S & M 3824

Medical Image Segmentation on Attention-based Gaussian Blurring

Yue-Tian Mao,[1] Qing-Song Liu,[1] Yin-Feng Fang,[1*] Guo-Zhang Jiang,[2] and Du Jiang[3]

[1]School of Communication Engineering, Hangzhou Dianzi University, Hangzhou310018, China
[2]Hubei Key Laboratory of Mechanical Transmission and Manufacturing Engineering, Wuhan University of Science and Technology, Wuhan 430081, China
[3]Key Laboratory of Metallurgical Equipment and Control Technology of Ministry of Education, Wuhan University of Science and Technology, Wuhan 430081, China

(Received July 5, 2024; accepted August 30, 2024)

Keywords: eye-gaze, attention, eye-tracking sensor, segmentation, deep learning

In this study, we explore the integration of eye-gaze (EG) information obtained via eye-tracking sensors to enhance medical image segmentation. EG is additional information on an image, and to incorporate EG information in medical image segmentation, we apply Gaussian blurring masked by the collected attention maps generated by eye tracking. The variance of distribution on each pixel is adjusted in a certain way by EG information. After applying Gaussian blurring, classic models including UNet, FCN, and other models were trained and compared. The results indicate that incorporating EG information in addition to preprocessing data yields superior performance on certain metrics, demonstrating notable advantages in accurately identifying isolated polyps that grow on the surface of a human tissue. This innovative approach highlights the potential of combining sensor-derived data with advanced image processing techniques to improve medical diagnostics.

1. Introduction

Medical image segmentation is a vital task in medical diagnosis, treatment planning, image-guided intervention, and other medical treatments. Over the past few decades, machine learning and deep learning methods have shown their dominance in this field.(1–3) They can analyze the complex images collected by different capture devices and provide us with satisfactory results under fluctuating circumstances. However, deep learning methods require a large number of annotated datasets to train so that they can be reliable. Moreover, it takes much effort for people with a medical background to annotate images so that the number of datasets of different diseases is limited. The number of images in each dataset is relatively small for the same reason. This leads to the urgent need for accelerating training convergence and learning features. To address this issue, certain types of data preprocessing, including the integration of sensor-derived data such as eye-tracking information, have been proposed to expedite training and improve results.(4) Eye tracking is a sensor technology that detects a person's presence and tracks

[*]Corresponding author: e-mail: yinfeng.fang@hdu.edu.cn

https://doi.org/10.18494/SAM5189

what they are looking at in real time. Over the past decades, eye-gaze (EG) attention has been analyzed to improve the segmentation and classification of diseases. Apparent linkages among EG information, image content, and diagnostic results have been found in a previous study by pooling the data from all test takers, and some other studies have demonstrated that EG information can generate segmentation masks suitable for deep learning segmentation.(5,6) By incorporating eye-tracking sensors to collect gaze data, we can generate attention maps that guide preprocessing techniques such as Gaussian blurring, ultimately enhancing the performance of segmentation models.

Conventional data preprocessing on images includes resizing, color space conversion, normalization, and filtering. Filtering is an effective way to reduce noise or artifacts when applying different degrees of blurring on images.(7) It is a good way to enhance the robustness of the models. To make a network focus on detecting lesion features, some researchers have proposed new data preprocessing methods using masks. Some of them use the probability distribution of organizations on images according to the labels made by experts(8) and some use saliency detection on edge boxes.(9) However, the limitation that these mentioned data preprocessing methods requires a certain amount of work by experts still exists, which makes it difficult to acquire an appropriate number of images and label them to form the probability distribution or saliency object positioning.

Compared with the aforementioned masking methods, masks based on eye-tracking information can be more easily combined with data preprocessing. Eye-tracking sencors are used to collect direct reactions of collectors, which cannot be achieved by other approaches. The eye-track recording process utilizing eye-tracking sensors is time-efficient and easy to apply. While accurate segmentation labeling requires experts, eye-track recording does not need to be performed by people with medical backgrounds, because lesions such as ulcers or polyps can be detected by anyone; thus, this method can be utilized in every medical segmentation training process with little effort.

2. Materials and Methods

In this section, we introduce the dataset used in experiments and the application of basic image preprocesses. The attention maps used in Gaussian blurring are also used in the data preprocessing stage of image segmentation.(10) The comparison between fundamental data preprocessing methods and the preprocessing method intergrating EG information will be illustrated.

The overall process of integrating EG information into medical image segmentation is depicted in Fig. 1. Initially, original medical images are obtained and used as the primary data source for analysis. EG data are then collected from observers viewing these images using eye-tracking sensors, generating attention maps that highlight areas of visual foci. These attention maps are used to generate a standard deviation map, which guides the application of Gaussian blurring on an original image, creating an EG-information-incorporated image. Finally, after incorporating attention maps in the data preprocessing stage, the images are used to train different models.

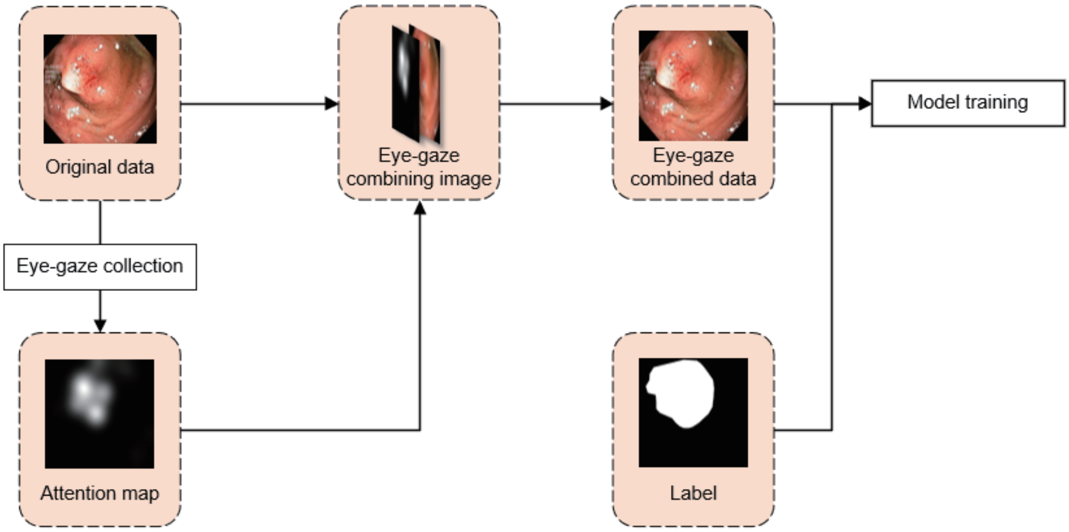

Fig. 1. (Color online) Pipeline of image segmentation integrating EG information.

2.1 Dataset and EG collection

Colorectal polyp segmentation in medical images is a demanding task. Kvasir-SEG is an endoscopic dataset for the pixel-level segmentation of colonic polyps.(11) The dataset includes 1000 images and labels, including polyps in different places and with different numbers, and each image includes at least one polyp. Polyps can be distinguished from the normal mucosa by color and surface pattern, but they vary in shape and color, so it is still challenging to separate them using deep learning networks. The dataset has flat, elevated, or pedunculated polyps in different images so that the effectiveness of the segmentation method can be tested. The resolutions of the images in the dataset vary, ranging from 332 × 487 to 1920 × 1072 pixels. When sent to different networks to train, each image is normalized to 512 × 512 pixels.

ET is collected with the Tobii eye tracker (Tobii AB, Sweden). The eye tracker can be attached to most computers and located below or above the screen to detect eyesight, as shown in Fig. 2. During eye-track recording, medical images are displayed on the computer while the eye tracker captures the gaze points in real time. The observation time for each image is approximately 8 s. The number of images observed depends on the actual application scenario. The medical images can be the most common RGB images, can be converted into color spaces such as LUV (cieluv, cie1976) and HSV (Hue, Saturation and Value), or can be multimodal. Then, 10 to 15 eye movement data are collected and reprocessed using curve fitting methods to highlight the areas where observers gaze for a longer period and generate eye movement heatmaps (expressed in the form of grayscale images).

2.2 Fundamental data preprocess

During experiments, images, labels, and attention maps are all reshaped to 512 × 512. To prevent overfitting, the brightness, contrast, saturation, and hue of the images are randomly

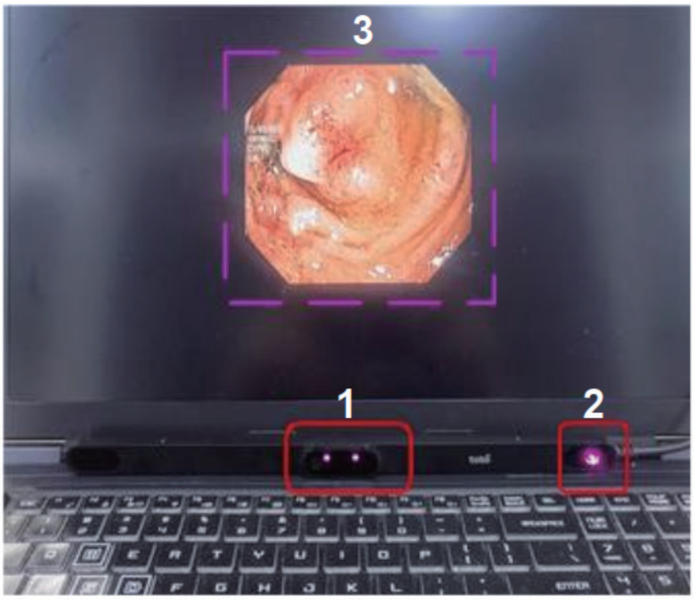

Fig. 2. (Color online) Layout of Tobii eye tracker. 1 and 2 capture EG at 60 HZ, and 3 is the medical image shown on the screen.

changed, and the images are randomly flipped horizontally and vertically. The brightness of the image is randomly changed by a factor uniformly chosen in the range [0.6, 1.4], the contrast by [0.5, 1.5], the saturation by [0.75, 1.25], and the hue by [0.001, 0.01]. Moreover, to hasten convergence when training neural networks, image tensors are normalized. The average and standard deviation of each color channel in RGB are adjusted to 0.5 to train the models to detect useful features more rapidly.

2.3 Attention-based Gaussian blurring

Gaussian blurring is a common image processing technique used to reduce noise and smoothen the appearance of images. In Gaussian blurring, each pixel in the image is replaced by a weighted average of its neighbors and itself, with the weights determined by a 2D Gaussian function. The Gaussian blurring process involves convolving an image with a Gaussian kernel, which is a matrix of values that represent the weights for each pixel in the kernel.

$$P(x,y)=\sum_{j=y-n/2}^{y+n/2}\sum_{i=x-n/2}^{x+n/2}W(i,j)P(i,j) \tag{1}$$

Here, n represents the kernel size and (x, y) are the coordinates, and

$$W(i,j)=\frac{1}{2\pi\sigma^2}e^{-\frac{i^2+j^2}{2\sigma^2}}, \tag{2}$$

where σ denotes the standard deviation.

The kernel is typically a 2D matrix, and the size of the kernel and the variance of the distribution of the function generating the kernel can vary depending on the desired level of blurring. The approach to combining attention maps with Gaussian blurring is to determine the variance we use on each pixel by attention level. After adjusting the variance of the Gaussian kernel on pixels, an image with variant clarity is generated. In positions where the attention level is higher, the image is clearer, and in places where the attention level is lower, the image appears blurred to a greater extent.

After processing, the features of a certain part are more easily detected if the attention is focused on that part. However, the original purpose of Gaussian blurring still needs to function. To enhance the robustness of the model, different images still need to be blurred to different extents. Therefore, the base of the variance is randomly set within a certain range.

$$\sigma_{base} = rand(min, max), \tag{3}$$

where *min* and *max* are the minimum and maximum of the base standard deviation, respectively.

We add the numerical value that is formerly illustrated on the base.

$$\sigma_{+}\left(x, y\right) = w\left(1 - p\left(x, y\right)\right), \tag{4}$$

where $p(x, y)$ is the attention level on each pixel and w is the weight that determines the clarity contrast between different attention levels.

The final σ on each pixel when applying Gaussian blurring is

$$\sigma = \sigma_{base} + \sigma_{+}. \tag{5}$$

Figure 3 shows examples of images to which Gaussian blurring is applied, where the weight is set to 0.3, σ_{base} equals 1, and the kernel size is set to 15.

It is evident that after applying Gaussian blurring with masking with attention maps, the resulting image is clear with a higher attention level, retaining a more detailed texture with features for models to learn.

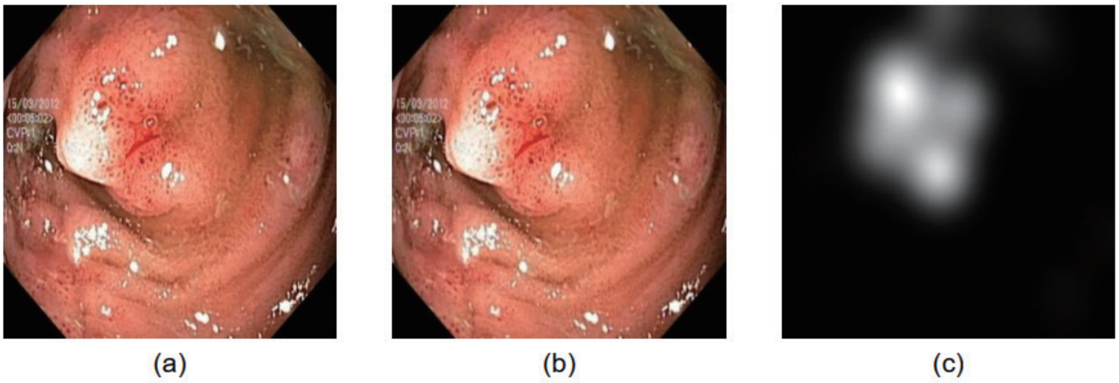

Fig. 3. (Color online) (a) Original image, (b) image obtained after attention-based Gaussian blurring, and (c) attention map.

After applying positively masked Gaussian blurring where images are clearer at gazed areas, an additional experiment is conducted. In the experiment, the parts with a higher attention level appear blurred to a greater extent and the parts with a lower attention level have a higher clarity.

$$\sigma = \sigma_{base} - \sigma_{+} \tag{6}$$

Figure 4 shows an example of reversely adjusted blurring, where we detract σ_{+} from σ_{base}. The weight is set to 0.3, the base equals 1, and the kernel size is set to 15.

The new method is the reverse of the original one. It is performed to determine the effects of attention maps on model training.

2.4 Segmentation model

Eye tracking is combined with data preprocessing, so the method can be utilized in all segmentation models. In the experiment, 10 image segmentation models are tested to see the performances of segmentation results on the 10 models, namely, UNet,(12) FCN,(13) TransUNet,(14) ResAttUNet,(15) UNet++,(16) FPN,(17) Manet,(18) Linknet,(19) PSPNet,(20) and DeepLabV3.(21) In all experiments, the learning rate is set at l × 10^{-4} and the batch size is set at 32. The Dice loss and Adam optimizer are used in the training.

2.5 Experimental setup and evaluation indicators

To perform the experiment, a public colon polyp dataset Kvasir-SEG is employed. The dataset contains 1000 polyp images with labeling. A classic ratio of the training, validation, and test sets (6:2:2) is utilized in the experiment. 600 randomly chosen images are included in the training set, 200 in the validation set, and the rest in the test set. After implementing enough sets of σ_{base}, the parameter set giving the best performance is chosen. σ_{base} is set to the range (0, 1) and the weight to 0.3. The training process is implemented on an Ubuntu 16.04.4 LTS OS

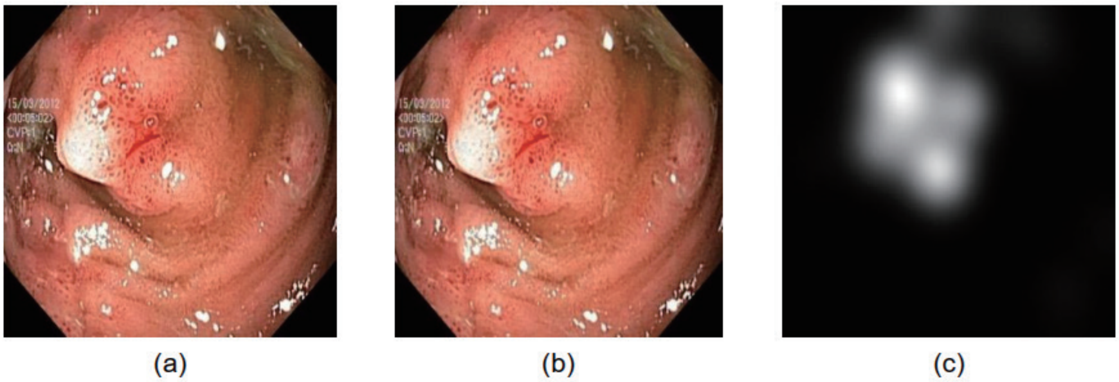

Fig. 4. (Color online) (a) Original image, (b) resulting image after reverse attention-based Gaussian blurring, and (c) attention map.

running on a local cluster on a server with an RTX 4090 graphics card, 16-core GPU, and 90 GB of memory.

Four metrics are employed to evaluate the training performance as listed below. In the following equations, N represents the number of images, TP_i the number of true-positive pixels, FN_i the number of false-negative pixels, and FP_i the number of false-positive pixels, where i represents the i-th image.

2.5.1 Mean Dice coefficient (*mDice*)

The Dice similarity coefficient is a measure of the overlap between the predicted segmentation mask and the ground truth mask.

$$mDice = \frac{1}{N}\sum_{i=1}^{N}\frac{2TP_i}{2TP_i + FP_i + FN_i} \tag{7}$$

2.5.2 Mean Intersection over Union (*mIou*)

MIou calculates the average of the Intersection over Union (*Iou*) values for each object in a dataset, where *Iou* is the ratio of the intersection of the predicted value to that of the ground truth.

$$mIou = \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i}{TP_i + FP_i + FN_i} \tag{8}$$

2.5.3 Mean precision (*mPrecision*)

Precision is calculated by dividing the number of correctly segmented pixels by the total number of pixels segmented as positive.

$$mPrecision = \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i}{TP_i + FP_i} \tag{9}$$

2.5.4 Mean recall (*mRecall*)

Recall is calculated by dividing the number of correctly segmented pixels by the total number of pixels that should have been segmented as positive.

$$mRecall = \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i}{TP_i + FN_i} \tag{10}$$

3. Results

In this section, we present the outcomes of our image segmentation experiments. The results are crucial to understanding the effectiveness and tendencies of the different approaches we implemented. This part of the article will detail the performance comparison and offer quantitative and visual comparisons between segmentation results before and after combining EG information. Tables 1–4 show the mean *Dice*, *Iou*, *Precison*, and *Recall* performance on the dataset. In each table, the results for the original images, images after Gaussian blurring positively adjusted by attention level, and images reversely adjusted by attention level are compared.

The integration of EG information through attention-based Gaussian blurring shows an improvement in segmentation accuracy, particularly in terms of Dice coefficient and *Precision* across several models. It can be observed that the models show significant gains when attention-based blurring is applied, especially for UNet with a 12.25% improvement in *Precision*. This suggests that the method enhances the models' ability to focus on relevant features, reducing noise and improving feature detection.

Table 1
Comparison of Dice performance among the original, positively Gaussian blurred, and reversely adjusted Gaussian blurred images.

Method	Original	Positive	Reverse
UNet	0.7462	0.7825	0.7626
FCN	0.8246	0.8401	0.8253
TransUNet	0.8803	0.8645	0.8713
ResAttUNet	0.8035	0.7915	0.8019
UNet++	0.8743	0.8877	0.8789
FPN	0.8743	0.8917	0.8782
MANet	0.8824	0.8620	0.8724
LinkNet	0.8743	0.8817	0.8782
PSPNet	0.7591	0.7378	0.7542
DeepLabV3	0.8776	0.8848	0.8855

Table 2
Comparison of *Iou* performance among the original, positively Gaussian blurred, and reversely adjusted Gaussian blurred images.

Method	Original	Positive	Reverse
UNet	0.6362	0.7356	0.6578
FCN	0.7392	0.7360	0.8253
TransUNet	0.8106	0.7874	0.7981
ResAttUNet	0.7153	0.7243	0.7149
UNet++	0.8098	0.8294	0.8162
FPN	0.8066	0.8320	0.8157
MANet	0.8191	0.7986	0.8111
LinkNet	0.8106	0.8193	0.8134
PSPNet	0.6544	0.6322	0.6496
DeepLabV3	0.8119	0.8223	0.8205

Table 3
Comparison of *Precision* performance among the original, positively Gaussian blurred, and reversely adjusted Gaussian blurred images.

Method	Original	Positive	Reverse
UNet	0.7721	0.8356	0.8130
FCN	0.8443	0.9191	0.8253
TransUNet	0.8970	0.9036	0.9306
ResAttUNet	0.8220	0.8557	0.8196
UNet++	0.9076	0.9036	0.8937
FPN	0.8901	0.8977	0.9047
MANet	0.8998	0.9152	0.9334
LinkNet	0.9061	0.9055	0.9333
PSPNet	0.8553	0.8589	0.8488
DeepLabV3	0.8751	0.9229	0.9217

Table 4
Comparison of *Recall* performance among the original, positively Gaussian blurred, and reversely adjusted Gaussian blurred images.

Method	Original	Positive	Reverse
UNet	0.8111	0.7695	0.7938
FCN	0.8607	0.7888	0.8253
TransUNet	0.8902	0.8581	0.8432
ResAttUNet	0.8525	0.8232	0.8529
UNet++	0.8806	0.9021	0.8976
FPN	0.8963	0.9152	0.8908
MANet	0.9011	0.8417	0.8491
LinkNet	0.8897	0.8417	0.8644
PSPNet	0.7564	0.7180	0.7508
DeepLabV3	0.9173	0.8849	0.8749

For some of the models such as Linknet, they show better results in the reversely adjusted blurring experiment. The reason for this is probably that the reverse Gaussian blurring makes it even more difficult for the model to learn the features of lesion parts, which in turn enhances the robustness of the model, making it easier to detect clear polyp images on test data.

To have a clearer presentation of the new method, Fig. 5 shows the average of the four metrics for 10 models, comparing between original data and data with EG information incorporated. As can be discovered, three out of the four metrics have been improved to some degree, among which *Precision* shows the biggest improvement. The standard deviation of *Precision* on data with EG information incorporated is also the smallest. This strongly suggests that the *Precision* of most of the models increases and EG information has helped the models learn the features and characteristics of polyps better.

Other than that, we discovered that after integrating EG information, the models have become more sensitive to isolated polyps. The models trained on original data tend to wrongly segment polyps that are in fact normal tissues. The models trained on data after attention-based Gaussian blurring can effectively prevent that. As an example, it can be seen from Fig. 6 that the segmentation result on original data shows that the model incorrectly detected a bubble as an isolated polyp that is not in the ground truth. However, the same model trained on images after

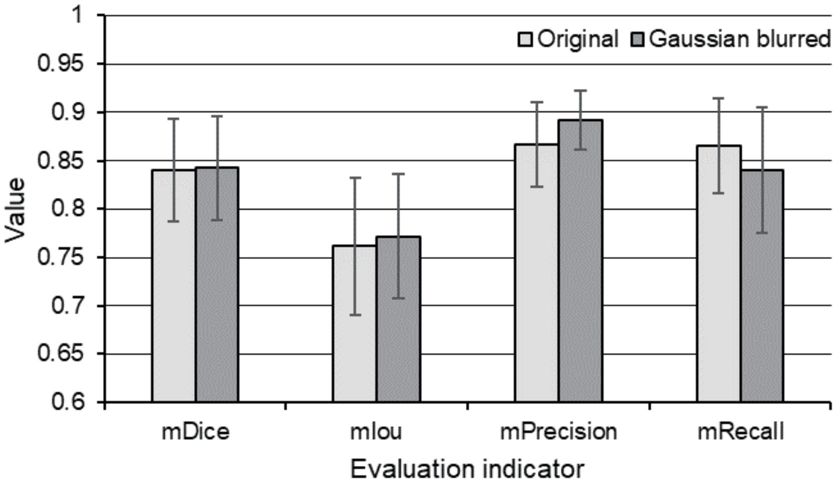

Fig. 5. Each column shows the average metric value for over 10 models. Light gray columns represent results for the original data, and dark gray columns represent results for the EG information incorporated.

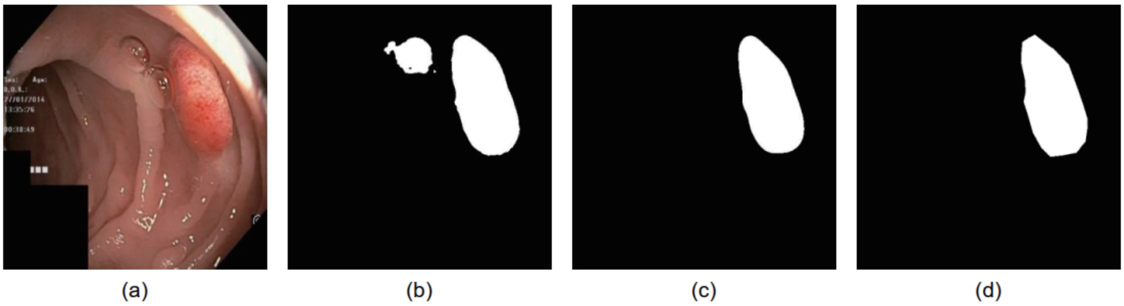

Fig. 6. (Color online) Segmentation results of Unet: (a) test image, (b) result for original, (c) result for EG information incorporated, and (d) ground truth.

positive attention-based Gaussian blurring does not make this mistake, the result is basically the same as the ground truth.

4. Conclusions and Future Works

The study introduces a novel method of improving deep learning training performance by incorporating eye-tracking sensor data from non-experts. This method provides a universal way to enhance the training process during image analysis. The standard deviation of Gaussian blurring on each pixel is adjusted by the attention level collected in data preprocessing. By integrating EG information into the preprocessing pipeline, we can improve the accuracy of medical image segmentation models. Regarding the basic and frequently used model UNet, the average over the four evaluation metrics is improved by 3.94%. Most of the models tested show better results after EG information is incorporated. This approach leverages natural human attention patterns collected by sensors to guide machine learning models.

The integration of EG information into data preprocessing can be separated from the model training process so that the cost of integrating attention-based Gaussian blurring is relatively low. Moreover, if conditions permit, it can be applied in real-time examination with only an additional eye tracker. When doctors perform colonoscopies, the eye tracker can collect EG from doctors looking at the screen and combine the information in data preprocessing before sending the data to the pretrained network. The result can be shown on the screen in real time, helping doctors concentrate on the lesion parts.

However, how to adjust the parameters to obtain the best performance still needs experience and the results need to be analyzed more precisely and effectively as well. In our future work, we will continue refining the integration of sensor data and exploring the application of this new method incorporating EG information in data preprocessing across various imaging tasks. In addition, we plan to explore how the information can be incorporated in different parts of the training process such as model building. This study lays the groundwork for future innovations at the intersection of sensor technology, human–computer interaction, and image processing, highlighting the potential for combining sensor-derived data with advanced machine learning techniques.

Acknowledgments

The project was supported in part by the National Key Research and Development Program of China (No. 2023YFE0114900), the Open Fund of Hubei Key Laboratory of Mechanical Transmission and Manufacturing Engineering at Wuhan University of Science and Technology (No. MTMEOF2022A01), and Xinmiao Talent Plan (No. 2023R407016). We thank the reviewers for their useful feedback that helped improve this paper.

References

1 I. Galić, M. Habijan, H. Leventić, and K. Romić: Electronics **12** (2023) 4411. https://doi.org/10.3390/electronics12214411
2 M. Z. Khan, M. K. Gajendran, Y. Lee, and M. A. Khan: IEEE Access **9** (2021) 83002. https://doi.org/10.1109/ACCESS.2021.3086530
3 Z. Ma, J. M. R. S. Tavares, R. N. Jorge, and T. Mascarenhas: Comput. Methods Biomech. Biomed. Eng. **13** (2010) 235. https://doi.org/10.1080/10255840903131878
4 R. Zhou, S. K. Ng, J. J. Y. Sung, W. W. B. Goh, and S. H. Wong: Comput. Struct. Biotechnol. J. **21** (2023) 4804. https://doi.org/10.1016/J.CSBJ.2023.10.001
5 J. N. Stember, H. Celik, E. Krupinski, P. D. Chang, S. Mutasa, B. J. Wood, A. Lignelli, G. Moonis, L. H. Schwartz, S. Jambawalikar, and U. Bagci: J. Digital Imaging **32** (2109) 597. https://link.springer.com/article/10.1007/s10278-019-00220-4
6 G. Tourassi, S. Voisin, V. Paquit, and E. Krupinski: J. Am. Med. Inf. Assoc. **20** (2013) 1067. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3822113/
7 R. Kaur and K. Kaur: IJEM-Int. J. Eng. Manuf. (IJEM) **6** (2016) 38. https://doi.org/10.5815/ijem.2016.06.04
8 B. Garcia-Zapirain, A. Shalaby, A. El-Baz, and A. Elmaghraby: Comput. Biol. Med. **90** (2017) 137. https://doi.org/10.1016/j.compbiomed.2017.09.015
9 H. Yu, M. Xian, and X. Qi: 2014 IEEE Int. Conf. Image Processing (ICIP, 2014) 4412–4416. https://doi.org/10.1109/ICIP.2014.7025895
10 T. G. Devi, N. Patil, S. Rai, and C. S. Philipose: Life **13** (2023) 348. https://doi.org/10.3390/LIFE13020348

11 P. H. Smedsrud, V. Thambawita, S. A. Hicks, H. Gjestang, O. O. Nedrejord, E. Næss, H. Borgli, D. Jha, T. J. D. Berstad, S. L. Eskeland, M. Lux, H. Espeland, A. Petlund, D. T. D. Nguyen, E. Garcia-Ceja, D. Johansen, P. T. Schmidt, E. Toth, H. L. Hammer, T. de Lange, M. A. Riegler, and P. Halvorsen: Sci. Data **8** (2021) 142. https://doi.org/10.1038/S41597-021-00920-Z

12 O. Ronneberger, P. Fischer, and T. Brox: arXiv (2015). https://arxiv.org/abs/1505.04597

13 J. Long, E. Shelhamer, and T. Darrell :arXiv (2014). https://arxiv.org/abs/1411.4038

14 J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou: arXiv (2021). https://arxiv.org/abs/2102.04306

15 A. Mohammed: arXiv (2022). https://arxiv.org/abs/2210.08506

16 Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, held in conjunction with MICCAI 2018, Granada, Spain, (2018) 11045 13–11. https://arxiv.org/pdf/1807.10165

17 T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie: arXiv (2016). https://arxiv.org/abs/1612.03144

18 P. He, L. Jiao, R. Shang, S. Wang, X. Liu, D. Quan, K. Yang, and D. Zhao: IEEE Trans. Geosci. Remote Sens. **60** (2022) 1. https://ieeexplore.ieee.org/abstract/document/9785828

19 A. Chaurasia and E. Culurciello: arXiv (2017). https://arxiv.org/abs/1707.03718

20 H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia: arXiv (2016). https://arxiv.org/abs/1612.01105

21 L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam: arXiv (2017). https://arxiv.org/abs/1706.05587