2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 25-29, 2020, Las Vegas, NV, USA (Virtual)

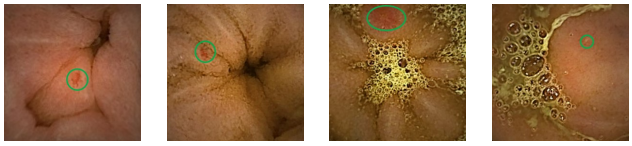
Diagnose like a Clinician: Third-order Attention Guided Lesion Amplification Network for WCE Image Classification

Xiaohan Xing[1], Yixuan Yuan[2*], and Max Q.-H. Meng[1,3*], *Fellow, IEEE*

***Abstract*— Wireless capsule endoscopy (WCE) is a novel imaging tool that allows the noninvasive visualization of the entire gastrointestinal (GI) tract without causing discomfort to the patients. Although convolutional neural networks (CNNs) have obtained promising performance for the automatic lesion recognition, the results of the current approaches are still limited due to the small lesions and the background interference in the WCE images. To overcome these limits, we propose a Third-order Attention guided Lesion Amplification Network (TALA-Net) for WCE image classification. The TALA-Net consists of two branches, including a global branch and an attention-aware branch. Specifically, taking the high-level features in the global branch as the input, we propose a Third-order Attention (ToA) module to generate attention maps that can indicate potential lesion regions. Then, an Attention Guided Lesion Amplification (AGLA) module is proposed to deform multiple level features in the global branch, so as to zoom in the potential lesion features. The deformed features are fused into the attention-aware branch to achieve finer-scale lesion recognition. Finally, predictions from the global and attention-aware branches are averaged to obtain the classification results. Extensive experiments show that the proposed TALA-Net outperforms state-of-the-art methods with an overall classification accuracy of 94.72% on the WCE dataset.**

I. INTRODUCTION

Wireless capsule endoscopy (WCE) [1] has been widely adopted for early screening of gastrointestinal (GI) diseases. Compared with traditional endoscopies, WCE provides painless and noninvasive visualization of the entire GI tract. In the screening for each patient, WCE will generate a large number of images, usually more than 55,000, which is time-consuming and tedious for clinicians to manually review all these images. Additionally, various shapes, textures, and sizes make it quite challenging for the clinicians to correctly identify lesion regions. Even well-trained clinicians may produce different diagnostic results. Therefore, automated recognition algorithms are highly demanded for efficient and accurate diagnosis of WCE images.



(a) (b) (c) (d)

Fig. 1: Examples of WCE images. (a)-(b) contain vascular lesions. (c)-(d) show inflammatory images. The lesion areas are annotated by the green circles.

Vascular lesion and inflammatory are two common GI diseases, which are also important syndromes or indicators of other GI abnormalities such as bleeding, ulcers and Crohn's diseases. In recent years, many efforts [2]–[4] have been dedicated to developing deep learning-based algorithms for autonomously recognizing these two diseases. However, these methods usually utilize off-the-shelf deep models without taking into account the challenging characteristics of WCE images, thus leading to limited performance and generalization capability. Challenges of WCE lesion recognition mainly lie in two aspects. Firstly, as shown in Fig. 1, lesions usually take up tiny regions in WCE images and show obscure boundaries with the background normal textures, which make them hardly recognizable. Secondly, for images of different categories, the background regions show a highly similar appearance and impede the extraction of class distinctive features. In the classification of abnormal frames, feature embeddings extracted from the entire images might be dominated by the background interference, thus leading to unsatisfactory performance.

In the clinical practice, clinicians usually first browse the whole image to localize potential lesion regions, then zoom in these tiny lesion regions for more detailed inspection, and make final diagnostic decisions based on the global and amplified lesion information. Inspired by this working mechanism, we propose a two-branch Third-order Attention guided Lesion Amplification Network (TALA-Net) to achieve more accurate WCE classification by automatically highlighting the potential lesion regions and emphasizing the features from these regions. Specifically, the global branch takes the WCE images as input and produces attention maps based on a novel Third-order Attention (ToA) module. With the guidance of these attention maps, multiple level features in the global branch are deformed through the proposed Attention Guided Lesion Amplification (AGLA) module. The AGLA module imitates clinicians to zoom in

This project is partially supported by National Key R&D program of China with Grant No. 2019YFB1312400, Shenzhen Science and Technology Innovation Projects JCYJ20170413161503220, and Hong Kong RGC CRF grant C4063-18GF awarded to Prof. Max Q.-H. Meng.

Xiaohan Xing is with the Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong, China (e-mail: xhxing@ee.cuhk.edu.hk).

Yixuan Yuan is with the Department of Electrical Engineering, City University of Hong Kong, Hong Kong, China (e-mail: yxyuan.ee@cityu.edu.hk).

Max Q.-H. Meng is with the Department of Electronic and Electrical Engineering of the Southern University of Science and Technology in Shenzhen, China, on leave from the Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong, and also with the Shenzhen Research Institute of the Chinese University of Hong Kong in Shenzhen, China (e-mail: max.meng@ieee.org).

Yixuan Yuan[2*] and Max Q.-H. Meng[1,3*] are co-corresponding authors of this work.

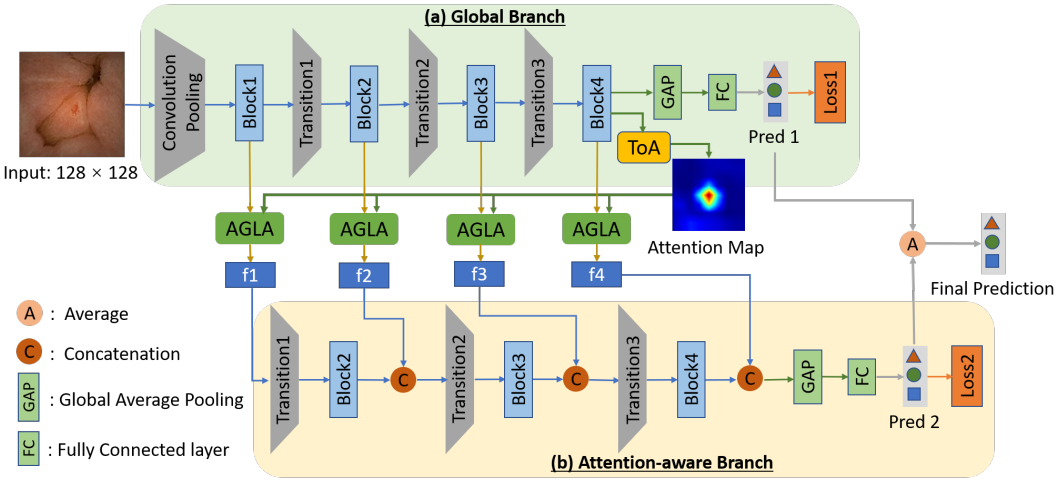


Fig. 2: Overview of the Third-order Attention guided Lesion Amplification Network (TALA-Net). ToA and AGLA stand for the proposed Third-order Attention module and Attention Guided Lesion Amplification module. [Best viewed in color.]

the lesion regions and zoom out the irrelevant background areas, thus producing deformed feature maps with enhanced representation of lesions. By fusing the deformed features into the attention-aware branch, more discriminative feature representations are extracted for finer-scale lesion recognition. Finally, the prediction scores of the two branches are averaged to obtain final classification results.

Our main contributions are summarized as follows:

- We design a two-branch TALA-Net to emphasize the lesion features and promote the WCE lesion recognition in an end-to-end training manner.
- We propose a ToA module that can produce attention maps to highlight the potential lesion regions and guide the lesion amplification.
- A novel AGLA module is proposed to zoom in the suspected lesion features and produce more discriminative feature representations.
- Effectiveness of our proposed ToA and AGLA modules are validated on a WCE dataset. Extensive experiments show that the proposed TALA-Net outperforms state-of-the-art WCE classification methods.

The rest of the paper is organized as follows: Section II reviews the related work, Section III presents the proposed methods. The experimental results are discussed in Section IV and we draw some conclusions in Section V.

II. Related Work

Deep Learning for WCE Image Classification: Due to the strong feature representation and discrimination abilities, deep learning-based algorithms have been widely utilized in WCE image classification [2]–[11]. Some researchers [2], [3] employed AlexNet [12] to automatically recognize abnormal images. Alaskar et al. [4] utilized a combination of the AlexNet and GoogLeNet [13] to distinguish ulcer images. Jeon et al. [6] trained two parallel GoogLeNet models to extract features from the images in RGB and CIELab color spaces, respectively. However, the above works made decisions based on the global images, thus the extracted features were dominated by the background interference and led to limited performance. To mitigate the background interference, Xing et al. [7] utilized saliency maps to indicate the abnormal regions and constructed a saliency-aware input for the WCE classification. Guo et al. [8] proposed a trainable abnormal-aware attention module to enhance the recognition of abnormalities. The performance of [7], [8] were improved since they enhanced the lesion features and suppressed the background interference.

Attention Mechanism: Inspired by the human visual system, attention mechanism has been widely used in the classification tasks of natural images [14]–[17] and medical images [18]. The related papers are reviewed from the perspectives of attention generation and attention utilization.

Since the activation value of a neuron is roughly proportional to its importance, some researchers [14]–[17] proposed activation based methods to construct attention maps. The methods in [14], [15] utilized stacked convolution operations in the CNN model to produce attention maps but required additional trainable parameters. To reduce parameter overheads, some methods generated spatial attention maps using the first-order statistics, such as channel-wise average pooling [16] or max pooling [17] of the feature maps. In our method, the ToA attention calculated by third-order statistics is free of parameters and outperforms the first-order attentions.

In the previous work, attention maps were usually utilized to recalibrate the feature intensities [14], [15], [18]. By such, the lesion features are enhanced to promote the classification performance. Though sharing the similar motivation of emphasizing the lesion features, our proposed AGLA module is intrinsically different from the existing methods since it is the first work that proposes to imitate the working mechanism of human doctors by zooming in the potential lesion regions.

III. Method

The proposed two-branch TALA-Net is illustrated in Fig. 2. For a given WCE image, it is resized to 128×128 and fed into the global branch, which is constructed by a

densely connected convolutional network (DenseNet) [19]. The DenseNet consists of four blocks, including $4, 8, 12, 8$ convolution layers, respectively. Then, the ToA module takes the feature maps in the block4 as an input and generates the attention map with large values at the discriminative lesion regions. The generated attention map is utilized to guide the deformation of different level features through the AGLA module. Compared with the original features in the global branch, lesion features transformed through the AGLA are spatially amplified and better represented. The transformed features are then hierarchically fused to the attention-aware branch to achieve more accurate classification. The entire network is differentiable and is trained by the cross-entropy losses (Loss1, Loss2) in an end-to-end manner.

A. Third-order Attention (ToA)

In the WCE classification networks, attention maps can roughly highlight potential lesion regions and emphasize the lesion features. However, the existing first-order attentions [16], [17] that rely on the local features inside the limited receptive fields usually produce inaccurate attention maps due to the following two reasons. Firstly, since the lesions on one WCE image may distribute in several non-contiguous regions, features from the unobvious segments of lesions may be occluded by the surrounding normal features, thus resulting in small attention values (false-negative). Secondly, hard mimics may share a similar appearance with lesions and produce relatively large attention values (false-positive).

To reduce the false-negative and false-positive responses on the attention maps, we propose a novel ToA module to enhance the attention values of abnormal regions and suppress the responses of hard mimics by aggregating long-range dependent features. Specifically, as shown in Fig. 3, for the feature $F \in R^{W\times H\times C}$ with C channels and size $W \times H$, we reshape it into $X \in R^{HW\times C}$ and calculate the second-order [20] spatial correlation matrix $M \in R^{HW\times HW}$ as

$$M = X\bar{I}X^T, \tag{1}$$

where $\bar{I} = \frac{1}{C}(I - \frac{1}{C}\mathbf{1}\mathbf{1}^T)$, I is the $C \times C$ identity matrix, and the vector $\mathbf{1} = [1, 1, ..., 1]^T$. The matrix M can capture long-range dependencies, with each entry $M_{i,j}$ represents the feature correlation between the i-th and j-th positions. Regardless of the spatial distance, a pair of features from the same class has a higher correlation while lesion features show lower correlations with normal features.

Then, the long-range dependent features are aggregated through the third-order feature aggregation defined as

$$X' = softmax(M)X = softmax(X\bar{I}X^T)X, \tag{2}$$

where the second-order correlation matrix M is first normalized through the row-wise softmax, hence the values in each row sum up as 1. Then the normalized correlation matrix is multiplied by the input feature X to produce the third-order feature $X' \in R^{HW\times C}$, with each entry $X'_{i,c}$ calculated as

$$X'_{i,c} = \sum_{j=1}^{HW} M_{i,j}X_{j,c}. \tag{3}$$

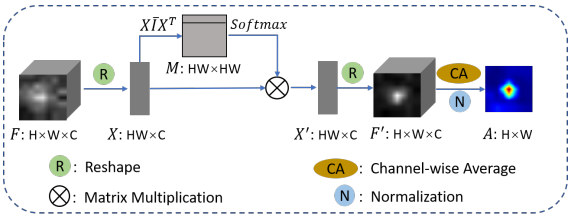


Fig. 3: Illustration of the Third-order Attention (ToA) module. F denotes the feature of the block4 in the global branch, A is the ToA attention map. [Best viewed in color.]

With this formula, the feature response at each position $X'_{i,c}$ is modified by the weighted aggregation of the features at all positions, which results in two benefits. On one hand, the responses at the potential false-negative positions are enhanced by aggregating the features from other lesions that might be spatially distant but share the similar semantic features with them. On the other hand, the features of hard mimics are suppressed by aggregating normal features with larger weights, thus reducing the false-positive responses. As illustrated in Fig. 3, compared with the original feature F, the third-order feature F' has enhanced responses at the lesion regions and suppressed activations at the normal areas. Finally, the ToA attention map $A \in R^{H\times W}$ is generated by compressing F' through channel-wise average pooling and then normalized into the range of [0, 1]. Attention values of the suspected lesion areas are close to 1 while the responses at the normal regions are close to 0. Although supervised with image-level labels only, the ToA attention map can indicate lesion regions and provide visual explanations for the classification results, which are crucial for the clinical applications of the deep learning-based lesion recognition algorithms.

Compared with the common nonlocal [21] attention module, the proposed ToA has two advantages. First, the ToA module does not require additional trainable parameters while several convolution layers are included in the nonlocal block. Second, the pairwise correlation matrix in the nonlocal module is calculated by the matrix multiplication between two different transformations of the input feature. While in our ToA module, the correlation matrix is computed as the second-order covariance of the input feature, which can capture more accurate pairwise feature correlations.

B. Attention Guided Lesion Amplification (AGLA)

To achieve the enhanced feature representation and finer-scale inspection of the small lesions on WCE images, we propose a novel AGLA module to zoom in the lesion regions indicated by the ToA attention maps.

Although a similar idea of zooming in the discriminative image parts has been studied in natural image analysis [22], [23], the deformation of input images may produce spatial distortion and involve additional interference, thus degrading the robustness and reliability of the network. Compared with natural image analysis, network reliability in the medical domain is more crucial since it is directly related to the diagnosis and survival of the patients. Therefore, we propose

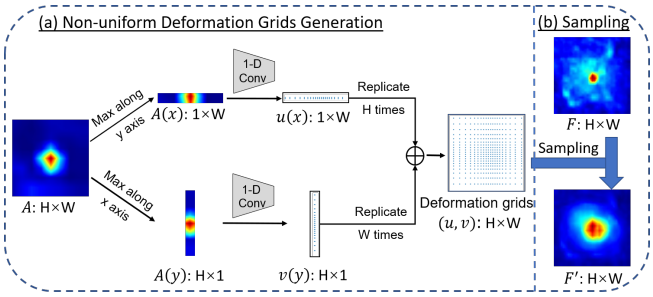


Fig. 4: Illustration of the Attention Guided Lesion Amplification (AGLA). [Best viewed in color with zoom in.]

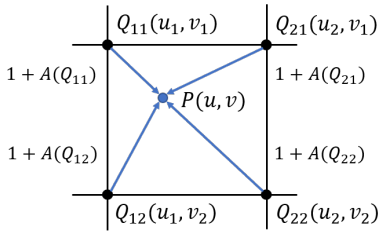


Fig. 5: Illustration of the attention based bilinear interpolation. P is the sampling point with fractional coordinates, and $Q_{11}, Q_{12}, Q_{21}, Q_{22}$ are the four neighbors of P. [Best viewed in color.]

the AGLA module that makes two main modifications to transfer the idea to the WCE domain. First, the AGLA module deforms the hierarchical feature maps rather than the input images. Compared with the WCE images, the abstracted feature maps contain less low-level features (e.g., edges and boundaries) and are less likely to suffer from the spatial distortion, thus the AGLA module can amplify the lesion features without distorting the input WCE images. Second, the AGLA presents a novel deformation method to reduce the spatial distortion effects.

As depicted in Fig. 4, the AGLA module consists of two parts: (a) generating non-uniform deformation grids (u, v) that are densely distributed in high-attention regions; (b) sampling the positions (u, v) from the original feature maps to generate the deformed features. To produce non-uniform deformation grids that are less likely to cause spatial distortion, we propose to calculate the row-wise mapping $(x \mapsto u)$ and column-wise mapping $(y \mapsto v)$ independently. In this way, each entire row or entire column is either sampled or discarded, thus the deformed features can keep the spatial structure of the original features. Specifically, we first decompose the attention map into marginal attention distributions $A(x)$ and $A(y)$ through

$$A(x) = \max_{1 \le y \le H} A(x, y); \quad A(y) = \max_{1 \le x \le W} A(x, y). \tag{4}$$

In order to produce u and v that are proportional to the attention values, we formulate the mapping problem as to find $u(x)$ and $v(y)$ that satisfy

$$\int_0^{u(x)} A(u)du = x; \qquad \int_0^{v(y)} A(v)dv = y. \tag{5}$$

Let's take the mapping from x to $u(x)$ for example, in an area with a higher attention $A(u)$, the increase of x corresponds to a smaller increment of $u(x)$, thus leading to more densely distributed u, and vice versa. Then, the solution $u(x)$ and $v(y)$ of Eq. (5) are calculated as

$$\begin{cases} u(x) = \frac{\sum_{x'} A(x')k(x,x')x'}{\sum_{x'} A(x')k(x,x')} \\ v(y) = \frac{\sum_{y'} A(y')k(y,y')y'}{\sum_{y'} A(y')k(y,y')}, \end{cases} \tag{6}$$

which is implemented by the 1D convolutions with Gaussian kernels. The underlying idea of Eq. (6) is that each 1-D pixel x' or y' pulls its neighbors with a force of $A(x')$ or $A(y')$. As a result, the 1-D deformation grids $u(x)$ and $v(y)$ are decided by the integration of the pulling effects from all positions. As shown in Fig. 4 (a), the distribution of $u(x)$ and $v(y)$ are proportional to the attention values. Subsequently, the 1-D deformation grids $u(x)$ and $v(y)$ are duplicated to produce 2-D column-wise grids $u(x, y)$ and row-wise grids $v(x, y) \in R^{H \times W}$. Finally, they are concatenated to form the deformation grids (u, v). Since pixels (x, y) in the same column (row) are mapped into (u, v) in the same column (row), thus the deformation is performed in a more structured way and suffers less from spatial distortion.

Then, as shown in Fig. 4 (b), by sampling features from the positions (u, v) on the original feature map F, we get the deformed feature map F'. As the coordinates of the deformation grids are fractional, we propose a novel attention based bilinear interpolation to calculate the feature $f(u, v)$ as

$$f(u, v) = \frac{[u_2 - u \quad u - u_1]}{(u_2 - u_1)(v_2 - v_1)} F(Q) \begin{bmatrix} v_2 - v \\ v - v_1 \end{bmatrix}, \tag{7}$$

where

$$F(Q) = \begin{bmatrix} (1 + A(Q_{11})) * f(Q_{11}) & (1 + A(Q_{12})) * f(Q_{12}) \\ (1 + A(Q_{21})) * f(Q_{21}) & (1 + A(Q_{22})) * f(Q_{22}) \end{bmatrix}, \tag{8}$$

where the positions $Q_{11}(u_1, v_1), Q_{12}(u_1, v_2), Q_{21}(u_2, v_1), Q_{22}(u_2, v_2)$ are the four nearest neighbors of the sampling point $P(u, v)$; see the demonstration in Fig. 5. Compared with the traditional bilinear interpolation, we modify the matrix $F(Q)$ by modulating each neighboring feature $f(Q_{ij})$ with the weight $1 + A(Q_{ij})$. As defined in Eq. (7) and Eq. (8), the feature of the neighbor Q_{ij} that is closer to P or shows a higher attention $A(Q_{ij})$ is aggregated to $f(u, v)$ with a larger weight. Consequently, compared with the original feature F in Fig. 4 (b), the features with large attention (red color) are densely sampled and spatially amplified while the features with small attention (blue color) shrank on the output feature F'. Therefore, in the attention-aware branch, the discriminative lesion features with larger attention are better propagated in the forward pass and promote classification. During the back-propagation, the lesion areas get larger gradients and accelerate network optimization.

TABLE I: Comparison results for the classification of WCE images.

Methods	N-Rec (%)	V-Rec (%)	I-Rec (%)	OA (%)	Cohen's Kappa (%)
DenseNet (B1)	97.15±0.42	92.50±1.10	87.59±0.58	92.40±0.30	88.28±0.49
DenseNet*2 + ToA + AGLA (TALA-Net)	97.33±0.38	**94.89±0.59**	**91.93±1.01**	**94.72±0.15**	**92.08±0.22**
Global branch	97.32±0.28	93.00±1.15	91.04±1.32	93.92±0.38	90.87±0.57
Attention-aware branch	97.32±0.39	94.67±0.67	91.48±1.17	94.49±0.27	91.74±0.40
DenseNet*2 (B2)	97.76±0.90	92.00±0.77	90.38±0.90	93.38±0.15	90.07±0.22
DenseNet*2 + ToA + SBS (B3)	97.54±0.95	93.54±0.89	91.04±1.22	94.05±0.12	91.07±0.17
DenseNet*2 + CAP + AGLA (B4)	**98.21±0.59**	92.89±0.97	89.26±0.77	93.45±0.27	90.18±0.40
DenseNet*2 + CMP + AGLA (B5)	97.99±0.01	93.11±0.97	89.68±0.62	93.60±0.27	90.40±0.40
DenseNet*2 + nonlocal + AGLA (B6)	97.33±0.01	94.65±1.32	89.26±0.67	93.75±0.22	90.63±0.34

C. Training and Testing Strategies

Loss function. Since all component modules are differentiable, we optimize the proposed TALA-Net in an end-to-end manner. The overall loss function is defined as

$$L = -\sum_{j\in\{1,2\}}\sum_{i\in D} t_i log\frac{e^{z_i^{(j)}}}{\sum_i e^{z_i^{(j)}}}, \tag{9}$$

where $j \in \{1, 2\}$ is the index of the two branches, $i \in D$ denotes the index of training samples, t_i and z_i represent the ground truth label and output logits of the i-th sample, respectively.

An advantage of the end-to-end training strategy is that the global branch and the attention-aware branch can be mutually promoted through gradients propagation. For example, if the ToA attention map fails to highlight the lesion areas, this may lead to suppression rather than amplification of the lesion features, thus degrading the performance of the attention-aware branch. Hence, the optimization of the attention-aware branch will force the global branch to produce more precise ToA attention maps. The accurate attention maps can further promote the feature representation and classification abilities of both branches.

Inference. In testing phase, the classification result of each image is obtained by the average of predictions from the global and attention-aware branches.

IV. Experimental Results

A. Experiment Setup

Dataset: The proposed method was validated on the CAD-CAP WCE dataset [24] containing 1812 images with resolution 512× 512. It consists of 600 normal images, 605 vascular lesions, and 607 inflammatory frames. The dataset was randomly divided into a training set (75%) and a testing set (25%) to conduct experiments. Each experiment was repeated three times and the average results were reported to evaluate the network performance. In order to maintain the robustness and stability of the training process, the training data was augmented through flip and rotation.

Implementation: Our model was implemented using Tensorflow on a desktop with Intel Core i7-7820X3.60GHz processors and a NVIDIA GeForce GTX 1080 Ti with 32 GB of RAM. The model was trained for 60 epochs utilizing stochastic gradient descent (SGD) with Nesterov momentum. We set the momentum to 0.9 and mini-batch size to 8. The learning rate was initialized as 0.01, and dropped by 0.1 after 40 epochs. The performance of lesion classification was evaluated by overall accuracy (OA), recall of normal images (N-Rec), recall of vascular lesion images (V-Rec), recall of inflammatory images (I-Rec), and Cohen's Kappa score.

B. Evaluation of Network Design

We evaluated the proposed TALA-Net and reported the classification performance in TABLE I. Compared with the vanilla "DenseNet (B1)", our TALA-Net obtains an improvement of 2.32% and 3.80% in terms of OA and Cohen's Kappa, respectively. Besides, the recall of abnormal images (V-Rec & I-Rec) gains larger improvement than the normal recall (N-Rec). These results verify that the TALA-Net achieves better lesion recognition performance through the amplification of lesion features. In the following subsections, we conducted comparison experiments to validate the effectiveness of the proposed AGLA and ToA modules.

*1) **Evaluation of the AGLA Module**:* As shown in TABLE I, in the TALA-Net, the attention-aware branch outperforms the global branch with an improvement of 0.57% and 0.87% in terms of OA and Cohen's Kappa, respectively. The reason is that the features deformed by the AGLA module have an enhanced representation of lesions thus lead to a stronger discriminative ability. As shown in Fig. 6, the deformation grids (Fig. 6 (c)) are more densely distributed in the lesion areas with higher attention, thus these regions are more densely sampled and spatially amplified on the outputs shown in Fig. 6 (e, g, i). What's more, compared with the original features (Fig. 6 (d, f, h)), the AGLA transformed features (Fig. 6 (e, g, i)) show suppressed responses in the background regions, which is due to the attention based bilinear interpolation utilized in the AGLA. Therefore, more discriminative deformed features with enhanced representation of the lesions and suppressed responses of the irrelevant features are transferred to the attention-aware branch, thus yielding stronger discrimination power and performance gains.

As shown in TABLE I, although the global branch in the TALA-Net shares an identical structure with the "B1" model, it gains an accuracy promotion by 1.52%. This result is consistent with our analysis in Section III-C. The inaccurate attention maps of the global branch are rectified through the

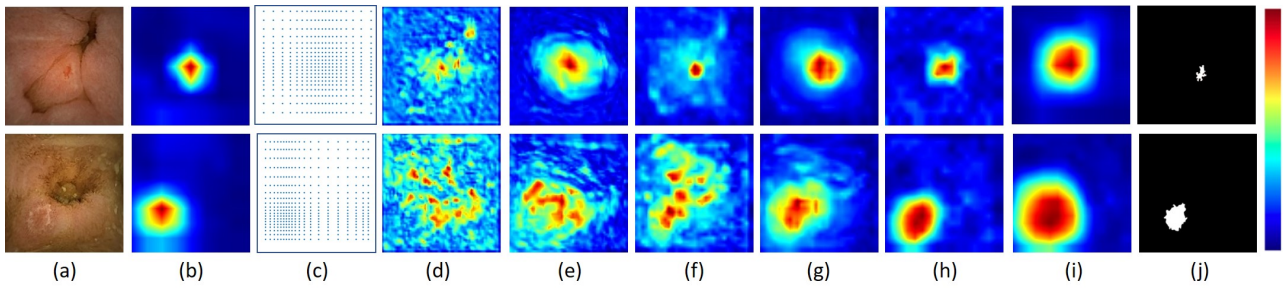


Fig. 6: Visualization of features before and after the AGLA module. (a) Input images. (b) ToA Attention maps. (c) Attention guided deformation grids. (d) Original feature maps in block1. (e) Feature maps in block1 after AGLA. (f) Original feature maps in block2. (g) Feature maps in block2 after AGLA. (h) Original feature maps in block3. (i) Feature maps in block3 after AGLA. (j) Ground truth masks. [Best viewed in color.]

end-to-end training, hence feature responses in the global branch are forced to concentrate on the lesion areas, which contributes to better classification performance.

Compared with the "B1", our TALA-Net contains an additional attention-aware branch with increased computation complexity. To prove that the performance gains of the TALA-Net are caused by the lesion amplification effect of the AGLA module rather than the additional computations, we constructed "DenseNet*2 (B2)" by ablating the AGLA module. In the "B2" model, hierarchical features from the global branch are directly transferred to the attention-aware branch. As shown in TABLE I, although the OA of "B2" is 0.98% higher than the "B1", it is much lower than the TALA-Net. This result proves the performance gains of the TALA-Net is mainly caused by the lesion amplification effect of the AGLA module.

Furthermore, since the saliency-based sampler (SBS) in [22] and our AGLA module share the similar idea of zooming in important regions, we constructed a comparison experiment "B3" by replacing the AGLA with the SBS module. As demonstrated in TABLE I, the accuracy of the "B3" is 0.67% lower than the TALA-Net, thus proving the superiority of the AGLA over the SBS module. The reason is that the AGLA module suffers less from spatial distortion and further enhances the lesion representations by utilizing attention based bilinear interpolation.

*2) **Evaluation of the ToA Module**:* To evaluate the performance of the ToA attention maps in indicating lesion positions, we compared them with the high-order Nonlocal attention [21] and the first-order attention maps generated through Channel-wise Average Pooling (CAP) [16] and Channel-wise Max Pooling (CMP) [17].

We constructed the models "B4", "B5", and "B6" by replacing the ToA module with the CAP, CMP, and Nonlocal modules, respectively. As shown in TABLE I, the "B4", "B5", and "B6" perform inferiorly against the TALA-Net with an overall accuracy of 93.45%, 93.60%, and 93.75%, respectively. A possible reason is that these three attention modules are insufficient to provide correct guidance for the attention-aware branch, thus leading to limited performance. To validate the above assumption, we compared the attention maps produced by different modules. The image in the first

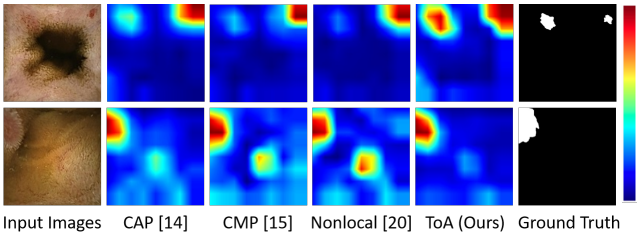


Fig. 7: Comparison between the attention maps produced by different attention modules. The top and bottom rows show examples of vascular lesion and inflammatory, respectively. [Best viewed in color.]

row of Fig. 7 contains two segments of vascular lesions. On the CAP and CMP attention maps, the lesion at the upper left corner has a relatively low attention value, this is due to its similarity with the surrounding normal features. On the corresponding ToA attention map, responses at that lesion area are enhanced. The underlying reason is that benefiting from the ToA module, low-response lesion features are enhanced by aggregating features from other long-range dependent lesion areas. For the inflammatory image shown in the second row of Fig. 7, false-positive responses are observed in the central part of the CAP and CMP attention maps. This indicates that the first-order attentions have limited capability of distinguishing lesions from hard mimics. In contrast, the responses of hard mimics are effectively suppressed on the ToA attention map. This is because the features from the background normal areas are aggregated to suppress the responses of the hard mimics. Note that although sharing a similar idea of aggregating long-range dependent features, the attention maps produced by the Nonlocal module fails to highlight the lesion regions correctly. The reason is that the multiplication of two different transformed features in the Nonlocal module is insufficient to capture the correct feature dependencies. This can further validate the superiority of the second-order covariance utilized in our ToA module.

According to the experimental results, we conclude that the ToA module can refine attention maps by suppressing background noises and enhancing lesion features, thus more accurate guidance for the AGLA module would help to effectively amplify the lesion features and boost classification.

TABLE II: Comparison with state-of-the art methods for classification of WCE images.

Methods	N-Rec (%)	V-Rec (%)	I-Rec (%)	OA (%)	Cohen's Kappa (%)
Fan et al. [2]	91.95±1.03	90.44±1.46	73.37±2.13	85.27±0.59	77.90±0.89
Iakovidis et al. [3]	77.23±0.05	78.79±0.64	75.00±3.12	77.01±1.24	65.52±1.86
Alaskar et al. [4]	89.49±1.47	89.33±1.68	77.40±2.52	85.40±0.49	78.13±0.73
Xing et al. [7]	95.99±1.15	91.30±0.35	88.61±1.41	91.96±0.22	87.95±0.34
Guo et al. [8]	97.10±1.15	93.30±1.65	90.18±0.65	93.53±0.36	90.29±0.55
TALA-Net (Ours)	**97.33±0.38**	**94.89±0.59**	**91.93±1.01**	**94.72±0.15**	**92.08±0.22**

C. Comparison with Other Methods

We further compared our method with five state-of-the-art deep learning-based image-level classification methods [2]–[4], [7], [8] in the WCE field. For a fair comparison, we used the implementations of other methods provided by the authors. Due to the relatively shallow off-the-shelf models, [2]–[4] showed limited discriminative capability and produced unsatisfactory performance. In contrast, the algorithms in [7], [8] produced relatively good performance, which can be attributed to their advanced backbone (i.e., DenseNet) and the delicately designed attention mechanisms. Compared with [8], our algorithm obtained performance gains of 1.19% in terms of OA. The comparison results validate the superiority of the proposed TALA-Net.

V. CONCLUSIONS

In this paper, we proposed a two-branch third-order attention guided lesion amplification network for the challenging classification task of WCE images. The main idea is to achieve better inspection of small lesions through amplification of the lesion features. Our proposed third-order attention can accurately highlight the potential lesion regions with image labels only. Then, a novel attention guided lesion amplification module was proposed to zoom in the suspected lesion features, thus leading to more discriminative feature representations and better classification performance. Extensive experiments on a publicly available WCE dataset validated the superiority of the proposed method which outperforms other state-of-the-art approaches.

REFERENCES

[1] Gavriel Iddan, Gavriel Meron, Arkady Glukhovsky, and Paul Swain. Wireless capsule endoscopy. *Nature*, 405(6785):417, 2000.

[2] Shanhui Fan, Lanmeng Xu, Yihong Fan, Kaihua Wei, and Lihua Li. Computer-aided detection of small intestinal ulcer and erosion in wireless capsule endoscopy images. *Phys. Med. Biol.*, 63(16):165001, 2018.

[3] Dimitris K Iakovidis, Spiros V Georgakopoulos, Michael Vasilakakis, Anastasios Koulaouzidis, and Vassilis P Plagianakos. Detecting and locating gastrointestinal anomalies using deep learning and iterative cluster unification. *IEEE Trans. Med. Imag.*, 37(10):2196–2210, 2018.

[4] Haya Alaskar, Abir Hussain, Nourah Al-Aseem, Panos Liatsis, and Dhiya Al-Jumeily. Application of convolutional neural networks for automated ulcer detection in wireless capsule endoscopy images. *Sensors*, 19(6):1265, 2019.

[5] Xiaohan Xing, Xiao Jia, and Max-HQ Meng. Bleeding detection in wireless capsule endoscopy image video using superpixel-color histogram and a subspace knn classifier. In *Proc. EMBC*, pages 1–4. IEEE, 2018.

[6] Yejin Jeon, Eunbyul Cho, Sehwa Moon, Seung-Hoon Chae, Hae Young Jo, Tae Oh Kim, et al. Deep convolutional neural network-based automated lesion detection in wireless capsule endoscopy. In *International Forum on Medical Imaging in Asia 2019*, volume 11050, page 110501N. International Society for Optics and Photonics, 2019.

[7] Xiaohan Xing, Yixuan Yuan, Xiao Jia, et al. A saliency-aware hybrid dense network for bleeding detection in wireless capsule endoscopy images. In *Proc. ISBI*, pages 104–107. IEEE, 2019.

[8] Xiaoqing Guo and Yixuan Yuan. Triple anet: Adaptive abnormal-aware attention network for wce image classification. In *Proc. MICCAI*, pages 293–301. Springer, 2019.

[9] Xiao Jia, Xiaohan Xing, Yixuan Yuan, Lei Xing, and Max Q-H Meng. Wireless capsule endoscopy: A new tool for cancer screening in the colon with deep-learning-based polyp recognition. *Proceedings of the IEEE*, 108(1):178–197, 2019.

[10] Xiaoqing Guo and Yixuan Yuan. Semi-supervised wce image classification with adaptive aggregated attention. *Medical Image Analysis*, page 101733, 2020.

[11] Xiaohan Xing, Yixuan Yuan, and Max Q-H Meng. Zoom in lesions for better diagnosis: Attention guided deformation network for wce image classification. *IEEE Trans. Med. Imag.*, 2020.

[12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Adv Neural Inf Process Syst*, pages 1097–1105, 2012.

[13] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proc. CVPR*, pages 1–9, 2015.

[14] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *Proc. CVPR*, pages 3156–3164, 2017.

[15] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proc. ECCV*, pages 3–19, 2018.

[16] Junsuk Choe and Hyunjung Shim. Attention-based dropout layer for weakly supervised object localization. In *Proc. CVPR*, pages 2219–2228, 2019.

[17] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016.

[18] Jo Schlemper, Ozan Oktay, Michiel Schaap, Mattias Heinrich, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention gated networks: Learning to leverage salient regions in medical images. *Medical image analysis*, 53:197–207, 2019.

[19] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proc. CVPR*, pages 4700–4708, 2017.

[20] Peihua Li, Jiangtao Xie, Qilong Wang, and Wangmeng Zuo. Is second-order information helpful for large-scale visual recognition? In *Proc. ICCV*, pages 2070–2078, 2017.

[21] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proc. CVPR*, pages 7794–7803, 2018.

[22] Adria Recasens, Petr Kellnhofer, Simon Stent, Wojciech Matusik, and Antonio Torralba. Learning to zoom: a saliency-based sampling layer for neural networks. In *Proc. ECCV*, pages 51–66, 2018.

[23] Heliang Zheng, Jianlong Fu, Zheng-Jun Zha, and Jiebo Luo. Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition. In *Proc. CVPR*, pages 5012–5021, 2019.

[24] Xavier Dray, Cynthia Li, Jean-Christophe Saurin, Franck Cholet, Gabriel Rahmi, JP Le Mouel, C Leandri, Stéphane Lecleire, Xavier Amiot, Jean-Michel Delvaux, 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.