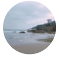
Triple ANet: Adaptive Abnormal-aware Attention Network for WCE Image Classification

Chapter · October 2019

DOI: 10.1007/978-3-030-32239-7_33



Xiaoqing Guo
City University of Hong Kong
24 PUBLICATIONS **273** CITATIONS

Yixuan Yuan
City University of Hong Kong
84 PUBLICATIONS **2,002** CITATIONS

Triple ANet: Adaptive Abnormal-aware Attention Network for WCE Image Classification

Xiaoqing Guo and Yixuan Yuan(✉)

Department of Electrical Engineering,
City University of Hong Kong, Kowloon, Hong Kong, China
xiaoqiguo2-c@my.cityu.edu.hk, yxyuan.ee@cityu.edu.hk

Abstract. Accurate detection of abnormal regions in Wireless Capsule Endoscopy (WCE) images is crucial for early intestine cancer diagnosis and treatment, while it still remains challenging due to the relatively low contrasts and ambiguous boundaries between abnormalities and normal regions. Additionally, the huge intra-class variances, alone with the high degree of visual similarities shared by inter-class abnormalities prevent the network from robust classification. To tackle these dilemmas, we propose an Adaptive Abnormal-aware Attention Network (Triple ANet) with Adaptive Dense Block (ADB) and Abnormal-aware Attention Module (AAM) for automatic WCE image analysis. ADB is designed to assign one attention score for each dense connection in dense blocks and to enhance useful features, while AAM aims to adaptively adjust the respective field according to the abnormal regions and help pay attention to abnormalities. Moreover, we propose a novel Angular Contrastive loss (AC Loss) to reduce the intra-class variances and enlarge the inter-class differences effectively. Our methods achieved 89.41% overall accuracy and showed better performance compared with state-of-the-art WCE image classification methods. The source code is available at https://github.com/Guo-Xiaoqing/Triple-ANet.

1 Introduction

Wireless capsule endoscopy (WCE) is a noninvasive, wireless imaging tool that allows direct visualization of the entire gastrointestinal (GI) tract without discomfort to patients. Despite their prevalent and good application, the collected WCE videos are manually reviewed, which is time-consuming and extremely laborious for clinicians. Moreover, the GI diseases usually demonstrate various characteristics of shape, texture, and size. Even well trained clinicians may produce different diagnostic results. Therefore it is highly desirable to develop a computer-aided detection (CAD) method to automatic diagnose diseases with satisfying accuracy.

Common GI diseases includes inflammatory, vascular lesion and polyp. Effectively recognizing the abnormalities is a very challenging task due to intra-class

D. Shen et al. (Eds.): MICCAI 2019, LNCS 11764, pp. 293–301, 2019.
https://doi.org/10.1007/978-3-030-32239-7_33

variances, inter-class similarities and existence of artifacts. Numbers of scholars have been dedicated to solving these issues and proposing automatic algorithms for abnormality recognition in WCE images [3,5,7,9]. Fan et al. [3] directly utilized AlexNet to automatically recognize ulcer and erosion in WCE images. Jia et al. [5] extracted WCE image features with AlexNet and conducted automatic bleeding detection strategy based on support vector machine. Seguí et al. [7] proposed an early fusion approach, in which the Laplacian and Hessian streams were integrated with original WCE images as input of a VGG-based neural network for abnormality classification. In our previous work [9], we proposed a rotation-invariant and image similarity constrained neural network for polyp recognition by dealing with object rotation problems of the collected WCE images.

Despite the relatively good performance, deep learning based methods [3,5,7,9] could not localize abnormalities accurately and extract sufficiently distinguishable features with only image-level labels. Moreover, the performance of existing multi-class WCE classification task is not satisfactory. The challenges associated with these methods may lie in the following three parts. Firstly, the latest research [9] performed WCE image classification by utilizing the Densely Connected Convolutional Network (DenseNet) [4], in which every layer are concatenated to each other layer to aggregate information and learn features. However, this model treats every feature layer equally and ignores the importance variation of different connections. Second, existing works for WCE image classification extracted features directly from the whole image and assumed that different image parts contribute equally to the feature learning [3,5,7,9]. However, the background and noisy parts in the image may introduce redundant information for the network and lead to bad performance while some abnormal informative parts are the important cues for doctors to make diagnostic decisions. While the recent self-attention module [8] generated attention map by calculating the correlation matrix of the feature map to highlight the important regions in images, it has not been applied in the WCE image analysis filed. Moreover, this attention module only considers the pixel-to-pixel relationship, overlooking the context information. Thirdly, softmax loss was commonly utilized to learn WCE image features [3,7,9]. Thus, the calculated deep features could not support feature correlations across the entire data space. In reality, images within the same class should share similar feature information while the image features from different classes should be distinctly different.

To address the aforementioned challenges, we propose an Adaptively Abnormal Aware Network (Triple ANet) for WCE image classification. Our contributions lie in the following four points. (1) An Adaptive Dense Block (ADB) is developed to adaptively assign one attention score for each dense connected layer in dense blocks, and the score reveals importance of feature maps in different depth. In this way, ADB can selectively aggregate the most useful information of the images. (2) We propose an Abnormal-aware Attention Module (AAM) that can gradually adjust the respective field according to the abnormal regions. This AAM is aimed to combine local information with context and help network pay attention to the abnormal region. (3) A novel angular contrastive loss

(AC Loss) is proposed to reduce the intra-class variances and enlarge the inter-class differences effectively. Therefore, Triple ANet can better characterize and distinguish features of different diseases rendered in WCE images. (4) We validate the robustness of our proposed Triple ANet by conducting comprehensive experiments, and our method achieves the state-of-the-art WCE classification performance compared with existing methods.

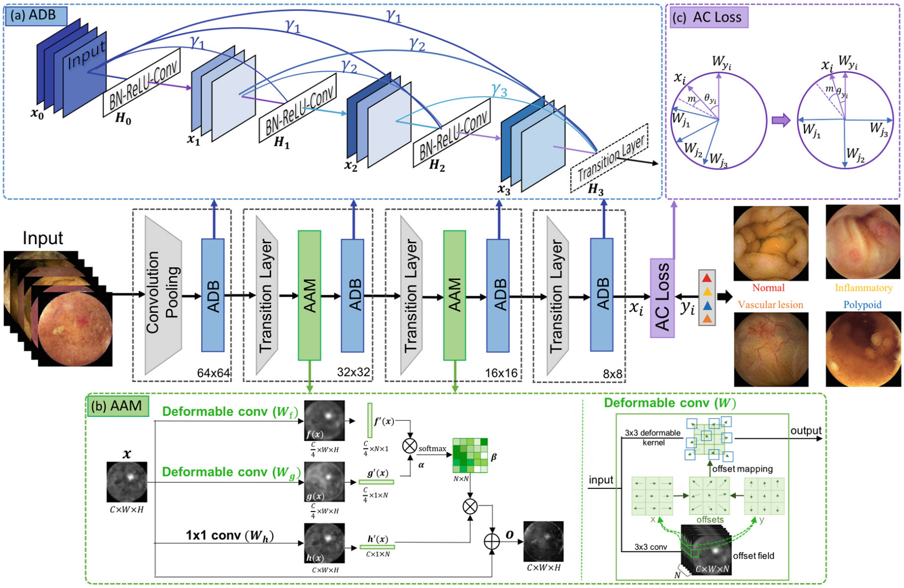


Fig. 1. Illustration of the proposed Triple ANet.

2 Method

In this paper, we propose the Triple ANet model to automatically differentiate inflammatory, vascular lesion and polyp from normal WCE images. The overall framework of our method is illustrated in Fig. 1. Given a WCE image, the deep features are extracted by alternate-cascaded ADBs, AAMs and transition layers. The size of feature maps in the four ADBs are 64×64, 32×32, 16×16, and 8×8, which makes network adaptively aggregate useful information at different scales. Before the first ADB, a 3×3 convolution with 24 output channels is performed on the input images. We use 1×1 convolution followed by 2×2 average pooling as transition layers following ADBs to reduce the number and dimension of feature maps. Two AAMs are inserted at 32×32 and 16×16 scales for capturing long range dependency information. At the end of the last ADB, a global average pooling is performed to squeeze feature maps into vectors for further classification. AC Loss is proposed to optimize the whole network and converge to learn discriminative features.

2.1 Adaptive Dense Block

To effectively train the deep neural network and aggregate information, we propose an ADB module in our classification network to introduce direct and suitable connections from any layer to all subsequent layers. The architecture of an ADB module is displayed in Fig. 1(a). Let x_l denote the output of lth layer and γ_l is the corresponding weight scalar, then adaptive dense connectivity can be formulated as $x_l = H_l([\gamma_0 x_0, \gamma_1 x_1, \cdots, \gamma_{l-1} x_{l-1}])$. Specifically, operator $[\gamma_0 x_0, \gamma_1 x_1, \cdots, \gamma_{l-1} x_{l-1}]$ refers to the adaptive concatenation of the feature maps produced in layers $0, 1, \cdots, l-1$. $H_l(\cdot)$ is a composite function with Batch Normalization (BN), Rectified Linear Units (ReLU) and Convolution (Conv), and each $H_l(\cdot)$ produces $k = 12$ feature maps.

Our proposed ADB module combines information from different layers adaptively, therefore encourages feature reuse, ensures maximum information flow between layers and aggregates useful information effectively. All the weights are initialized as 1s and optimized with iteration, which makes the useful convolutional signals gradually enhanced. Specifically, our proposed framework includes four ADBs as shown in Fig. 1, and they are respectively comprised of 6, 12, 24, 16 densely connected layers.

2.2 Abnormal-aware Attention Module

Considering that features in the neighbourhood of abnormalities also make contribution to the identification, we propose AAM to calculate the region-to-region correlation and highlight features in the most important region. As shown in Fig. 1(b), AAM includes three branches. In particular, the first and second branches are implemented by deformable convolution, while the third branch is implemented by 1×1 convolution. The deformable convolution adds an additional convolutional layer to learn offsets from preceding feature maps, and the offsets include offsets in horizontal and vertical direction. The weight in this additional layer is initialized to $N(0, \sigma^2)$ with $\sigma \ll 1$, and the bias is initialized to zeros. Through offset mapping, deformable convolution adds the learned offsets to the regular grid sampling locations in the standard convolution kernel, which enables the receptive fields to gradually expand around the abnormalities. Through these branches, the input features $x \in \mathbb{R}^{C \times W \times H}$ are gathered into three feature spaces f, g and h. The channels of feature maps $f(x)$ and $g(x)$ are reduced to $\frac{C}{4}$, while channels of $h(x)$ remains to be C. Then the gathering feature maps are arranged into sequences $f^{'}(x), g^{'}(x) \in \mathbb{R}^{\frac{C}{4} \times N}$ and $h^{'}(x) \in \mathbb{R}^{C \times N}$, and the response at a position in a sequence can be represented by a cross correlation matrix with $\alpha_{i,j} = f^{'}(x)^{\top} \times g^{'}(x)$. After that, the obtained cross correlation matrix is spatially normalized to be $\beta_{j,i} \in \mathbb{R}^{N \times N}$, representing the correlativity between regions in feature map respect to the other regions. Note that $\sum_i \beta_{i,j} = 1$. Then the spatial attention maps are computed by $\sum_{i=1}^{N} h(x)\beta_{j,i}$, which indicates the region-to-region correlation in a spatial feature map. The final attention maps is obtained by

$$O = x + \sum_{i=1}^{N} h(x)\beta_{j,i}. \tag{1}$$

With this formula, local information is combined with context information, and we could highlight the abnormality itself as well as the features around the boundaries of abnormalities. Compared with self-attention [8], our AMM could adaptively strengthen the response of abnormalities and suppress noise in normal regions with the expanded receptive fields. Additionally, the proposed AAM could be flexibly inserted to any deep neural networks for capturing long-range dependencies.

2.3 Angular Contrastive Loss

The softmax cross entropy loss is widely used to evaluate the classification loss and it can be formulated as $L_{softmax} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{e^{W_{y_i}^{\top} x_i + b_{y_i}}}{\sum_{j=1}^{n} e^{W_j^{\top} x_i + b_j}}$, where x_i denotes the extracted features of the i^{th} samples, and y_i is the ground truth. W_j is the weight for j^{th} class in the fully connection (FC) layer, and b_j is the bias. N and n represents batch size and number of classes respectively. However, softmax loss function does not explicitly optimize the feature to ensure higher differences for inter-class features and similarity for intra-class features.

To address the issue mentioned above, we propose AC Loss. For simplicity, we discard bias term as in [1] and rewrite the formula of FC layer as $F(W_j, x_i) = ||W_j|| \cdot ||x_i|| \cos\theta_j$, where θ_j is the angle between the features x_i and the weight W_j. Then x_i and W_j are respectively normalized to $||x_i|| = 1$ and $||W_j|| = 1$ by L_2 normalization. A hyper parameter s is introduced to rescale the length of x_i to s, which could control the magnitude of loss value. This normalization method makes the predictions only relied on angle between the features and weights. Our proposed AC Loss takes the general form of $F(W_j, x_i) = s||W_j|| \cdot ||x_i|| \cdot A(\theta_j)$, where $A(\cdot)$ is a angular activation function. Obviously, appropriate angular activation function may lead better classification performance. In this paper, we define the angular activation function as:

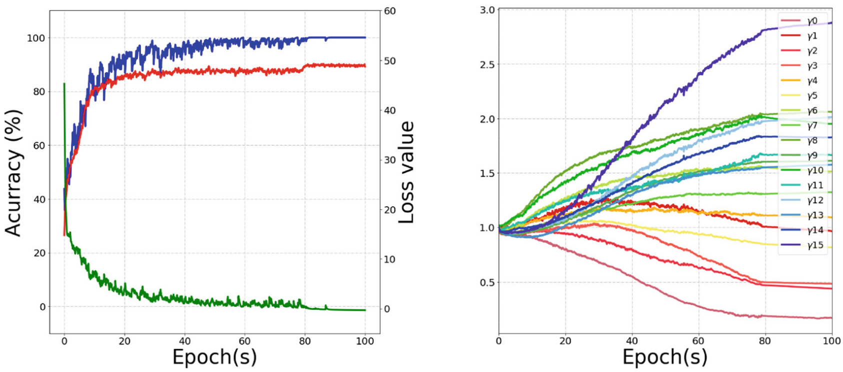
$$A(\theta_j) = \frac{1 + e^{(-\frac{\pi}{2k})}}{1 - e^{(-\frac{\pi}{2k})}} \cdot \frac{1 - e^{(\frac{\theta_j}{k} - \frac{\pi}{2k})}}{1 + e^{(\frac{\theta_j}{k} - \frac{\pi}{2k})}}, \tag{2}$$

where k is a hyper parameter to control the gradient of angular activation function and the performance of loss function directly. This characteristic makes our angular activation function more general compared with original cosine function [1]. In this paper, we choose $k = 0.3$. With $k = 0.3$, the loss function is more smooth when the loss value is large, which reduces the contribution of outliers and increases the robust of training. In the meanwhile, the loss function is more steep when the loss value is not large, to accelerate convergence of the network.

In order to further reduce the intra-class variances and inter-class similarities, we introduce an angle margin penalty m and a regularization term to enhance the discrepancy of weights in FC layer for different classes. Therefore, the proposed AC loss function for Triple ANet training can be formulated as

$$L_{AC} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{sA(\theta_{y_i}+m)}}{e^{sA(\theta_{y_i}+m)}+\sum_{j\neq y_i}^{n}e^{sA(\theta_j)}}+\frac{1}{n}\cdot\frac{1}{n-1}\sum_{y_i=1}^{n}\sum_{j\neq y_i}^{n}W_{y_i}^{\top}W_j. \quad (3)$$

Within a certain range, larger m leads to more similar and compact intra-class features. As shown is Fig. 1(c), converged L_{AC} maximizes the separability of inter-class features and enables intra-class features to cluster toward the weight of their corresponding class.



(a) (b)

Fig. 2. (a) Accuracy and loss values for different epochs. The blue and red color curves respectively represent the train and test accuracy, while the green color one shows the training loss. (b) γ values in the fourth ADB for different epochs. (Color figure online)

3 Experiments and Results

We evaluated Tripe ANet model on a combined WCE dataset from CAD-CAP [2], KID [6] and our collected polyp dataset. It consists of 2846 WCE images, including 771 normal frames, 728 inflammatory ones, 762 vascular lesion ones and 585 polyps. Considering different image resolution and quality, we resized the WCE images to 128×128 and applied a uniform circle mask to the dataset. Flip and rotation were implemented to augment the training data.

We implemented our model using TensorFlow. NVIDIA TITAN XP GPU and CUDA 8.0 are used for the training acceleration. Adam is chosen for optimization with $\beta_1 = 0.5$ and $\beta_2 = 0.999$. All training steps used batch size of 16. The initialized learning rate is set to 0.001, and is dropped by 0.1 at 80 epochs. Hyper parameters of s, m, k are set as 64, 0.5, 0.3, respectively. Fourfold cross validation was adopted to evaluate our methods. The performance of classification was evaluated by per-class accuracy, overall accuracy (OA) and Cohen's Kappa score.

We first analyzed the learning process of Triple ANet. Figure 2(a) shows the train accuracy, test accuracy and train loss value. Our method rapidly reduces the loss values and reaches a relatively steady state after 80 epochs, which indicates that the network is successfully optimized and verifies the effectiveness of Triple ANet for WCE image classification.

Then we analyzed the influence of ADB module. Figure 2(b) shows the change of γ values in the 4^{th} ADB with 16 connected layers, and γ_i indicates the learned attention score for the i^{th} layer. The assigned attention values for different layers vary significantly. It can be seen that the deeper the layer is, the higher the value of γ is, indicating the deep feature will make more contributions for the WCE image classification. Actually, the deeper features usually contain more context and spatial structure information, while the shallow features contain more color and texture information. Thus we can make a conclusion that structure information plays more important role in classification.

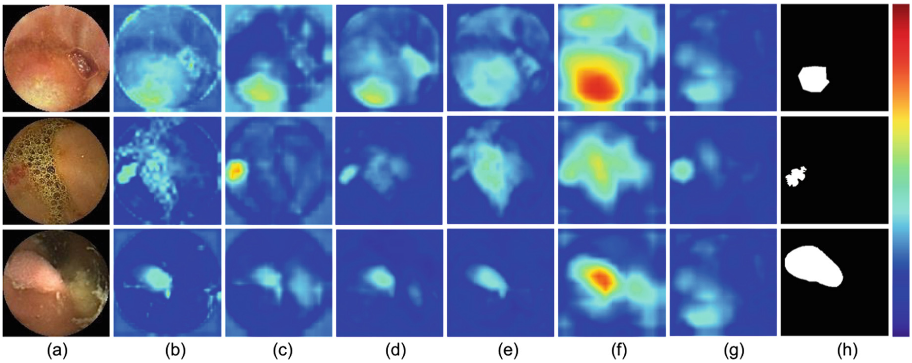


Fig. 3. From top to bottom, they are respectively inflammatory, vascular lesion and polyp samples. (a) Original image. (b)(c) show attention maps of the 1st and 2nd AAM. (d)(e) show offset fields in the 1st branch and 2nd branch of the 1st AMM while (f)(g) show offset fields in the 1st branch and 2nd branch of the 2st AMM. (h) Ground truth of mask

We further showed extracted attention maps and learned offset fields in Fig. 3. Figure 3(b–c) show the feature maps extracted from the two AAMs, and we could find that AAMs make feature maps highlight the abnormal regions successfully. Figure 3(d–g) show the offset fields obtained in these AAMs, and red regions indicate the large receptive fields. It can be seen that the respective fields are tend to be larger at abnormalities compared with normal regions. Therefore, context information of abnormalities are incorporated with local information to lead better classification performance.

Table 1. Comparison results for WCE image classification. w/ADB, w/AAM and w/AC Loss denote DenseNet with ADB (instead of original dense block), DenseNet with AAM and DenseNet minimized by AC Loss (instead of original softmax cross entropy loss), respectively.

Methods	Normal ACC (%)	Inflammatory ACC (%)	Vascular lesion ACC (%)	Polyp ACC (%)	OA	Cohen's Kappa
DenseNet [4]	92.69 ± 0.49	90.12 ± 0.54	93.71 ± 0.52	97.59 ± 0.39	87.05 ± 0.45	82.66 ± 0.60
w/ADB	93.03 ± 0.40	89.94 ± 0.18	93.61 ± 0.15	97.71 ± 0.27	87.14 ± 0.21	82.78 ± 0.28
w/AAM	**94.06 ± 0.17**	91.49 ± 0.81	94.47 ± 0.77	97.78 ± 0.42	88.89 ± 0.52	85.13 ± 0.69
w/AC Loss	93.59 ± 0.30	91.20 ± 0.33	94.94 ± 0.34	97.69 ± 0.46	88.70 ± 0.20	84.87 ± 0.27
Triple ANet	94.03 ± 0.09	**91.73 ± 0.29**	**95.26 ± 0.33**	**97.81 ± 0.20**	**89.41 ± 0.23**	**85.82 ± 0.31**
Fan et al. [3]	85.44 ± 1.43	83.09 ± 0.79	90.19 ± 0.96	95.47 ± 0.89	77.10 ± 1.14	69.30 ± 1.58
Jia et al. [5]	86.16 ± 1.07	83.37 ± 0.71	90.32 ± 0.88	95.81 ± 0.59	77.83 ± 1.28	70.31 ± 1.74
Seguí et al. [7]	92.11 ± 0.60	89.71 ± 0.48	94.21 ± 0.57	97.31 ± 0.12	86.67 ± 0.84	82.15 ± 1.12
Yuan et al. [9]	93.44 ± 0.30	90.79 ± 0.26	93.91 ± 0.17	97.73 ± 0.35	87.93 ± 0.07	83.84 ± 0.08

To individually demonstrate the effectiveness of the proposed ADB, AAM and AC Loss, we conducted several comparison experiments and the results were shown in Table 1 *row* 1–4. In general, it is clear that the proposed ADB, AAM and AC Loss make contribution to the promotion of performance, because involving any of them leads to relatively better performance compared with traditional DenseNet [4]. Especially, AAM and AC Loss show significant improvements. The AAM module was verified to be effective in improving the classification performance, with 1.84%, 2.47% increment in OA and Cohen's Kappa. This increment is due to that AAM captured the long range dependent information and amplified the effects of abnormal regions. We also replaced original softmax cross entropy loss with AC Loss in DenseNet to evaluate the performance of AC Loss, which is denoted as 'w/AC Loss' in Table 1. The great improvement of 1.65%, 2.21% in OA and Cohen's Kappa compared with DenseNet indicates that AC Loss can facilitate the distinction of learned features and strengthen the robustness of Triple ANet.

We further assessed the performance of the our Triple ANet (*row* 5) by comparing it with state-of-the-art methods [3,5,7,9] for WCE image classification. We implemented these methods on our datasets and the comparison results are shown in Table 1 *row* 6–9. The proposed method shows superior performance with an increment of 12.31%, 11.58%, 2.74%, 1.48% in OA, 16.52%, 15.51%, 3.67%, 1.98% in Cohen's Kappa compared with methods [3,5,7,9], respectively. This result validates the proposed Triple ANet possesses superior ability to aggregate abnormal information and extract discriminative features for WCE images.

4 Conclusion

Automatic abnormality classification is a challenge task due to the diverse characteristics rendered on WCE images. Our method is fundamentally different from the previous works with traditional convolutional network applications. Instead, we proposed a novel Triple ANet with Adaptive Dense Block (ADB), Abnormal-aware Attention Module (AAM) and Angular Contrastive loss (AC Loss). Our methods can be flexibly transferred to a wide range of medical image classification tasks to extract discriminative features and boost the classification performance.

Acknowledgments. This work was supported by Sichuan Provincial Science and Technology Research Grant 2019YJ0632 and TSG 6000690. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU for this research.

References

1. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: ArcFace: additive angular margin loss for deep face recognition. In: CVPR, pp. 4690–4699 (2019)
2. Dray, X., et al.: Cad-cap: une base de données française à vocation internationale, pour le développement et la validation d'outils de diagnostic assisté par ordinateur en vidéocapsule endoscopique du grêle. Endoscopy **50**(03), 000441 (2018)
3. Fan, S., Xu, L., Fan, Y., Wei, K., Li, L.: Computer-aided detection of small intestinal ulcer and erosion in wireless capsule endoscopy images. Phys. Med. Biol. **63**(16), 165001 (2018)
4. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: CVPR, pp. 4700–4708 (2017)
5. Jia, X., Meng, M.Q.H.: A deep convolutional neural network for bleeding detection in wireless capsule endoscopy images. In: EMBC, pp. 639–642 (2016)
6. Koulaouzidis, A., et al.: Kid project: an internet-based digital video atlas of capsule endoscopy for research purposes. Endosc. Int. Open **5**(06), E477–E483 (2017)
7. Seguí, S., et al.: Generic feature learning for wireless capsule endoscopy analysis. Comput. Biol. Med. **79**, 163–172 (2016)
8. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: CVPR, pp. 7794–7803 (2018)
9. Yuan, Y., Qin, W., Ibragimov, B., Han, B., Xing, L.: RIIS-DenseNet: rotation-invariant and image similarity constrained densely connected convolutional network for polyp detection. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11071, pp. 620–628. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00934-2_69