Eye-Guided Dual-Path Network for Multi-organ Segmentation of Abdomen

Chong Wang, Daoqiang Zhang, and Rongjun Ge(✉)

College of Computer Science and Technology,
Nanjing University of Aeronautics and Astronautics, Nanjing, China
rongjun.ge@nuaa.edu.cn

Abstract. Multi-organ segmentation of the abdominal region plays a vital role in clinical such as organ quantification, surgical planning, and disease diagnosis. Due to the dense distribution of abdominal organs and the close connection between each organ, the accuracy of the label is highly required. However, the dense and complex structure of abdominal organs necessitates highly professional medical expertise to manually annotate the organs, leading to significant costs in terms of time and effort. We found a cheap and easily accessible form of supervised information. Recording the areas by the eye tracker where the radiologist focuses while reading abdominal images, gaze information is able to force the network model to focus on relevant objects or features required for the segmentation task. Therefore how to effectively integrate image information with gaze information is a problem to be solved. To address this issue, we propose a novel network for abdominal multi-organ segmentation, which incorporates radiologists' gaze information to boost high-precision segmentation and weaken the demand for high-cost manual labels. Our network includes three special designs: 1) a dual-path encoder to further integrate gaze information; 2) a cross-attention transformer module (CATM) that embeds human cognitive information about the image into the network model; and 3) multi-feature skip connection (MSC), which combines spatial information during down-sampling to offset the internal details of segmentation. Additionally, our network utilizes discrete wavelet transform (DWT) to further provide information on organ location and edge in different directions. Extensive experiments performed on the publicly available Synapse dataset demonstrate that our proposed method can integrate effectively gaze information and achieves Dice similarity coefficient (DSC) up to 81.87% and Hausdorff distance (HD) reduction to 11.96%, as well as gain high-quality readable visualizations. Code will be available at https://github.com/code-Porunacabeza/gaze_seg/.

Keywords: Eye-tracking · Multiorgan segmentation · Computer-Aided Diagnosis

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14226, pp. 23–32, 2023.
https://doi.org/10.1007/978-3-031-43990-2_3

1 Introduction

The automatic segmentation of abdominal multiple organs is clinically significant in extremely that can significantly reduce clinical resource costs. However, the task of abdominal organ segmentation is difficult. The number of abdominal organs is large, and these multiple organs show diverse characteristics among themselves. For example, the shape of the stomach varies greatly even in the same individual at different times, making precise pixel segmentation extremely challenging. Accurate and automatic segmentation of readable results from abdominal multiple organs can provide accurate evidence of reality for surgical navigation, visual enhancement, radiation therapy, and biomarker measurement systems. Therefore, how to accurately make the segmentation results more readable in the case of multiple organs influencing each other has a great contribution to clinical examination and diagnosis.

Abdominal multi-organ network models based on deep neural networks (DNN) are difficult to train. Training such a good enough model usually requires a large amount of labeled data, or the model performance is likely to meet a heavy drop. However, manual annotation of organs requires doctors to make accurate judgments based on their professional knowledge and rich experience, this leads to making manual labeling both expensive and time-consuming. In addition to pixel-level annotated datasets, deep neural networks can also benefit from other types of supervision. For example, boundary-level annotation can provide more detailed boundary information. In addition, weakly supervised [6,12,14] learning techniques can be used, such as training with pixel-level labels and unlabeled data. Additionally, visual perceptual [1,7] supervision can be employed by utilizing visual perceptual theory in the training of deep networks to increase their sensitivity to image features. Furthermore, pre-trained models can be utilized for transfer learning, which allows the model to learn features from previous tasks and improve its performance. In summary, deep neural networks can benefit from various types of supervision, which can improve their performance in a variety of visual tasks. These studies have demonstrated that incorporating finer-grained additional supervision can enhance the accuracy of deep neural networks and improve the interpretability of network models.

However, the practical process of collecting additional annotations remains challenging, as it may require clinicians to repeatedly provide specific and refined annotations to fine-tune the network model. There is a need to minimize the impact of the annotation process on clinical work. To address this, we investigate novel annotation information that can be used for abdominal multi-organ segmentation. In the context of medical image analysis, it has been observed that radiologists tend to focus their attention on specific regions of interest (ROIs) or lesions when interpreting medical images. Specifically, our method utilizes eye gaze information collected by an eye-tracker during radiologists' image interpretation as a source of additional supervision. In clinical practice, experienced radiologists can usually quickly locate specific organs when reading abdominal images. In this process, the doctor's eye movement information can reflect the location information of organs to a certain extent. Compared with manual label-

ing, this information is cheap and fast and can be used as effective supervision information to assist the localization and segmentation of each organ. The literature studies have implied that the potential of the radiologist's gaze data can be high in improving disease diagnosis [2,17]. Recently, Wang et al. [16] applied eye-tracking technology to diagnose knee osteoarthritis, while Men et al. [9] used eye-trackers to provide visual guidance to sonographers during ultrasound scanning. It can be seen that the use of eye movement attention information has great value and potential in automated auxiliary diagnosis.

In this paper, we propose a novel eye-guided multi-organ segmentation network for diverse abdominal organ images. The network model is forced to focus on relevant objects or features required for the segmentation task by fully and synergistically utilizing the radiologist's cognitive information about the abdominal image. This method of information collection is convenient and can make the positioning of each organ more accurate. The overall architecture is shown in Fig. 1. The proposed network has three special designs: 1) a dual-path encoder that integrates human cognitive information; 2) a cross-attention transformer module (CATM) that communicates information in network semantic perception and human semantic perception; and 3) multi-feature skip connection (MSC), which effectively combines spatial information during down-sampling to offset the internal details of segmentation.

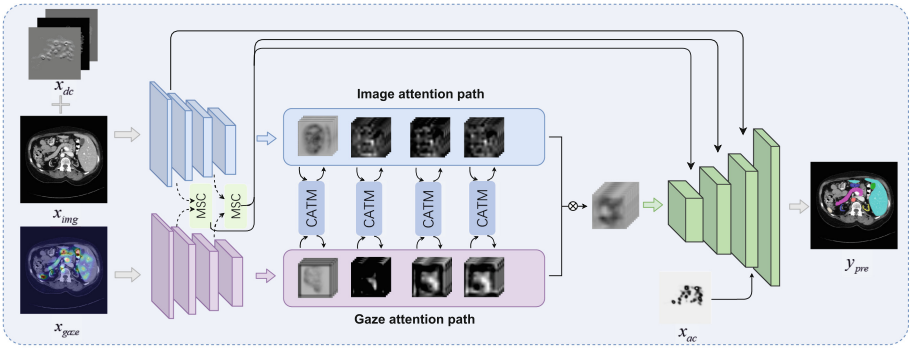

Fig. 1. Overview of the network architecture, detail coefficients x_{dc}, image x_{img} and gaze information x_{gaze} are input into the network for segmentation. In the decoding phase, approximation coefficients x_{ac} are fused to compensate the global information. MSC: multi-feature skip connection. CATM: cross-attention transformer module

2 Methodology

As shown in Fig. 1. The proposed network adopts an encoder-decoder structure, where the encoder part consists of parallel dual paths that utilize multi-feature skip connection (MSC) to combine spatial information during down-sampling to offset the internal details of segmentation. A cross-attention transformer module

(CATM) is designed at the bottleneck stage to effectively communicate information in network perception and human perception.

2.1 Wavelet Transform for Composite Information

Wavelet transform is able to obtain global information and edge information in different directions in gaze attention heatmaps so that the network can effectively fuse the composite information in the heatmaps. In the clinic, when radiologists read abdominal images, the more important location, the longer the radiologists' gaze. We convert this information into a heatmap representation. The heatmap reflects the rough position information of the target to be segmented. The single heatmap is unable to reflect the composite information it contains, therefore DWT is utilized for extracting it. Discrete wavelet transform [8] (DWT) is applied to decompose the approximation coefficients and detail coefficients of abdominal organ distribution information on the gaze attention heatmap to locate the position and edge of multiple organs in the abdomen. In the decoding phase, we fuse the approximation coefficient in the gaze heatmap so that compensates for the global topological information of decoding features at the final segmentation. The detail coefficients are input into the image encoder together with the original image, which is used to guide the dual-path encoder to reserve detailed information.

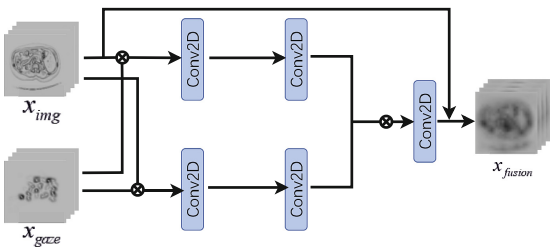

Fig. 2. MSC fuses the encoding features on the two paths concatenated with features in the up-sampling process to offset the internal details of segmentation.

2.2 Multi-feature Skip Connection

Multi-feature skip connection (MSC) comprehensively utilizes multiple features to guide the segmentation results of each abdominal organ toward accurate internal details. As shown in Fig. 2, we choose to integrate the encoding features on the two paths concatenated with features in the up-sampling process to offset the internal details of segmentation. Instead of using a simple concatenated and fusion strategy, we use multiple composite splicing and fusion to obtain matching features. The residual connection can aggregate the features of different levels

to avoid additional noise and thus improve the network performance. The MSC can be expressed as follows:

$$F_{feat} = F_i + Conv(Conv(Conv(F_i \oplus F_g)) \oplus Conv(Conv(F_g \oplus F_i))), \quad (1)$$

where F_i and F_g represent the output features from the down-sampling layers of the two paths, and F_{fusion} denotes the final multiple fusion features. Following the MSC, the dimension of the concatenated multiple features remains the same as the dimension of the upsampled features.

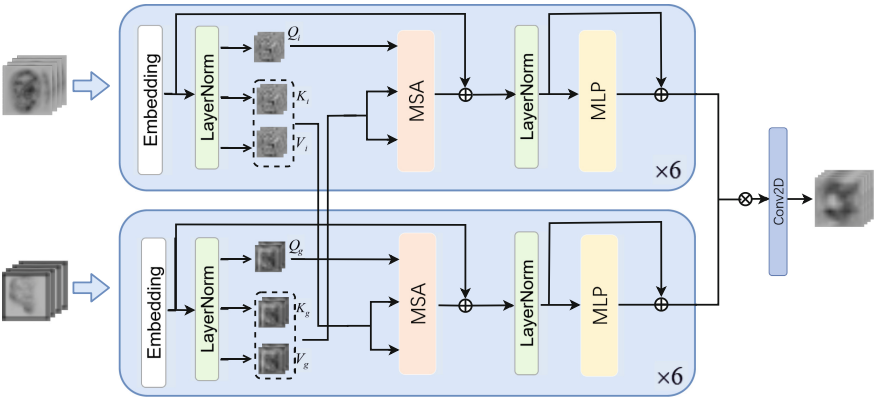

Fig. 3. CATM utilizes cross-attention efficiently enhances information interactive collaboration between network semantic perception and human semantic perception. The convolution operation fuses the feature information from two different paths.

2.3 Cross-Attention Transformer Module

The cross-attention transformer module (CATM) creatively enables the communication between network semantic perception and human semantic perception. Different from the traditional self-attention mechanism in transformer block [15], by using CATM, information interactive collaboration on two paths is enabled effectively. CATM, which is a multi-path structure, is embedded in the bottleneck layer between the encoder and decoder. As shown in Fig. 3, it consists of two paths: the image attention path and the gaze attention path. In our work, CATM is composed of L (we set $L = 6$) cross-attention transformer blocks (CTB) and Conv2D. The expression of CATM can be represented as:

$$F_{out} = Conv((CTB(F_i) \oplus CTB(F_g))_i), \quad (2)$$

where F_i and F_g represent the final output features of the encoder on the image attention and gaze attention encoding pathways. Our experimental results demonstrate that the cross-attention operation within the CATM design efficiently enhances the communication of information between these two paths.

The convolution operation of CATM is used to fuse the feature information from two paths, which makes up for the possible information shortage in the decoding process. The CTB is the core design of the CATM, which is a variant of the Transformer [15] that exchanges the network semantic perception and human semantic perception on the two different paths of the image and gaze attention. The key of cross-attention is to exchange Q, K and V of respective features between different path features and fuse them. As shown in Fig. 3, the K and V in the image attention path exchange with the gaze attention path, and each original Q fuse with the exchanged K and V. It is represented by a formula:

$$Attention(Q_i, K_g, V_g) = Softmax(Q_i K_g/\sqrt{d} + B)V_g, \tag{3}$$

$$Attention(Q_g, K_i, V_i) = Softmax(Q_g K_i/\sqrt{d} + B)V_i, \tag{4}$$

Q_i, K_i, V_i and Q_g, K_g, V_g represent Q, K, and V in the image and gaze attention path, respectively; B denotes the learnable relative positional encoding; d is the dimension of K, and we set the number of head of multi-headed self-attention is 12.

3 Experiments

3.1 Datasets and Evaluation

Our experiments use the Synapse multi-organ segmentation dataset(Synapse). Each CT volume consists of $85-198$ slices of 512×512 pixels, with a voxel spatial resolution of $([0.54-0.54] \times [0.98-0.98] \times [2.5-5.0])\,\mathrm{mm}^3$. We use the 30 abdominal CT scans and split it 18 training cases and 12 testing cases randomly. Following [3,4], all 3D volumes are inferenced in a slice-by-slice fashion and the predicted 2D slices are stacked together to reconstruct the 3D prediction. We use the average Dice-Similarity coefficient(DSC) and average Hausdorff distance (HD) as the evaluation metric to evaluate our method on the full resolution of the original slice.

Table 1. Comparison on the Synapse multi-organ CT dataset (average dice score % and average Hausdorff distance in mm, and dice score % for each organ).

Methods	DSC↑	HD↓	Aorta	Gallbladder	Kidney(L)	Kidney(R)	Liver	Pancreas	Spleen	Stomach
V-Net [10]	68.81	-	75.34	51.87	77.10	80.75	87.84	40.05	80.56	56.98
DARR [5]	69.77	-	74.74	53.77	72.31	73.24	94.08	54.18	89.90	45.96
R50 U-Net [4]	74.68	36.87	87.74	63.66	80.60	78.19	93.74	56.90	85.87	74.16
U-Net [13]	76.85	39.70	89.07	**69.72**	77.77	68.60	93.43	53.98	86.67	75.58
U-Net$_{gaze}$	77.94	35.50	89.66	65.04	81.25	75.91	93.57	59.94	84.66	73.53
R50 Att-UNet [4]	75.57	36.97	55.92	63.91	79.20	72.71	93.56	49.37	87.19	74.95
Att-UNet [11]	77.77	36.02	89.55	68.88	77.98	71.11	93.57	58.04	87.30	75.75
R50 ViT [4]	71.29	32.87	73.73	55.13	75.80	72.20	91.51	45.99	81.99	73.95
TransUnet [4]	77.48	31.69	87.23	63.13	81.87	77.02	94.08	55.86	85.08	75.62
TransUnet$_{gaze}$	78.29	27.00	88.57	61.89	83.07	75.99	94.56	59.44	85.54	77.27
SwinUnet [3]	79.13	21.55	85.47	66.53	83.28	79.61	94.29	56.58	**90.66**	76.60
SwinUnet$_{gaze}$	80.02	19.47	86.58	66.97	84.05	81.80	94.22	61.33	89.22	76.00
Ours	**81.87**	**11.96**	**89.88**	64.16	**86.62**	**83.89**	**94.77**	**66.97**	87.53	**81.16**

3.2 Results and Analysis

Overall Performance. As the last row shown in Table 1. Our experimental results demonstrate that leveraging gaze attention as an auxiliary supervision mechanism for network training achieves superior segmentation performance, as evidenced by segmentation accuracies of 81.87% (Dice similarity coefficient) and 11.96% (Hausdorff distance).

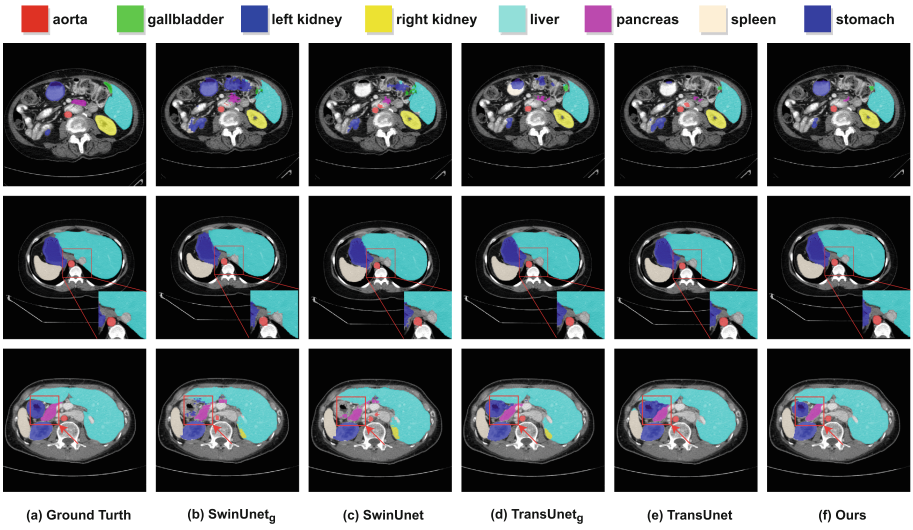

Fig. 4. 1). 1st row of the figure highlights that the approach leveraging gaze attention as auxiliary supervision yields fewer erroneous labels in comparison to the other methods; 2). TransUnet and SwinUnet (without gaze attention information) predict coarser edges and shapes compared to our method; 3). In the 3rd row, our method correctly identifies the stomach, while SwinUnet (with and without gaze attention information) failed to predict the shape of the stomach.

Comparison with Existing Methods Performance. As shown in Table 1, We also train Unet$_{gaze}$, TransUnet$_{gaze}$, and SwinUnet$_{gaze}$ networks with gaze attention information by simply concatenating the gaze attention information with the input image. Our experimental results reveal that this approach of introducing gaze attention information as an auxiliary supervision mechanism leads to an appreciable improvement in network segmentation performance. Specifically, Unet, TransUnet, and SwinUnet have an improvement of approximately 1% in terms of the DSC evaluation metric and 2–4% in terms of the HD evaluation metric by concatenating the gaze attention. Furthermore, our method surpasses the SwinUnet (without gaze attention information) by approximately 3% in terms of the DSC evaluation metric and approximately 10% in terms of the HD evaluation metric. Comparing our method with U-Net, TransUnet,

and SwinUnet (with gaze attention information), we also observe a significant improvement of approximately 2–4% in terms of the DSC evaluation metric and 7–24% in terms of the HD evaluation metric, respectively.

Qualitative Visualization Results. As shown in Fig. 4. 1st row of the figure highlights that the approach leveraging gaze attention as auxiliary supervision yields fewer erroneous labels in comparison to the other methods, implying that eye-tracking attention can aid the model in attending to relevant objects or features. TransUnet (with gaze attention information) predicts coarser edges and shapes compared to our method (e.g. in the 2nd row, the model's prediction for the liver). In the 3rd row, our method correctly identifies the stomach, while SwinUnet (with and without gaze attention information) failed to predict the shape of the stomach. The results demonstrate that our method can leverage gaze attention as auxiliary supervision and better segment and retain edge and shape information.

Table 2. Ablation study on the different variant network under the Synapse multi-organ CT dataset (average dice score % and average Hausdorff distance in mm).

Variant	Modules			Metrics	
	DWT	MSC	CATM	DSC	HD
w/o DWT	×	✓	✓	75.75	32.35
w/o MSC	✓	×	✓	81.21	8.99
w/o CATM	✓	✓	×	80.46	12.34
Full Version	✓	✓	✓	81.87	11.96

Ablation Study. We verified the DWT, MSC, and CATM separately using three different network configurations. We summarize the experimental results in Table 2. It can be seen that the w/o DWT performs the worst, indicating that the detail coefficients extracted from eye-tracking heatmaps can effectively locate organs in the image and provide strong support for edge segmentation. The w/o CATM does not effectively fuse the image features and eye-tracking attention features using CATM, resulting in a less-than-ideal improvement in segmentation results. In the results of w/o MSC and the full version of the network, we observed that MSC can further improve segmentation results by integrating down-sampling information from both paths.

4 Conclusion

In this paper, we propose a novel network that can realize the interactive communication between network semantic perception and human semantic perception, and apply it to the task of abdominal multi-organ segmentation for information

interactive collaboration. The network is innovatively built with 1) a dual-path encoder that integrates human cognitive information; 2) a cross-attention transformer module (CATM) that communicates information in network semantic perception and human semantic perception; and 3) multi-feature skip connection (MSC), which effectively combines spatial information during down-sampling to offset the internal details of segmentation. Extensive experiments with promising results reveal gaze attention has great clinical value and potential in multi-organ segmentation.

Acknowledgements. This study was supported by the National Natural Science Foundation (No. 62101249 and No. 62136004), the Natural Science Foundation of Jiangsu Province (No. BK20210291), and the China Postdoctoral Science Foundation (No. 2021TQ0149 and No. 2022M721611).

References

1. Bertram, R., et al.: Eye movements of radiologists reflect expertise in CT study interpretation: a potential tool to measure resident development. Radiology **281**(3), 805–815 (2016)
2. Brunyé, T.T., Drew, T., Weaver, D.L., Elmore, J.G.: A review of eye tracking for understanding and improving diagnostic interpretation. Cogn. Res. Princ. Implic. **4**(1), 1–16 (2019). https://doi.org/10.1186/s41235-019-0159-2
3. Cao, H., et al.: Swin-Unet: Unet-like pure transformer for medical image segmentation. In: Karlinsky, L., Michaeli, T., Nishino, K. (eds.) ECCV 2022. LNCS, vol. 13803, pp. 205–218. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-25066-8_9
4. Chen, J., et al.: Transunet: transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
5. Fu, S., et al.: Domain adaptive relational reasoning for 3D multi-organ segmentation. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12261, pp. 656–666. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59710-8_64
6. Herzig, J., Nowak, P.K., Müller, T., Piccinno, F., Eisenschlos, J.M.: Tapas: weakly supervised table parsing via pre-training. arXiv preprint arXiv:2004.02349 (2020)
7. Kundel, H.L., Nodine, C.F., Krupinski, E.A., Mello-Thoms, C.: Using gaze-tracking data and mixture distribution analysis to support a holistic model for the detection of cancers on mammograms. Acad. Radiol. **15**(7), 881–886 (2008)
8. Li, G., Lyu, J., Wang, C., Dou, Q., Qin, J.: WavTrans: synergizing wavelet and cross-attention transformer for multi-contrast MRI super-resolution. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022. LNCS, vol. 13436, pp. 463–473. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16446-0_44
9. Men, Q., Teng, C., Drukker, L., Papageorghiou, A.T., Noble, J.A.: Multimodal-guidenet: gaze-probe bidirectional guidance in obstetric ultrasound scanning. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022. LNCS, vol. 13437, pp. 94–103. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16449-1_10
10. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV), pp. 565–571. IEEE (2016)

11. Oktay, O., et al.: Attention U-net: learning where to look for the pancreas. arXiv preprint arXiv:1804.03999 (2018)
12. Ouyang, X., et al.: Learning hierarchical attention for weakly-supervised chest X-ray abnormality localization and diagnosis. IEEE Trans. Med. Imaging **40**(10), 2698–2710 (2020)
13. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
14. Shu, R., Chen, Y., Kumar, A., Ermon, S., Poole, B.: Weakly supervised disentanglement with guarantees. arXiv preprint arXiv:1910.09772 (2019)
15. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
16. Wang, S., Ouyang, X., Liu, T., Wang, Q., Shen, D.: Follow my eye: using gaze to supervise computer-aided diagnosis. IEEE Trans. Med. Imaging **41**(7), 1688–1698 (2022)
17. Wu, C.C., Wolfe, J.M.: Eye movements in medical image perception: a selective review of past, present and future. Vision **3**(2), 32 (2019)