T. Wu, and K. M. Bennett, "Small blob identification in medical images using regional features from optimum scale," *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 4, pp. 1051–1062, April 2015.

[30] Y. Song, L. Zhang, S. Chen, D. Ni, B. Lei, and T. Wang, "Accurate segmentation of cervical cytoplasm and nuclei based on multiscale convolutional network and graph partitioning," *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 10, pp. 2421–2433, Oct 2015.

[31] F. Xing, Y. Xie, and L. Yang, "An automatic learning-based framework for robust nucleus segmentation," *IEEE Transactions on Medical Imaging*, vol. 35, no. 2, pp. 550–566, Feb 2016.

[32] K. Sirinukunwattana and S. E. A. R. et. *al.*, "Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1196–1206, May 2016.

[33] S. Graham, Q. D. Vu, S. e Ahmed Raza, J. T. Kwak, and N. M. Rajpoot, "XY network for nuclear segmentation in multi-tissue histology images," *CoRR*, vol. abs/1812.06499, 2018. [Online]. Available: http://arxiv.org/abs/1812.06499

[34] P. Naylor, M. La, F. Reyal, and T. Walter, "Segmentation of nuclei in histopathology images by deep regression of the distance map," *IEEE Transactions on Medical Imaging*, vol. 38, no. 2, pp. 448–459, Feb 2019.

[35] H. Irshad, A. Veillard, L. Roux, and D. Racoceanu, "Methods for nuclei detection, segmentation, and classification in digital histopathology: A review, current status and future potential," *IEEE Reviews in Biomedical Engineering*, vol. 7, pp. 97–114, 2014.

[36] F. Xing and L. Yang, "Robust nucleus/cell detection and segmentation in digital pathology and microscopy images: A comprehensive review," *IEEE Reviews in Biomedical Engineering*, vol. 9, pp. 234–263, 2016.

[37] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[38] A. Krizhevsky, "Learning multiple layers of features from tiny images," *Technical Report*, 2009.

[39] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint*, vol. abs/1212.0402, 2012. [Online]. Available: http://arxiv.org/abs/1212.0402

[40] A. Janowczyk and A. Madabhushi, "Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases," *Journal of Pathology Informatics*, vol. 7, July 2016.

[41] H. Irshad, L. Montaser-Kouhsari, G. Waltz, O. Bucur, J. Nowak, F. Dong, N. W. Knoblauch, and A. H. Beck, "Crowdsourcing image annotation for nucleus detection and segmentation in computational pathology: evaluating experts, automated methods, and the crowd." in *Pacific Symposium on Biocomputing*, 2015, pp. 294–305.

[42] E. D. Gelasca, B. Obara, D. Fedorov, K. Kvilekval, and B. Manjunath, "A biosegmentation benchmark for evaluation of bioimage analysis methods," *BMC Bioinformatics*, vol. 10, no. 1, p. 368, 2009.

[43] A. Vahadane, T. Peng, A. Sethi, S. Albarqouni, L. Wang, M. Baust, K. Steiger, A. M. Schlitter, I. Esposito, and N. Navab, "Structure-preserving color normalization and sparse stain separation for histological images," *IEEE Transactions on Medical Imaging*, vol. 35, no. 8, pp. 1962–1971, Aug 2016.

[44] M. Macenko, M. Niethammer, J. S. Marron, D. Borland, J. T. Woosley, X. Guan, C. Schmitt, and N. E. Thomas, "A method for normalizing histology slides for quantitative analysis," in *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. IEEE, 2009, pp. 1107–1110.

[45] E. Reinhard, M. Adhikhmin, B. Gooch, and P. Shirley, "Color transfer between images," *IEEE Computer graphics and applications*, vol. 21, no. 5, pp. 34–41, 2001.

[46] S. M. Pizer, E. P. Amburn, J. D. Austin, R. Cromartie, A. Geselowitz, T. Greer, B. ter Haar Romeny, J. B. Zimmerman, and K. Zuiderveld, "Adaptive histogram equalization and its variations," *Computer vision, graphics, and image processing*, vol. 39, no. 3, pp. 355–368, 1987.

[47] F. Yu, D. Wang, E. Shelhamer, and T. Darrell, "Deep layer aggregation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2403–2412.

[48] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8759–8768.

[49] V. Iglovikov and A. Shvets, "Ternausnet: U-net with vgg11 encoder pre-trained on imagenet for image segmentation," *arXiv preprint arXiv:1801.05746*, 2018.

[50] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.

[51] P. Jaccard, "Étude comparative de la distribution florale dans une portion des alpes et des jura," *Bull Soc Vaudoise Sci Nat*, vol. 37, pp. 547–579, 1901.

[52] A. Janowczyk. (2018) On stain normalization in deep learning. [Online]. Available: http://www.andrewjanowczyk.com/on-stain-normalization-in-deep-learning/

[53] A. F. Frangi, W. J. Niessen, K. L. Vincken, and M. A. Viergever, "Multiscale vessel enhancement filtering," in *International conference on medical image computing and computer-assisted intervention*. Springer, 1998, pp. 130–137.

[54] A. S. et. *al.*, "Empirical comparison of color normalization methods for epithelial-stromal classification in h and e images," *Journal of Pathology Informatics*, vol. 7, 2016.

[55] T.-Y. Lin and M. e. e. Maire, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

Enhancing Endoscopic Image Classification with Symptom Localization and Data Augmentation

Trung-Hieu Hoang
hthieu@selab.hcmus.edu.vn
University of Science, VNU-HCM
Vietnam

Hai-Dang Nguyen
nguyenhd@eurecom.fr
Eurecom
France

Viet-Anh Nguyen
nvanh160013@ump.edu.vn
University of Medicine and Pharmacy,
Ho Chi Minh city
VietNam

Thanh-An Nguyen
ntan@selab.hcmus.edu.vn
University of Science, VNU-HCM
Vietnam

Vinh-Tiep Nguyen
tiepnv@uit.edu.vn
University of Information Technology,
VNU-HCM
Vietnam

Minh-Triet Tran
tmtriet@fit.hcmus.edu.vn
University of Science, VNU-HCM
Vietnam

ABSTRACT

Inspired by recent advances in computer vision and deep learning, we proposed new enhancements to tackle problems appearing in endoscopic image analysis, especially abnormalities findings and anatomical landmarks detection. In details, a combination of **Residual Neural Network** and **Faster R-CNN** are applied jointly in order to take all of their advantages and improve the overall performance. Nevertheless, novel data augmentation has been designed and adapted to corresponding domains. Our approaches prove their competitive results in term of not only the accuracy but also the inference time in Medico: The 2018 Multimedia for Medicine Task and The Biomedia ACM MM Grand Challenge 2019. These results show the great potential of the collaborating between deep learning models and data augmentation in medical image analysis applications.

KEYWORDS

endoscopic image, symptoms localization, anatomical landmarks detection, object detection, image classification

1 INTRODUCTION

The abnormalities finding and landmark detection in endoscopy images challenge aim to bring new achievements in computer vision, image processing and machine learning to the next level of computer and multimedia assisted diagnosis. In order to encourage the research community to tackle the problems with endoscopy images, besides proposing new endoscopic images dataset [8], several challenges are conducted, such as the Medico: Multimedia Task at MediaEval 2018 [7] and The Biomedia ACM MM Grand Challenge 2019 [13]. The goal of these challenges is encouraging research communities to establish an assisted diagnosis systems that can detect abnormalities and anatomy landmarks in human gastrointestinal tract automatically in an efficient way with as less training data as possible.

In our approach, we introduce a stacked model consisting of two deep networks, a Residual Neural Network (Resnet) [2] followed by a Faster Region-based Convolutional Neural Network (Faster R-CNN) [10]. As our observation, the Resnet mostly focuses on deep global features of image, it fails to classify images that symptoms of *abnormal symptoms* or *instruments* appear as small objects on diversity backgrounds. Therefore, we aim to apply the Faster R-CNN to re-classify the images of some classes that Resnet usually mis-classify.

Besides, due to the limitation and the imbalance between classes in the training samples and using extra source of endoscopic data is infeasible, we proposed a novel data augmentation mechanism which can enhance the performance of both models. Nevertheless, a multi-tasks classifiers is introduced to reduce the confusion level of classifier module in cases that multiple symptoms of diseases appeared in the same image.

2 RELATED WORK

Early approaches that tackle with endoscopic image analysis usually work on the bleeding detection problem. By using color threshold and applying some handcrafted features, Shah et al.[6] and Jung et al.[5] propose several solutions using color domain and region segmentation. Later, supervised learning machine techniques are used jointly with superpixel saliency to identify bleeding regions of by recognizing blood color patterns [3].

Since introduced, deep neural networks have been used in order to solve several problems in the field of analyzing endoscopic images of the gastrointestinal (GI) tract. Particularly, to localize and identify polyps within real-time constraint, deep CNNs has recently shown an impressive potential when achieving up to 96.4% accuracy - published in 2018 by Urban G et al. [14]. Another interesting article of Satoki Shichijo et al. [12] also applies multiple deep CNNs to diagnose Helicobacter pylori gastritis based on endoscopic images. Further, gastrointestinal bleeding detection using deep CNNs on endoscopic images has been successfully done and published by Xiao Jia et al. [4].

3 APPROACH

3.1 Overview of proposed methods

As mentioned before, a stacked model consists of a Residual Neural Network (ResNet) and a Faster Region-based Convolutional Neural Network (Faster R-CNN). In order to training the Faster R-CNN model, addition information regarding to the location of abnormal symptom need to be annotated which is described in details in

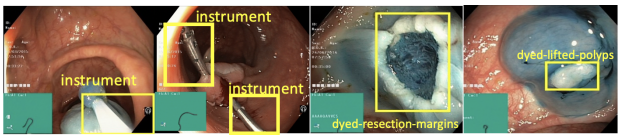


Figure 1: Our proposed symptoms region localization, annotated images with bounding boxes and classes name.

section 3.2. Additionally, unbalancing between classes can make deep models bias to some classes than others, which reduce the overall accuracy. In the context of these challenges that require using as less as training data as possible, therefore, using extra data from other dataset is infeasible. In Section 3.3, we also provide some augmentation mechanism on the given training dataset in order to provide a better version for the training step of both modules.

Nevertheless, inference time must be taken into account, while the Faster R-CNN needs longer time to process than that of ResNet. Instead of feeding all images through the ResNet module, we should have different strategies in Section 3.5 that reduce the number of images need to go through the Faster R-CNN module and as result, balancing between the accuracy and inference time.

Last but not least, an other improvement of our method is using a multi-task classification architecture that can predict multiple classes at the same time, which can reduce the confusing level of the image classifier model in these difficult cases. Further information about this architecture is presented in Section 3.6.

3.2 From Classification to Symptoms Region Localization

An object detection module can be really useful to detect small abnormal symptoms and diseases, such as *polyps, instruments, dyed-lifted-polyps* and *dyed-resection-margins*. Besides, in some cases that multi-symptoms appear in the same image, identify all of them is necessary to draw the final conclusion. Nevertheless, system that can not only predict the abnormalities but also propose the corresponding positions of that is more reliable and more convenient for endoscopists.

Due to the limitation of the given training dataset, there is no extra information about positions of abnormalities corresponding to each endoscopic image. Therefore, we decide to annotate all the abnormal symptoms in every images of the following classes: *dyed-resection-margins, dyed-lifted-polyps, instruments* and *polyps*. Examples of our proposed bounding box can be seen in Figure 1.

Totally, 5241 images belonged to the mentioned classes, are annotated with 5715 bounding boxes from the Kvasir dataset [8] and Medico 2018 development dataset. Although, the number of training samples we annotated is much larger than the number of actual training samples used in training phase, it is still useful for future works.

3.3 *Instruments, polyps* dataset Augmentation

Noticeably, *Instruments* - the second highest priority class has only 36 training samples. In order to maintain the balancing between all of these classes and also improve the diversity of the *instruments* images, we aim to augment the given dataset by generate more

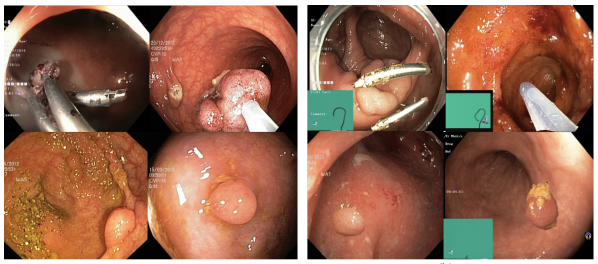


Figure 2: Dataset augmentation result. (a) original samples (b) augmented samples generated by our method

images for the *instruments* based on the current given development set.

From a given image, we carefully crop the region that contains symptom of diseases or *instruments* along their edges. Then, we randomly select images from other classes and use them as the background of the cropped instruments. Random affine transformation are also applied. With this strategy, we are able to generate more than 800 images for the target classes. As can be seen in Figure 2, there is not significant visual differences between original and augmented samples.

The augmented dataset can be used to enhance the robustness of the classification model. On the other hand, we can easily get the position of the foreground object on generated images, which can be utilized to augment the training dataset of the objects detection module.

The augmented dataset can be used for both sub-task. By adding more positive examples to the training phase of the classification model, the performance of this model can be more robust. On the other hand, we can easily get the position of the foreground object on generated images, which can be utilized to augment the training dataset of the objects detection module (Faster R-CNN).

3.4 Fine-tuning Deep Neural Network on Endoscopic images

Besides high computational cost, one of the main drawback of deep learning architecture is that it requires a large amount of training data. Moreover, labeled medical data for supervised learning is limited and manual labelling of medical images is a difficult task. Therefore, it requires a lot of effort and time to train the network, which would depend on the size of training data used. However, there is a possible solution to deal with these limitations is using transfer learning, where a pre-trained network on a large dataset (such as ImageNet [11]) is used.

In our approach, both Residual Network with 101 layers and Faster R-CNN [1] are both share a same features encoder. Therefore, it is necessary to propose a features encoder that is specialized on endoscopic images. This is the reason that we fine-tuned our deep neural network models (pre-trained on ImageNet) by using our modified development dataset. After training the whole neural network and then we freeze several first layers in its architecture and fine-tune the remains with small learning-rate. We also tried to train the network from scratch and all of our experiments point

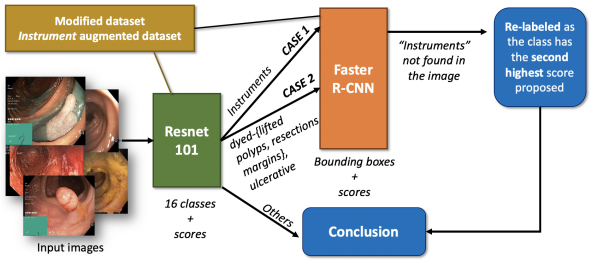


Figure 3: Configuration of the Third scenario (best performance).

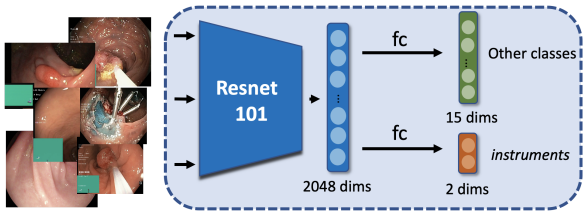


Figure 4: Overview of the multi-classes classifier network.

out that in term of using convolution neural network for medical images, knowledge transferring from natural images to medical images is possible, even though there is a large difference between the source and the target databases.

This idea is also mentioned in [9], it is especially useful in the case of small dataset of images provided. Fine-tuning on the ImageNet pre-trained model significantly improves the efficient of deep learning model on medical domains.

3.5 Configuration of Conditional Scenarios

As mentioned before, there is a trade-off between the inference time and the accuracy. The following section describes the pipeline for each of our scenarios that we proposed in order to evaluate the performance in term of the accuracy and inference time.

First scenario: instruments, polyps double-checked. Residual network with 101 layers model are fine-tuned on the original development set provided by the task organizers along with our instruments increased dataset. After passed through ResNet101, output images classified as special classes become the input of Faster R-CNN network, which is trained for detecting instruments in images.

- First case: Images predicted as *instruments* by Resnet101 are double-checked. In case instruments are not detected by Faster R-CNN in those images, they are re-labeled as the class of their second highest score proposed by Resnet101.
- Second case: Images predicted as *dyed-lifted-polyps, dyed-resection-margins, ulcerative colitis* by ResNet101 are fed forward through Faster R-CNN network to detect *instruments*. They are classified as *instruments* if detected or keep the original prediction otherwise.

Second scenario: instruments double-checked. Feeding forward a large number of images in the three classes through Faster R-CNN causes a bottle-neck of inference time, as Faster R-CNN has high time complexity. Therefore, in this second , we limited the images passed through Faster R-CNN by only performing the first case of the first scenario.

Third scenario: instruments double-checked and data augmentation. The configuration of the third scenario is as same as the second scenario which is illustrated in Figure 3. Instead of using the original training set mentioned in the first scenario, we train our model on the re-labeled development set combined with the augmented instrument set.

Forth scenario: 75% of training set. In this scenario, we reduce the number of images used for training by selecting randomly 75% images of each class in the same training set as the third scenario. Other processing steps are also configured in the same way.

Fifth scenario: esophagitis priority. Throughout our experiments, *normal-z-line* and *esophagitis* are the top most confusing classes not only for Resnet101 but also for human to distinguish them. In the priority list, *esophagitis* has a higher rank than *normal-z-line's*. Thus, after several times evaluating our model on the development dataset, we propose a condition for these two classes when they are predicted by ResNet101. As ResNet101 provides a probability distribution over the 16 classes for each image, whenever the *normal-z-line* appears to be the highest class, we add a small bias 0.3 to the probability of the *esophagitis*. Hence, the model is more likely to emit the *esophagitis* class. This intuitively means that our model prefers *esophagitis* to *normal-z-line* when it is confused between these classes.

3.6 Solving Multi-classes Problem with Multi-tasks Classifier

After Medico 2018, another improvement of our work is to reduce the confusion level of the deep neural network model in cases that various type of abnormalities appeared in a same image. Instead of using only one classifier and forcing the deep neural network model to follow the priority list in these cases by feeding a number of positive and negative samples, we narrow down this job to multiple classifiers. For instance, given an image that both *esophagitis* and *instruments* appear simultaneously, the job to determine whether or not *instruments* appear inside the image is then left for a 2-classes classifier. This classifier can output the probability that the given image contains *instruments*. Meanwhile, the second classifier works independently, which can output the probability for other classes, except *instruments*.

Architecture. In order to reduce the inference time, we decide to share the weights of backbone ResNet 101 for both classifier. The overview of this module can be seen in Figure 4. In general, the proposed multi-task classification model consists of a ResNet 101 architecture except the last fully connected layer, working as a features extractor module that can output a 2048 dimensions vector for each input image. There are two fully connected branches on top of the output of that features extractor in order to get the prediction of *instruments* class and other 15 *classes*. The number of classifiers is extendable in the future.

Table 1: Official evaluation result of Medico: : The 2018 Multimedia for Medicine Task for both sub-tasks (provided by the organizers) and speed (fps) on Tesla K80 GPU

RunID	PREC	REC	ACC	F1	MCC	RK	FPS
Run01	94.245	94.245	99.281	94.245	93.861	93.590	6.589
Run02	93.959	93.959	99.245	93.959	93.556	93.273	**23.191**
Run03	**94.600**	**94.600**	**99.325**	**94.600**	**94.240**	**93.987**	23.148
Run04	93.043	93.043	99.130	93.043	92.579	92.257	22.654
Run05	94.508	94.508	99.314	94.508	94.142	93.884	21.413

Loss function. Two classifiers are trained simultaneously with the overall loss function is a weighted sum of each of their loss, given as follow

$$\mathcal{L}(p_i, p_o, p_i^*, p_o^*) = \lambda \sum_i \mathcal{L}_{instr}(p_i, p_i^*) + (1-\lambda) \sum_i \mathcal{L}_{others}(p_o, p_o^*) \tag{1}$$

where (p_i, p_i^*) and (p_o, p_o^*) denote the prediction and the ground-truth of *instr* class and other classes, respectively. $\mathcal{L}$ is Cross Entropy Loss function. λ is a combination weight.

Final prediction. Since, there are two output vectors from the model, the final prediction can be determined by

$$y_f = \begin{cases} argmax(p_o) & p_i < 0.5 \\ y_{instruments} & p_i \geq 0.5 \end{cases} \tag{2}$$

where y_f stands for the final prediction of input image, p_o is a 15 dimensions vector indicates the probability that image is likely to belong to. p_i is the probability that *instruments* appear inside the image and $y_{instruments}$ is the corresponding label of *instruments* class.

4 EXPERIMENTAL SETUP

Medico: The 2018 Multimedia for Medicine Task

The configuration of *Run*01 to *Run*05 corresponding to five scenarios descried in section 3.5.

Biomedia ACM MM Grand Challenge 2019

We continue to tackle existing problems that we did not solve efficiently in Medico 2018, which are the confusing between *normal-z-line* and *esoiphagitis*; multi-classes problem with *instruments* classes. Besides increasing the size of training data with our augmentation strategy, there are two major improvements in this challenge.

(1) With the multi-classes problem, **we applied the Multi-tasks Classifier** (*MUL*) which is introduced in Section 3.6.
(2) With the *normal-z-line* and *esophagitis*, with a help of a medical expert, we re-annotate the labels of the original dataset and train our models on the modified version of the dataset (*RE_LBL*).

5 RESULTS

As illustrated in Table 1, there is a trade-off between speed and accuracy when comparing the result of *Run*01 and *Run*02. When reducing a large number of images passing through Faster R-CNN for the sake of time, so its performance seems to be relatively worse than *Run*01's.

Table 2: Official evaluation result of The Biomedia ACM MM Grand Challenge 2019 for both sub-tasks (provided by the organizers) and speed (fps) on GTX 1080 Ti GPU

RunID	PREC	REC	F1	MCC	FPS
SIN	0.8565	0.8503	0.8458	0.9126	3.5461
MUL	**0.8763**	**0.8771**	**0.8746**	**0.9406**	**3.6101**
SIN+RE_LBL	0.8667	0.8567	0.8483	0.9168	3.5842
MUL+RE_LBL	0.8698	0.8573	0.8491	0.9202	3.5842

As we mentioned earlier, training samples takes an important role in building a deep-neural network model. Through our experiments, in the case of less training data, the augmented dataset helps us improve the performance of deep-neural network model. Run03 and Run05 show impressive results comparing to the first two runs. This implies that training on our re-labeled development set provides better models.

On the other hand, using the Residual neural network cannot classify efficiently the two classes *esophagitis* and *normal-z-line*. The same problem also occurs between the *dyed-resection-margins* and *dyed-lifted-polyps* classes.

Additionally, the configuration of *Run*05 intuitively prefers *esophagitis* to *normal-z-line*, which may leads to an increasing of the false-positive cases in the result.

By comparison to the others, *Run*04 has the lowest precision since it uses 75% of training data. Decreasing the amount of training samples of course affects the performance in deep-learning models. Nevertheless, the result is still acceptable when it decreases only a few percentages and its configuration is as same as *Run*03.

Regarding to the evaluation results in Table 2, the performance of the Multi-tasks classifiers has successfully proved to have better performance than the single-task classifier used in previous experiments.

Although, having competitive results on our own validation set, re-labeling the development dataset on *esophagitits* and *normal-z-line* approach seem to have worse performance on the official test set. Distinguishing these classes is an challenging problem, especially for computers since there are still disagreements between experts in some cases.

6 CONCLUSION AND FUTURE WORKS

Endoscopic image classification is a challenging problem because of the fine-grained images, less training data and require high accuracy. In our approach, we focus on enhancing the performance of image classification model, such as Residual Neural Network by using object detection model, such as Faster R-CNN that can localize small symptoms of diseases, which are useful evidences. Besides, data augmentation strategy can be applied to solve the limitation of training samples which is commonly occurred in medical datasets. Accuracy and inference time that we reach is acceptable and appropriate for real-time constraint. However, for future works, additional medical testing must be taken into account besides visual information to create a more robust approach to exploit the distinction between easy-confused classes.

ACKNOWLEDGMENTS

We would like to express our appreciation to Honors Program of Computer Science, Software-engineering Laboratory, University of Science, VNU-HCM

REFERENCES

[1] Xinlei Chen and Abhinav Gupta. An implementation of faster rcnn with study for region sampling. *arXiv preprint arXiv:1702.02138*, 2017.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[3] Dimitris Iakovidis, Dimitris Chatzis, Panos Chrysanthopoulos, and Anastasios Koulaouzidis. Blood detection in wireless capsule endoscope images based on salient superpixels. volume 2015, 08 2015.

[4] Xiao Jia and Max Q.-H. Meng. A deep convolutional neural network for bleeding detection in wireless capsule endoscopy images. *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2016.

[5] Y. S. Jung, Y. H. Kim, D. H. Lee, and J. H. Kim. Active blood detection in a high resolution capsule endoscopy using color spectrum transformation. In *2008 International Conference on BioMedical Engineering and Informatics*, volume 1, pages 859–862, May 2008.

[6] Subodh K Shah, Pragya P Rajauria, Jeongkyu Lee, and M. Emre Celebi. Classification of bleeding images in wireless capsule endoscopy using hsi color domain and region segmentation. 07 2019.

[7] PËŽal Halvorsen Thomas de Lange Kristin Ranheim Randel Duc-Tien Dang-Nguyen Mathias Lux-Olga Ostroukhova Konstantin Pogorelov, Michael Riegler. Medico multimedia task at mediaeval 2018. *Media Eval' 2018*, 2018.

[8] Konstantin Pogorelov, Kristin Ranheim Randel, Carsten Griwodz, Sigrun Losada Eskeland, Thomas de Lange, Dag Johansen, Concetto Spampinato, Duc-Tien Dang-Nguyen, Mathias Lux, Peter Thelin Schmidt, Michael Riegler, and Pål Halvorsen. Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection. In *Proceedings of the 8th ACM on Multimedia Systems Conference*, MMSys'17, pages 164–169, New York, NY, USA, 2017. ACM.

[9] Adnan Qayyum, Syed Anwar, Muhammad Majid, Muhammad Awais, and Majdi Alnowami. Medical image analysis using convolutional neural networks: A review. 42, 09 2017.

[10] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 91–99. Curran Associates, Inc., 2015.

[11] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Fei-Fei Li. Imagenet large scale visual recognition challenge. *CoRR*, abs/1409.0575, 2014.

[12] Aoyama Kazuharu Nishikawa Yoshitaka Miura Motoi Shinagawa Takahide Takiyama Hirotoshi Tanimoto Tetsuya Ishihara Soichiro-Matsuo Keigo-Tada Tomohiro Shichijo Satoki, Nomura Shuhei. Application of convolutional neural networks in the diagnosis of helicobacter pylori infection based on endoscopic images. *EBioMedicine*, 25:106–111, Nov 2017.

[13] Pia Smedsrud Trine B. Haugen Kristin Randheim Randel Konstantin Pogorelov HÃëkon Kvale Stensland Duc-Tien Dang-Nguyen Mathias Lux Andreas Petlund Thomas de Lange Peter Thelin Schmidt PÃël Halvorsen Steven Hicks, Michael Riegler. Acm mm biomedia 2019 grand challenge overview. 2019.

[14] Gregor Urban, Priyam Tripathi, Talal Alkayali, Mohit Mittal, Farid Jalali, William Karnes, and Pierre Baldi. Deep learning localizes and identifies polyps in real time with 96% accuracy in screening colonoscopy. *Gastroenterology*, 155(4), 2018.