Nuclei Detection Using Residual Attention Feature Pyramid Networks

Panagiotis Dimitrakopoulos[1], Giorgos Sfikas[1,2], and Christophoros Nikou[1]

[1]Dpt. of Computer Science and Engineering, University of Ioannina, 45110 Ioannina, Greece
[2]Information Technologies Institute, CERTH, 57001 Thessaloniki, Greece
Email: {pdimitrakopoulos, sfikas, cnikou}@cs.uoi.gr

***Abstract*—Detection of cell nuclei in microscopy images is a challenging research topic due to limitations in acquired image quality as well as due to the diversity of nuclear morphology. This has been a topic of enduring interest with promising success shown by deep learning methods. Recently, attention gating methods have been proposed and employed successfully in a diverse array of pattern recognition tasks. In this work, we introduce a novel attention module and integrate it with feature pyramid networks and the state-of-the-art Mask R-CNN network. We show with numerical experiments that the proposed model outperforms the state-of-the-art baseline.**

***Index Terms*—Nuclei detection, Attention gates, Feature Pyramid Network, Mask R-CNN**

I. Introduction

Cellular image analysis is a research area that is increasingly taking advantage of developments in generic machine vision and pattern recognition methods. Automated methods have been proposed and applied with sucess in various tasks, including image classification, segmentation, detection and tracking [1], [2]. In this paper, we are interested in automatic detection of cell nuclei in microscopy images. Challenges of this task include limitations in cellular image quality and diversity of nuclear morphology, which includes varying nuclei shapes, sizes, and overlaps between multiple cell nuclei. This has been a topic of enduring interest with recent success shown by deep learning methods [3].

The application of deep learning in digital image processing usually involves the use of a network architecture that can be categorized as a feed-forward, convolutional network. Under this family of models, information typically flows from the input towards the output from layer to layer in a sequential fashion [4]. An apparently simple, yet practically important development in terms of convolutional network architecture has been the introduction of skip connections [5], [6]. This involves the idea of combining feature maps at different scales / resolutions, with intermediate layers being fed input also from layers that do not directly precede them. Fully Convolutional Networks [6] and the celebrated U-Net architecture [5] were among the first works to have popularized this idea. The more recent Feature Pyramid Networks (FPNs) [7], also utilizing skip connections, have been successfully used as a network backbone of the Mask R-CNN instance segmentation model [8], in order to perform detection and localization of nuclei in cell images [3], [9], [10].

A more recent development is the concept of the Attention Gate (AG) [11]. Attention Gates act as a soft mask on concatenated intermediate layer inputs. These eventually weigh up activations over regions of interest, while weighing down task unrelated to the task. In the current work, we are using attention gates in order to build more sophisticated skip connections that can be used for detection tasks and improve overall efficiency. We propose a novel attention module architecture, applicable in perspective to a generic deep-learning based detection model. We fuse the proposed attention module to the FPN architecture as part of a baseline Mask R-CNN model. Numerical experiments show that the proposed attention-based detector can detect nuclei in a wide range of microscopy images of cell nuclei, outperforming the state-of-the-art Mask R-CNN detector.

The remainder of the paper is structured as follows. In the following section, section II, we discuss related work with respect to the key aspects of the current work: nuclei detection, feature pyramids and attention gates. In section III we review the basics concerning attention mechanisms and in section IV we discuss the state-of-the-art detection model that we aim to extend with attention, which is Mask-RCNN with a FPN backbone. We present the proposed attention mechanism in section V, and evaluate it with numerical experiments in section VI. The paper is concluded with a discussion on the paper contribution and future work in section VII.

II. Related work

Mask R-CNN is a state-of-the-art neural network that has originally been proposed for instance segmentation [8]. Its original formulation can take into account different classes of objects to be detected; it has already been used successfully as a nuclei detector in digitized cell images, where the class of interest is only one (vs background) [3], [9], [10]. The Mask R-CNN network architecture can be analyzed into two distinct components: a convolutional backbone and a network head. In the original paper [8], the most efficient backbone is found to be the Feature Pyramid Network backbone, which includes skip connections between layers at different scales. In the current work, we extend this backbone by fusing it with the proposed attention module.

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan XP GPU used for this research.

Attention mechanisms have been proposed and used for various learning tasks, ranging from natural language processing for image captioning and machine translation, to machine vision tasks such as detection, segmentation, person re-identification and many more [11]–[13]. Apart from fusions with Recurrent Neural Network (RNN) [14] standard Convolutional Neural Network (CNN) architectures [12], attention gates have been employed also in conjunction with Generative Adversarial Networks, as for example in [15] where the attention gate helps in weighing feature map regions related to an image translation task. Attention modules have also been generalized to a 'non-local' version, with non-local self-attention defined in [16], capable of capturing long-range spatial dependencies. Application-wise, perhaps one of the most related applications of an attention mechanism in a medical imaging context, is the Attention U-Net [11], fusing attention modules to a standard U-Net architecture. The integrated module is applied to pancreas segmentation.

III. Attention Gates

Attention in deep neural networks can be defined as a generic alignment score between two input signals [17]. Given two input signals x and y, attention computes the alignment score between x_i and y_i by a compatibility score function $f(x_i, y_i)$ where $i \in 1, .., n$ spatial locations. In general we write:

$$\alpha_i = f(x_i, y_i),$$

$$z_i = g(x_i)\alpha_i, \tag{1}$$

where the output of the compatibility function f is stored as a feature map α, with one attention value per pixel. Signals x and y are feature maps corresponding to intermediate layers and different scale. The attention per-pixel values for α are typically constrained to take values in $[0, 1]$, and multiplied by the transformed input x_i, they serve as local weights to feature map x. After training, these will be pushed towards unity for feature map regions of x that are relevant to the task, and close to zero for irrelevant regions.

The compatibility score function f can be defined in various ways [11], [18], [19], leading to different attention modules. For example, additive attention [11], [18] is defined as:

$$\alpha_i = \sigma(\psi(ReLU(\theta(x_i) + \phi(y_i)))) \tag{2}$$

where ψ, θ, ϕ are embedding functions which can be incorporated using 1×1 convolutions, and σ is the sigmoid function, mapping outputs to the required $[0, 1]$ range.

Used in conjuction with skip connections in a convolutional neural network, the coarse scale input is used to disambiguate irrelevant regions to the finer scale input [11]. This region highlighting and pruning process is relevant to the feed-forward pass as well as during training. In particular, in the context of back-propagation, attention gating results in gradients to be weighted according to their importance to the task at hand.

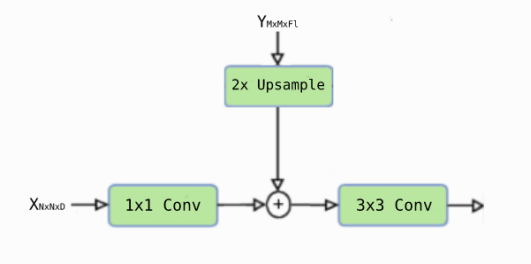

Fig. 1. Building block responsible for constructing the top-down feature maps.

IV. Feature Pyramid Networks and nuclei detection with Mask R-CNN

Feature pyramid networks (FPNs) [7] were designed as a solution for detecting the objects of an image at different scales efficiently by providing multi-scale feature representation of the input image. The main idea of these networks is to take advantage of the "pyramid-like" feature maps produced by a CNN and combine them to high-level semantic feature maps. The construction of the feature pyramid involves a bottom-up and a top-down pathway. The bottom-up pathway is the feed-forward computation of the backbone CNN, which computes a feature hierarchy consisting of feature maps at several scales with a scaling step of two.

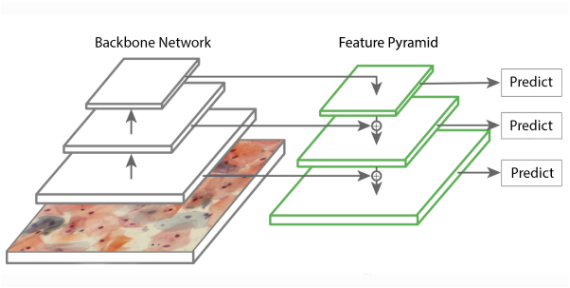

Fig. 2. Feature Pyramid Network Architecture. Bottom-up and Top-down pathway layers are connected via skip connections, and outputs are created at multiple resolutions.

A building block is responsible for constructing the top-down feature maps. The feature map at the coarsest resolution is upsampled by a factor of 2, then merged with the corresponding bottom-up map by element-wise addition. At each pair of corresponding blocks, a bottom-up feature map is semantically low-level, and the top-down map is semantically high-level. This merging is performed by so-called lateral or skip connections.

The Feature Pyramid Network is in the current context used as a backbone network for Mask R-CNN [8], a state-of-the-art instance segmentation / detection network. Mask R-CNN is a convolutional network that incorporates a multi-task loss of the form:

$$L = L_{class} + L_{box} + L_{mask}, \tag{3}$$

where L_{class} is the loss related to correct prediction of each object class, L_{box} depends on correct prediction of object

bounding boxes. L_{mask} depends on binarizing the detected bounding box correctly, so as to label only relevant pixels as the detected mask (and not the whole rectangle). Each of the losses of the multi-task loss L corresponds to a separate output in the network. While this architecture evidently can handle multiple object classes, in the context of nucleus segmentation we simply set the number of object classes to be detected as 1.

V. Proposed model

The proposed model is a based on the vanilla Mask-RCNN network, including however a number of important modifications as well as architectural choices. Following the standard feature pyramid architecture for a ResNet backbone [20], we are using 3 building blocks on top of the last residual block at each stage. We follow the original notation of the output of these last residual blocks, named as $C2, C3, C4, C5$ for $conv2, conv3, conv4$ and $conv5$ outputs. We do not include $conv1$ into the pyramid, due to its large memory footprint. Finally, the FPN outputs are denoted as 3 building blocks $P2, P3, P4$ where the network gives predictions independently at every scale.

Concerning attention, we have used an additive attention gate to define α, as presented in section III (cf. eq. 2). We also replaced the re-sampling module by using zero-padded convolutions. The information flow according to this gate can be examined at fig. 3.

We then combined the aforementioned FPN building blocks [7] and an additive attention gate to build the proposed final block:

$$z = h(\gamma\alpha(x, y)g(x) + g(x) + q(y)), \tag{4}$$

where block output is denoted as z. The alternated attention modules are incorporated into the standard feature pyramid architecture to highlight salient features that are passed through the skip connections. This scheme allows us to build a richer hierarchy that combines both non-local and local information. The proposed attention-based block can be examined at fig. 1.

Attention gates are integrated to all 3 Feature Pyramid levels. Functions θ, ϕ, ψ, g, q are all defined as $1 \times 1 \times N$ convolutions. Function h is also a $3 \times 3 \times N$ convolution, acting over the block output, effectively mitigating aliasing due to upsampling. (This is operation is akin to that of a deconvolutional layer [21], acting as a parameterized linear upsampler).

The significance of the γ residual coefficient in eq. 4 is related to a compromise between weights resulting from a pretrained, non-attention version of the backbone. Using weights obtained from pretraining on a very large dataset (e.g. ImageNet) is standard practice for deep neural network training. The coefficient γ is a scalar initialized as 0. Introducing γ as a learnable parameter, allows the network to first rely on the learned pretrained weights and then gradually and during training to take into account attention gating.

VI. Experiments

A. Setup

For our numerical experiments, we have used the publicly available microscopy imaging dataset BBBC038v1, part of the Broad Bioimage Benchmark Collection [22]. The data consists of 729 microscopy images, where pixel-level annotations of nucleus positions are provided. Dataset nuclei are imaged in a variety of conditions, including fluorescent and histology stains, several magnifications, and varying quality of illumination. A small sample of the dataset can be seen at fig. VI.

We have run experiments on various different versions of the Mask R-CNN model [8], integrated with a Feature Pyramid backbone and the proposed attention mechanism. Resnet-101-FPN is used as the Feature Pyramid backbone for Mask-RCNN. [1]. All the standard building blocks where replaced by the proposed residual attention block (fig. VI-A). This setting has been compared with numerical trials versus the baseline Mask-RCNN with no attention gating, as wells as various different variations of the attention-based model.

At each trial, we have trained the network for 50 epochs using Stochastic Gradient Descent (SGD) with momentum 0.9 and weight decay parameter set to 0.0001. We constrained the number of training RoIs per image to 600, as these images are small and tend to have fewer objects, allowing RoI sampling to pick 33% positive RoIs.

Furthermore, each batch was set include 2 input images, each at resolution of 512×512 pixels. As the dataset contains microscopy images at various different resolutions, we create fixed-size training images by sampling random 512×512 crops from the available data. Data augmentation has also been employed, with cropped inputs undergoing a random set of simple transforms to produce augmented data. In particular, we have used horizontal and vertical flips and random rotations.

In order to evaluate our results, we have computed average precision (AP) values at different intersection over union (IoU) thresholds, following the evaluation protocol of the Kaggle 2018 Data Science bowl competition [3]. In this context, AP is defined as

$$\frac{1}{thresholds} \sum_{t} \frac{TP(t)}{TP(t) + FP(t) + FN(t)} \tag{5}$$

where $TP(t)$, $FP(t)$, $FN(t)$ are numbers of True Positives, False Positives and False Negatives, computed with respect to IoU threshold t. The IoU threshold is set to vary on values from 0.5 up to 0.95, with a step size of 0.05.

B. Discussion of results

Numerical results can be examined in table I. We have tested the standard attention modules as proposed in [11] and the novel residual modules proposed in this paper. The attention-based variants that were compared (aside from the proposed variant presented in section V) are as follows:

[1]The implementation used is based on an the implementation at https://github.com/matterport/Mask_RCNN

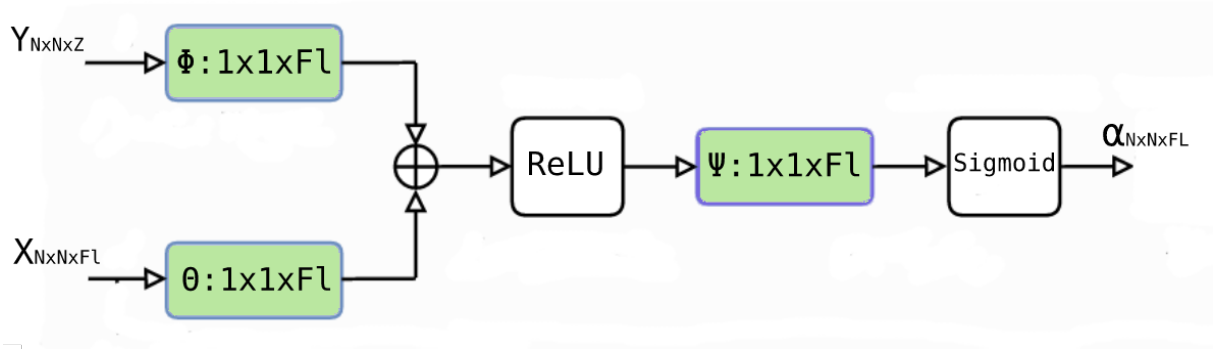

Fig. 3. Proposed attention module.

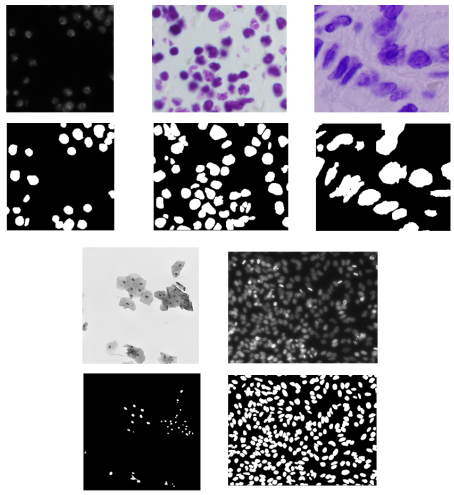

Fig. 4. Indicative samples images from dataset BBBC038v1, and their corresponding ground truth masks showing nucleus positions.

TABLE I
NUMERICAL RESULTS FOR NUCLEI DETECTION TRIALS. AVERAGE PRECISION (AP) FIGURES ARE SHOWN, COMPUTED FOR THE PROPOSED METHOD VERSUS A VANILLA MASK-RCNN MODEL AND OTHER ATTENTION-BASED VARIANTS OF THE PROPOSED METHOD.

Model	AP
No attention gating (vanilla Mask R-CNN)	0.625
Attention Mask - No Gamma	0.636
Attention Mask - Gamma for pixel vector	0.639
Attention Mask - Concat	0.637
Attention Mask - Proposed model	**0.641**

Attention Mask - No Gamma: The γ parameter is not included, hence block outputs are instead of eq. 4 computed as

$$z = h(\gamma\alpha(x, y)g(x) + g(x) + q(y)).$$

Attention Mask - Gamma for Pixel Vector: FPN blocks and attention is defined as in the proposed model, but functions ψ, θ, ϕ are $1 \times 1 \times 1$ (instead of $1 \times 1 \times N$), following [11].

Attention Mask - Concat: Attention under this variant is defined as:

$$\alpha_i = \sigma(\psi(ReLU([\theta(x_i), \phi(y_i)]))) \quad (6)$$

where $[\cdot, \cdot]$ corresponds to an input concatenation operation.

Detection results for the baseline Mask R-CNN detector without an attention module are also included in Table I. As we can see, in all cases accurate nuclei detection is quite a challenging problem. It is clear that a supervised segmentation techniques like a deep convolutional neural networks can benefit from the presence of attention modules since models with attention modules outperform the standard setting.

With respect to the type of model, the model that is integrated with the proposed attention module outperforms other variants. As we can see the effect of the trainable γ is crucial for the model's final result as experiments without this parameter give lower scores. Also, even though the concatenation attention module has more parameters than the other modules, it still does not perform as well as the additive module.

The usefulness and importance of the learned attention coefficients α on detection can be seen clearly in figure VI-A. As we can see, regions of the input image that are highly correlated with ground truth positions of nuclei attain higher attention coefficient values.

VII. CONCLUSION AND FUTURE WORK

We have presented a nuclei detection scheme that utilises a novel attention gating mechanism, integrated with feature pyramids and the state-of-the-art Mask R-CNN network. Numerical experiments show that the addition of the proposed attention module results in improving overall detection efficiency. Furthermore, visualizations of relevant feature map activations show that indeed the proposed attention gating manages to prune areas of the input that are irrelevant to the task. This results in subsequent network intermediate layers being given only parts of the feature map that are relevant, eventually leading to a more efficient detection task.

For future work, we aim at exploring other forms of gating functions, like non-local gating extensions [16]. We also look forward to extending the proposed attention mechanism to detecting $3D$ structures, by defining networks that are applied on 3D structures / volumes [23], [24] or sequences of frames [25] (for example, for cell tracking). Finally, we envisage fusioning the proposed detection mechanism with a probabilistic model approach, as popularized recently with Variational

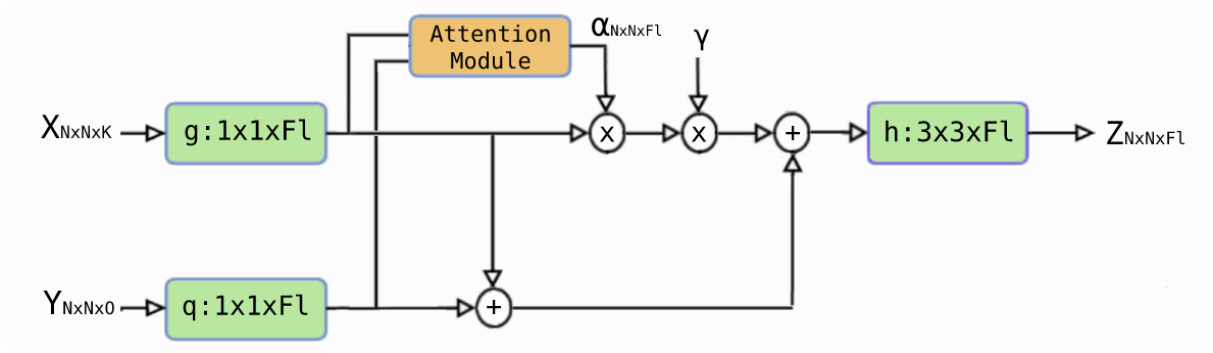

Fig. 5. Proposed feature pyramid residual attention block.

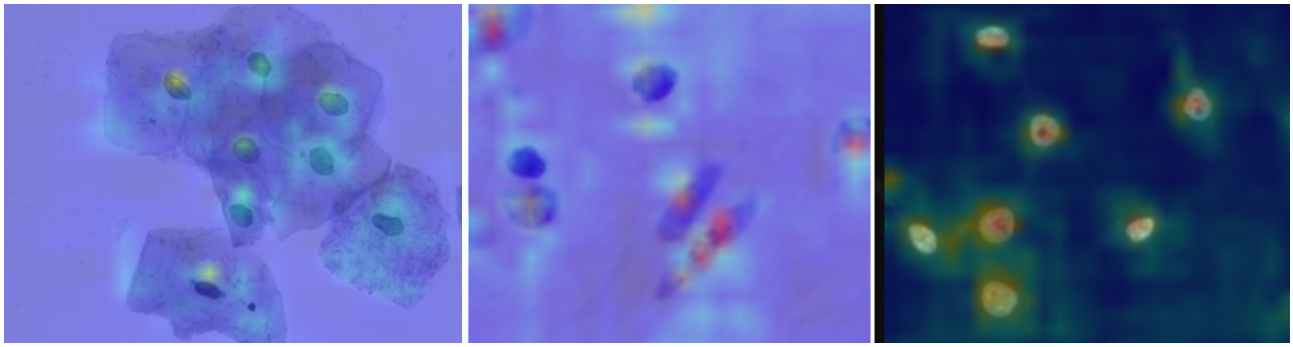

Fig. 6. Visualization of learned attention coefficients. Attention is visualized over indicative nuclei images, pseudocoloured as a heat map. Attention coefficients from the C4/P4 pyramid stage module are shown.

Autoencoder-based models [26] or the more classical Bayesian paradigm [27]–[29].

REFERENCES

[1] Marina E Plissiti, P Dimitrakopoulos, Giorgos Sfikas, Christophoros Nikou, O Krikoni, and Antonia Charchanti, "SIPAKMED: A new dataset for feature and image based classification of normal and pathological cervical cells in Pap smear images," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 3144–3148.

[2] Erick Moen, Dylan Bannon, Takamasa Kudo, William Graf, Markus Covert, and David Van Valen, "Deep learning for cellular image analysis," *Nature methods*, p. 1, 2019.

[3] Johan Scott Loudon, "Detecting and localizing cell nuclei in medical images.," M.S. thesis, NTNU, 2018.

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

[5] Olaf Ronneberger, Philipp Fischer, and Thomas 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.

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

[7] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie, "Feature pyramid networks for object detection," in *CVPR*, 2017, vol. 1, p. 4.

[8] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick, "Mask R-CNN," in *Computer Vision (ICCV), 2017 IEEE International Conference on*. IEEE, 2017, pp. 2980–2988.

[9] Xinpeng Xie, Yuexiang Li, Menglu Zhang, and Linlin Shen, "Robust segmentation of nucleus in histopathology images via mask r-cnn," in *International MICCAI Brainlesion Workshop*. Springer, 2018, pp. 428–436.

[10] Jeremiah W Johnson, "Adapting Mask R-CNN for automatic nucleus segmentation," *arXiv preprint arXiv:1805.00500*, 2018.

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

[12] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang, "Residual attention network for image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3156–3164.

[13] Wei Li, Xiatian Zhu, and Shaogang Gong, "Harmonious attention network for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2285–2294.

[14] Petros-Pavlos Ypsilantis and Giovanni Montana, "Learning what to look in chest X-rays with a recurrent visual attention model," *arXiv preprint arXiv:1701.06452*, 2017.

[15] Xinyuan Chen, Chang Xu, Xiaokang Yang, and Dacheng Tao, "Attention-GAN for object transfiguration in wild images," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 164–180.

[16] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He, "Non-local neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7794–7803.

[17] Tao Shen, Tianyi Zhou, Guodong Long, Jing Jiang, Shirui Pan, and Chengqi Zhang, "Disan: Directional self-attention network for rnn/cnn-free language understanding," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[18] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep

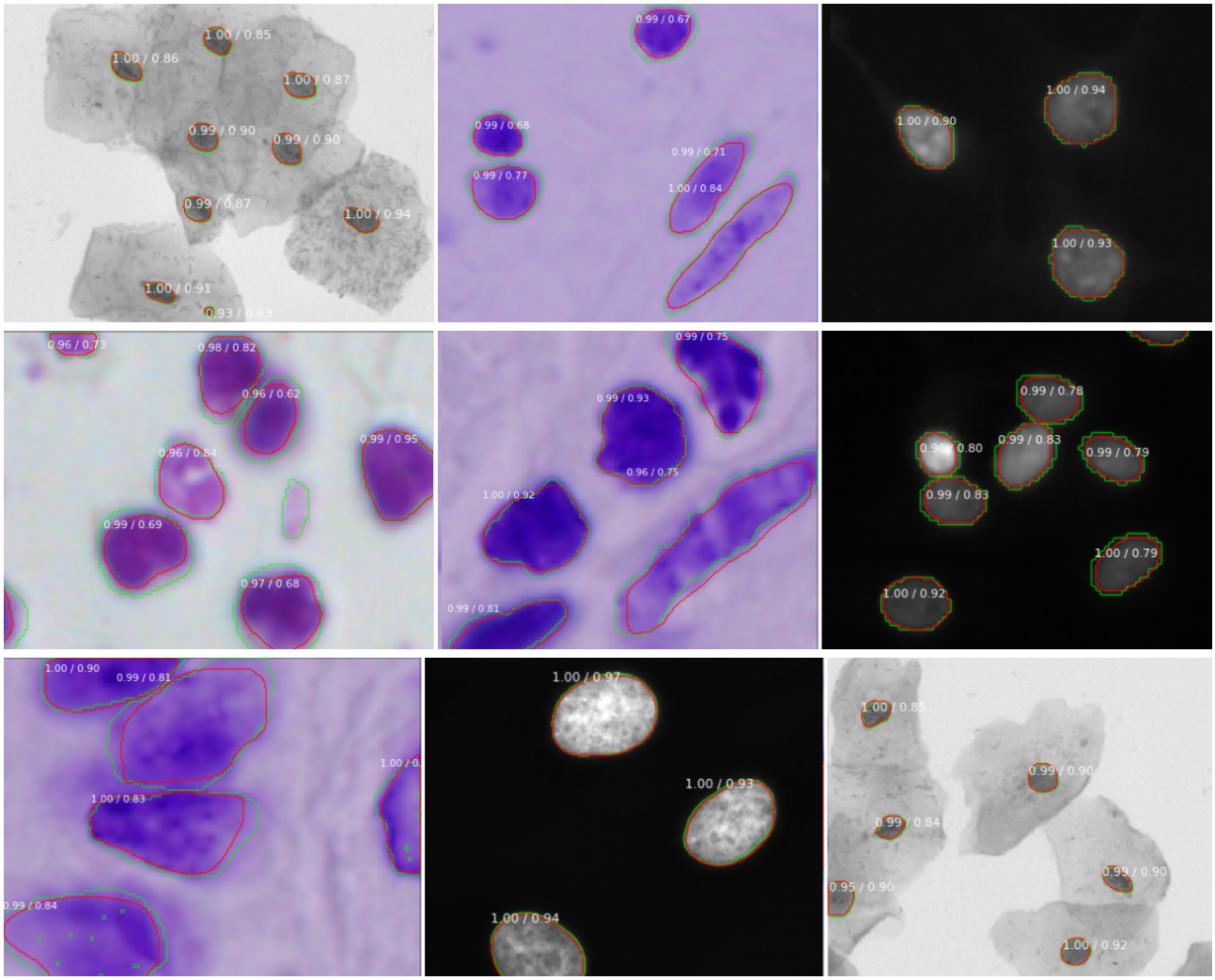

Fig. 7. Qualitative result of the proposed detection scheme. Ground truth and detected object borders are drawn with green and red color respectively. Numerical captions indicate model prediction score and IoU ratio of detection versus groud truth.

residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[21] Augustus Odena, Vincent Dumoulin, and Chris Olah, "Deconvolution and checkerboard artifacts," *Distill*, vol. 1, no. 10, pp. e3, 2016.

[22] Vebjorn Ljosa, Katherine L Sokolnicki, and Anne E Carpenter, "Annotated high-throughput microscopy image sets for validation," *Nat Methods*, vol. 9, no. 7, pp. 637, 2012.

[23] Konstantinos Kamnitsas, Christian Ledig, Virginia FJ Newcombe, Joanna P Simpson, Andrew D Kane, David K Menon, Daniel Rueckert, and Ben Glocker, "Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation," *Medical image analysis*, vol. 36, pp. 61–78, 2017.

[24] Georgia Gkioxari, Jitendra Malik, and Justin Johnson, "Mesh R-CNN," *arXiv preprint arXiv:1906.02739*, 2019.

[25] Panagiotis Kouzouglidis, Giorgos Sfikas, and Christophoros Nikou, "Automatic video colorization using 3D conditional generative adversarial networks," *arXiv preprint arXiv:1905.03023*, 2019.

[26] Suman Sedai, Dwarikanath Mahapatra, Sajini Hewavitharanage, Stefan Maetschke, and Rahil Garnavi, "Semi-supervised segmentation of optic cup in retinal fundus images using variational autoencoder," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2017, pp. 75–82.

[27] Nick Pawlowski, Matthew CH Lee, Martin Rajchl, Steven McDonagh, Enzo Ferrante, Konstantinos Kamnitsas, Sam Cooke, Susan Stevenson, Aneesh Khetani, Tom Newman, et al., "Unsupervised lesion detection in brain ct using bayesian convolutional autoencoders," 2018.

[28] Giorgos Sfikas and Christophoros Nikou, "Bayesian multiview manifold learning applied to hippocampus shape and clinical score data," in *Medical Computer Vision and Bayesian and Graphical Models for Biomedical Imaging*, pp. 160–171. Springer, 2016.

[29] Giorgos Sfikas, Christophoros Nikou, Nikolaos Galatsanos, and Christian Heinrich, "MR brain tissue classification using an edge-preserving spatially variant bayesian mixture model," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2008, pp. 43–50.