



This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Weakly-supervised Medical Image Segmentation with Gaze Annotations

Yuan Zhong¹, Chenhui Tang¹, Yumeng Yang², Ruoxi Qi², Kang Zhou¹, Yuqi Gong¹, Pheng Ann Heng¹, Janet H. Hsiao³(✉), and Qi Dou¹(✉)

¹ Dept. of Computer Science and Engineering, The Chinese University of Hong Kong

² Dept. of Psychology, The University of Hong Kong

³ Division of Social Science, The Hong Kong University of Science and Technology

Abstract. Eye gaze that reveals human observational patterns has increasingly been incorporated into solutions for vision tasks. Despite recent explorations on leveraging gaze to aid deep networks, few studies exploit gaze as an efficient annotation approach for medical image segmentation which typically entails heavy annotating costs. In this paper, we propose to collect dense weak supervision for medical image segmentation with a gaze annotation scheme. To train with gaze, we propose a multi-level framework that trains multiple networks from discriminative human attention, simulated with a set of pseudo-masks derived by applying hierarchical thresholds on gaze heatmaps. Furthermore, to mitigate gaze noise, a cross-level consistency is exploited to regularize overfitting noisy labels, steering models toward clean patterns learned by peer networks. The proposed method is validated on two public medical datasets of polyp and prostate segmentation tasks. We contribute a high-quality gaze dataset entitled **GazeMedSeg** as an extension to the popular medical segmentation datasets. To the best of our knowledge, this is the first gaze dataset for medical image segmentation. Our experiments demonstrate that gaze annotation outperforms previous label-efficient annotation schemes in terms of both performance and annotation time. Our collected gaze data and code are available at: <https://github.com/med-air/GazeMedSeg>.

Keywords: Gaze Annotation · Weakly-supervised Image Segmentation

1 Introduction

Recent studies have witnessed increasing interest in incorporating human factors into deep learning applications [13,24]. Eye tracking data, serving as a popular tool reflecting the underlying cognitive processes [27], has stood out as a promising and accessible media for human-AI interaction. Previous works commonly utilize gaze as auxiliary information to guide deep networks [9,15,22,23], with recent explorations of employing gaze as the sole supervision signal for label-efficient classification [19]. However, leveraging gaze for supervising image segmentation models remains under-explored yet valuable, since it alleviates

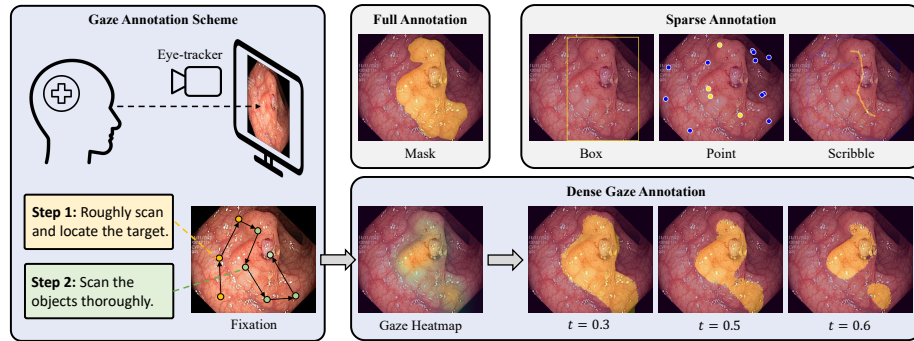


Fig. 1. Illustrations of full and different label-efficient annotation schemes. Dense binarized gaze pseudo-masks are generated with various thresholds t , which trade off the activation of the foreground and background.

annotators’ workload by alleviating the need for labor-intensive pixel-wise annotation. Unlike existing label-efficient annotation schemes that provide sparse supervision with bounding boxes [6,20], points [4] or scribbles [25], gaze data yield dense pixel-wise supervision signals, which is crucial for medical images featuring ambiguous boundaries and low contrast. These motivate us to investigate the potential of gaze-supervised medical image segmentation.

A straightforward way to gaze supervision is training with pseudo-masks generated by binarizing gaze heatmaps with a fixed threshold, where the dense gaze heatmaps contain continuous values indicating the degree of observational attention. A similar approach is widely employed in image-level semantic segmentation [3,26] for binarizing class activation maps (CAMs) [29]. In practice, this approach yields suboptimal performance for gaze supervision due to the distinct characteristics of gaze data as *discriminative* and *noisy*. Firstly, the quality of pseudo-masks is sensitive to the selection of threshold (Fig. 1). And it is unreasonable to give precise object boundaries with a global threshold over all images since human annotators usually pay discriminative attention even on different parts of a single object. Secondly, the error in eye tracking and human subjectivity makes gaze data a noisy supervision signal for segmentation. For example, the annotator may check every suspicious area when annotating the targets, thus some noisy gaze will be left. Current noise-robust approaches are based on the symmetric or asymmetric assumptions of simulated noise and design robust loss functions [8,28] or regularization [14]. For correlated real-world gaze noise, however, the assumptions on simulated noise do not necessarily hold.

The key to robust gaze supervision lies in the unity and consideration of the aforementioned two characteristics of gaze data. Inspired by multi-expert models [17] benefiting from comprehensively integrating knowledge from multiple experts, we propose to fuse multiple diverged networks learning from multi-level human attention, simulated by applying a set of hierarchical thresholds on gaze heatmaps. These networks are designed to learn heterogeneous knowledge from

discriminative human attention. Moreover, to mitigate gaze noise, we exploit the clean knowledge learned by peer networks of different levels to compensate for overfitting gaze noise with analysis on the memorization effect of deep networks.

In this paper, we propose a new gaze annotation scheme that collects dense annotation in an annotator-friendly and efficient manner for segmentation tasks. Utilizing the scheme, we introduce the gaze dataset **GazeMedSeg**, which extends the Kvasir-SEG [10] and NCI-ISBI [2] datasets with gaze data of multiple annotators. To train with gaze, we propose a multi-level approach that trains multiple divergent deep networks to ensemble information from different levels of human discriminative attention. In addition, a cross-level consistency regularization term over predictions smoothed by a local pixel propagation module is exploited to compensate for overfitting on noisy gaze labels. The advantage of the proposed neat approach is in its ability to seamlessly fit into standard training pipelines with no changes to model architectures. In experiments, we validate gaze annotation on polyp and prostate segmentation tasks using our GazeMedSeg dataset. Compared to the existing label-efficient annotation schemes, gaze supervision narrows the gaps with full supervision and consistently boosts performance by over 2.0% in Dice while being 15.4% faster to annotate, striking a sweet trade-off between performance and annotation time.

2 Gaze Annotation Collection

Gaze annotation scheme. We develop the eye-tracking program utilizing SR Research Experiment Builder platform. At the beginning of gaze annotation, each annotator goes through a 9-point gaze calibration process. Our gaze annotation collection consists of two stages. When presented with an image, the annotator (with eye-tracker) first roughly scans the image and locates the target objects. Following that, the annotator is requested to scan the objects thoroughly. Typically, participants start from central areas and then move on to the boundaries, ensuring that all parts of the target are covered. This step avoids partial activation by encouraging annotators to pay more attention to the target objects. Therefore, the noise will be relatively weakened when normalizing the heatmap. After finishing the annotating, a key is pressed to switch to the next image. More details on the eye-tracking settings can be found in the Appendix.

GazeMedSeg dataset. Our collected GazeMedSeg dataset includes gaze annotations for two public medical segmentation datasets. We use the Kvasir-SEG [10] dataset for polyp segmentation from gastrointestinal images, and the NCI-ISBI [2] dataset for prostate segmentation from T2-weighted MR images. The Kvasir-SEG dataset includes 900 training and 100 testing images and the NCI-ISBI dataset includes 60 training and 10 testing volumes, where we retain slices containing prostate and obtain 789 training and 117 testing images. One annotator finishes the annotation of all images in the datasets, and we use it in our major experiments. We also invite two additional annotators to annotate a subset for sensitivity studies (Sec. 4.2). All annotators are experienced in medical imaging and are well-trained for eye-tracking trials.

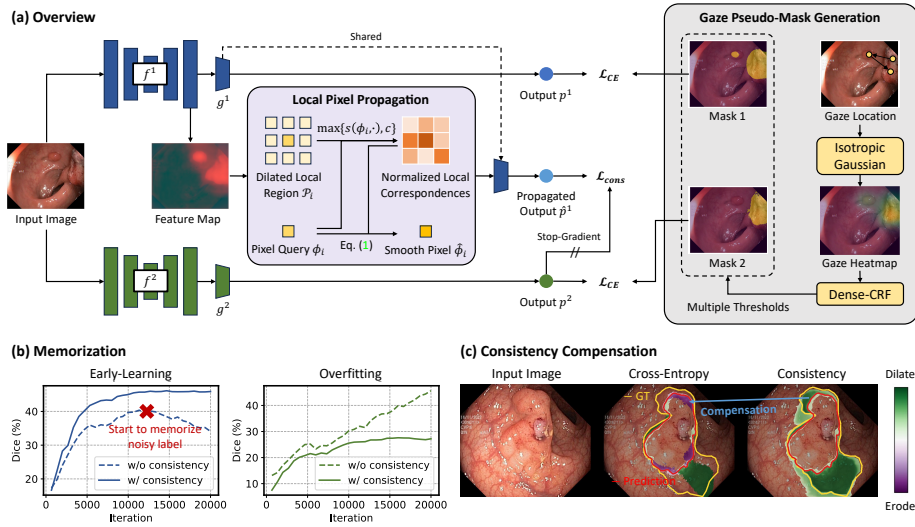


Fig. 2. (a) Overview of the proposed method. For simplicity, we present the case with two levels and \mathcal{L}_{cons} of network 1. The consistency loss is applied to all networks in the implementation. (b) We visualize the dynamics of *early-learning* (the Dice of output and ground-truth on wrongly annotated pixels) and *overfitting* (the Dice of output and noisy gaze pseudo-mask on wrongly annotated pixels) with and without the proposed consistency regularization on Kvasir-SEG [10] training data. The proposed consistency prevents overfitting on the noisy labels. We use two levels and plot the average of all levels in this experiment. (c) We visualize the gradients of cross-entropy and consistency terms in the training process. The gradients that encourage dilation and erosion in different colors. The cross-entropy term predicted target are scaled for visualization on the top of the target object, which is compensated by consistency with clean patterns of dilation learned by other networks.

3 Methodology

3.1 Multi-level Learning from Discriminative Attention

The original gaze data is a series of gaze positions collected at a certain frequency. Given the gaze positions, the attention heatmaps are obtained by convolving an isotropic Gaussian over them. We further apply dense conditional random fields (CRF) [11] to enhance the initial heatmaps and generate pseudo-masks. However, the quality of pseudo-masks varies significantly with distinct thresholds (Fig. 1). It is hard to balance the over-activation and under-activation of foregrounds via a fixed global threshold, because human subjectively pay discriminative attention to target objects. For instance, annotators may focus on the most discriminative part while only roughly scanning ambiguous parts of an object, which may be neglected with a fixed global threshold.

Our idea is to train m deep networks simultaneously supervised by pseudo-masks generated from m different thresholds (Fig. 2 (a)), simulating multi-level

human attention. Each network is independently initialized, resulting in varying learning capabilities. Being supervised by pseudo-masks with different activation degrees, the networks evolve to learn various representations of human attention.

In practice, we select a pair of thresholds based on annotators’ feedback to generate diverse heatmaps that closely resemble ground truth and complement each other, with one tending to erode (under-activate) and the other dilate (over-activate) targets. Additional thresholds are linearly interpolated from the pair. Empirical results indicate that two levels are sufficient for decent performance (Sec. 4.2). The final segmentation prediction is obtained by ensembling the predictions of these networks. We maintain the multi-level structure throughout both the training and inference stages.

3.2 Cross-level Consistency for Noise Compensation

Another essential aspect of gaze-supervised segmentation is the inevitable noise in the pseudo-masks. The noises can be introduced in the process of human interpretation, gaze estimation by eye-trackers, heatmap generation, etc. We observe a memorization effect [1] of deep networks when training with gaze data on the medical image segmentation task. As shown in Fig. 2 (b), the model captures clean patterns on incorrectly annotated pixels of pseudo-masks at the beginning of training, but eventually overfits on noisy labels.

In the multi-level framework, though the networks are diverged with distinct supervisions, they share the same input and various pseudo-masks supervise different representations of the shared gaze data. Based on the assumption that networks learn clean patterns at the beginning of the memorization, for each network, we propose to exploit the knowledge learned by peer networks of other levels to compensate for the noisy label via a consistency term.

To ensure noise-robust consistency, we first use a non-parametric local pixel propagation (LPP) module to filter the feature of each pixel by propagating the features of surrounding pixels in local regions inspired by recent works [5,7] proving noise-robust feature correspondence distillation. Given the feature map ϕ , for each pixel feature ϕ_p , the LPP module computes the transform $\widehat{\phi}_p$ as:

$$\widehat{\phi}_p := \sum_{q \in \mathcal{P}_p} \text{softmax}(\max\{\cos(\phi_p, \phi_q), 0\}) \cdot \phi_q, \quad (1)$$

where \cos denotes the cosine similarity, \mathcal{P}_p denotes the set of neighboring pixels (e.g., a 3×3 region with dilation 1) of pixel p . This refinement has a feature denoising/smoothing effect that reduces the outliers and enhances the features with local context by introducing spatial smoothness which encourages spatially close pixels to be similar. Given an arbitrary pair of levels $i, j \in \mathbb{Z}^+$, where $i, j \leq m$ and $i \neq j$, the consistency loss applied on the i -th level maximizes dot-product between the propagated prediction \widehat{p}^i of i -th level and non-propagated prediction p^j of j -th level, i.e., $\mathcal{L}_{\text{cons}}^{(i,j)} := -\widehat{p}^i \cdot p^j$. Notably, we have $\widehat{p}^i = g^i(\widehat{\phi}^i)$ for the propagated prediction and $p^i = g^i(\phi^i)$ for the non-propagated one, where g denotes the shallow classifier.

3.3 Overall Optimization of Gaze Supervision

The overall loss for the i -th level is the combination of the supervision term and the consistency term over all other peer-level j :

$$\mathcal{L}^i = \mathcal{L}_{\text{CE}}^i + \frac{\lambda}{m-1} \sum_{j=1, j \neq i}^m \mathcal{L}_{\text{cons}}^{(i,j)}, \quad (2)$$

where \mathcal{L}_{CE} is the cross-entropy loss and can be replaced by any other segmentation loss such as dice loss, and λ is empirically set to 3. Note that when optimizing the network of the i -th level, the parameters of all other networks are frozen.

The key to understanding the overall optimization process lies in the trade-off of cross-entropy and consistency terms. Intuitively, \mathcal{L}_{CE} trains a set of divergent networks utilizing multi-level pseudo-masks. However, this term tends to vanish after the early learning stage and each network starts to overfit on the respective noisy label. The consistency term $\mathcal{L}_{\text{cons}}$ compensates for it, implicitly forcing networks to continue learning from clean patterns learned by networks of other levels. The mechanism can be viewed as pushing networks to struggle to find a balance between divergence and consistency, in which the hyper-parameter λ controls this balance. It is worth noting that both divergence and consistency are equally essential for optimization. The consistency term ensures robustness to noise and the cross-entropy term expands and diversifies clean knowledge learned and prevents collapsing into a single network.

4 Experiments

We validate the proposed gaze annotation scheme on the aforementioned Kvasir-SEG [10] dataset for polyp segmentation and NCI-ISBI [2] dataset for prostate segmentation. For all datasets, we only utilize weak annotations for training and report performance on the testing set. We train a 2D UNet [18] from scratch for 15k iterations with a NVIDIA A40 GPU. We use Adam optimizer with batch size 8 and learning rate $4e^{-4}$. More details on the training recipe can be found in the Appendix.

4.1 Comparison Among Label-efficient Annotation Schemes

Comparison with state-of-the-art weakly-supervised methods. We compare the new gaze-based annotation scheme with full mask supervision and other state-of-the-art weakly-supervised methods using different sparse annotations including box, point, and scribble on two datasets. To compare annotation time, we also invite the same annotator to annotate a randomly sampled subset of the Kvasir-SEG dataset containing 100 images using other annotation schemes and report the annotation time in Table 1. Note that we annotate the bounding box using extreme points clicking [16], and annotate points using the scheme suggested by [4]. The estimated time is close to that reported in the literature [4,12,16,21], where the narrow gaps may come from the difference in the

Table 1. Comparison with full mask supervision and SOTA weakly-supervised methods using different annotation schemes. We report the mean and standard deviation of three runs with different seeds. Dice is used as the evaluation metric. The reported annotation time is estimated to annotate 900 images in Kvasir-SEG [10] training set.

Method	Sup.	Polyp		Prostate
		Anno. Time	Dice	Dice
Vanilla	Full	18.7 hrs	82.12 \pm 1.11	80.58 \pm 0.48
BoxInst [20]	Box	3.1 hrs	65.72 \pm 2.97	73.78 \pm 1.15
BoxTeacher [6]	Box	3.1 hrs	73.33 \pm 1.30	75.60 \pm 1.15
PointSup [4]	Point	4.8 hrs	73.05 \pm 1.64	73.46 \pm 4.71
AGMM [25]	Point	4.8 hrs	75.57 \pm 0.84	73.86 \pm 1.26
AGMM [25]	Scribble	2.6 hrs	67.23 \pm 1.02	72.70 \pm 1.03
Ours	Gaze	2.2 hrs	77.80\pm1.02	77.64\pm0.57

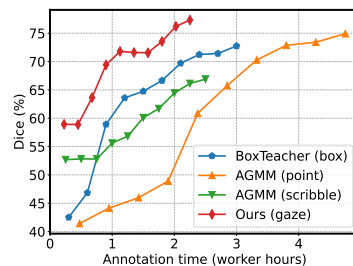


Fig. 3. Performance versus annotation time for different annotation schemes. To match annotation times among annotation forms, we train a 2D UNet model using from 10% to 100% of the Kvasir-SEG training set.

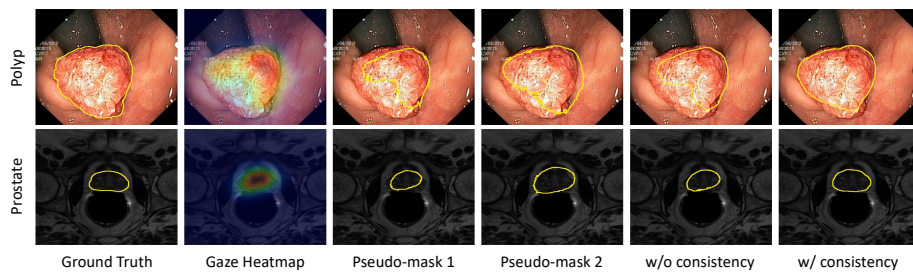


Fig. 4. Visualization of gaze data and predictions. The model without consistency term ensemble the noise of different levels. Instead, the model regularized by consistency learns clean patterns of pseudo-masks and demonstrates robustness to noises.

complexity of different target objects to be annotated. For all datasets, we simulate weak annotations based on the ground truth. Note that we randomly sampled 10 pixels and 10 background pixels inside and outside the bounding box respectively as suggested by [4] for point annotations, and we follow [21] to simulate scribble annotations. In Table 1, our results show that gaze supervision outperforms previous weakly-supervised methods trained with other sparse annotation schemes and achieves over 95% of the fully-supervised performance.

Trade-off between performance and annotation time. We compare the proposed gaze annotation scheme with other label-efficient sparse annotation schemes for image segmentation under the same annotation budget, *i.e.*, the time required to annotate training data. Fig. 3 presents our results on Kvasir-SEG [10], proving that gaze annotation boosts weakly-supervised segmentation by striking a sweet performance/annotation time trade-off by maximizing performance with the least annotation time among existing weak annotation schemes.

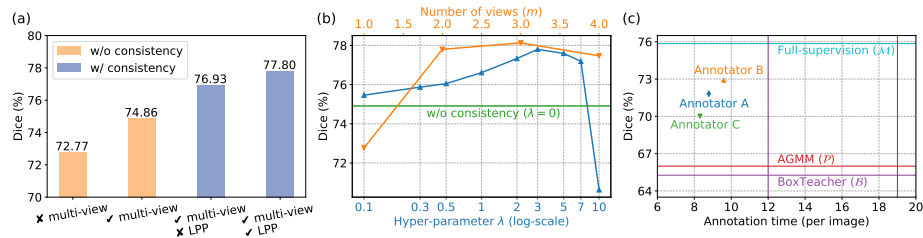


Fig. 5. Ablation studies on polyp segmentation. (a) Effects of proposed components. (b) Effects of the hyper-parameter m ($m = 2$ by default) and λ ($\lambda = 3$ by default). The orange and blue lines depict the results of varying m and λ , respectively. (c) Effects of three different annotators (major annotator A and additional annotators B and C) and comparison with other annotation schemes.

4.2 Ablation Studies

Modules and hyper-parameters. To evaluate the impact of critical components of the proposed method, we study the effectiveness of the proposed components for gaze supervision in Fig. 5 (a). We further present the result of different choices of hyper-parameter m (number of levels) and λ in Fig. 5 (b). Our experiments show that $m = 2$ is optimal for gaze training while increasing m gives limited benefits but leads to greater training and inference complexity. The results on λ echo the intuition of it in Sec. 3.3. We also observe that having λ greater than 7 results in a degenerated model that collapses to consistency with performance worse than simply ensembling without consistency regularization. We further visualize gaze-supervised predictions in Fig. 4, where the model with consistency regularization demonstrates resistance to gaze noise.

Sensitivity to annotator. We invite two additional annotators to annotate a subset of the Kvasir-SEG training set containing 500 images and train a UNet on this subset using different annotation schemes. Note that all annotators receive the same training for gaze annotating. The results presented in Fig. 5 (c) show that though eye-tracking is subjective, different annotators demonstrate comparable annotation time and supervision quality, consistently outperforming other annotation schemes.

5 Conclusion and Future Work

In this paper, we propose to train deep networks with gaze annotations efficiently collected using a new gaze annotation scheme for medical image segmentation. The proposed method can be seamlessly integrated into standard training pipelines. The results show that gaze annotation achieves a sweet performance and annotation time trade-off compared to other annotation forms.

Our explorations give rise to several potential directions for future work: (1) While we expect gaze supervision to be broadly applicable to medical applications, the multiplied complexities in the training and especially the inference

stage hamper real-time scenarios since multiple networks are maintained even though we have shown that $m = 2$ is sufficient. One potential direction is to aggregate networks at a certain frequency in the training and only keep an aggregated model for inference. (2) Though specialized hardware is required to collect gaze data, we foresee that eye-tracking will not pose a bottleneck for clinical practicality even at present with the advance of commercial VR/XR headsets with precise and affordable eye-tracking capabilities. (3) We focus on binary segmentation in this paper, and the extension to multiple cases is straightforward via annotating each class separately and deciding the label for each pixel as the class with the highest value in gaze heatmaps of different classes.

Acknowledgments. This work was supported by Hong Kong Research Grants Council (Project No. T45-401/22-N), and Science, Technology and Innovation Commission of Shenzhen Municipality (Project No. SGDX20220530111201008).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Arpit, D., Jastrzebski, S., Ballas, N., Krueger, D., Bengio, E., Kanwal, M.S., Maharaaj, T., Fischer, A., Courville, A., Bengio, Y., et al.: A closer look at memorization in deep networks. In: International conference on machine learning. pp. 233–242. PMLR (2017)
2. Bloch, B.N., Madabhushi, A., Huisman, H., Freymann, J., Kirby, J., Grauer, M., Enquobahrie, A., Jaffe, C., Clarke, L., Farahani, K.: Nci-isbi 2013 challenge: Automated segmentation of prostate structures (isbi-mr-prostate-2013) (2015). <https://doi.org/10.7937/K9/TCIA.2015.ZFOVLOPV>
3. Chang, Y.T., Wang, Q., Hung, W.C., Piramuthu, R., Tsai, Y.H., Yang, M.H.: Mixup-cam: Weakly-supervised semantic segmentation via uncertainty regularization. British Machine Vision Conference (2020)
4. Cheng, B., Parkhi, O., Kirillov, A.: Pointly-supervised instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2617–2626 (2022)
5. Cheng, H., Zhu, Z., Sun, X., Liu, Y.: Mitigating memorization of noisy labels via regularization between representations. International Conference on Learning Representations (2022)
6. Cheng, T., Wang, X., Chen, S., Zhang, Q., Liu, W.: Boxteacher: Exploring high-quality pseudo labels for weakly supervised instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3145–3154 (2023)
7. Hamilton, M., Zhang, Z., Hariharan, B., Snavely, N., Freeman, W.T.: Unsupervised semantic segmentation by distilling feature correspondences. International Conference on Learning Representations (2022)
8. Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., Tsang, I., Sugiyama, M.: Co-teaching: Robust training of deep neural networks with extremely noisy labels. Advances in neural information processing systems **31** (2018)
9. Huang, Y., Li, X., Yang, L., Gu, L., Zhu, Y., Seo, H., Meng, Q., Harada, T., Sato, Y.: Leveraging human selective attention for medical image analysis with limited training data. British Machine Vision Conference (2021)

10. Jha, D., Smedsrud, P.H., Riegler, M.A., Halvorsen, P., de Lange, T., Johansen, D., Johansen, H.D.: Kvasir-seg: A segmented polyp dataset. In: International Conference on Multimedia Modeling. pp. 451–462. Springer (2020)
11. Krähenbühl, P., Koltun, V.: Efficient inference in fully connected crfs with gaussian edge potentials. *Advances in neural information processing systems* **24** (2011)
12. Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Mallocci, M., Kolesnikov, A., et al.: The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision* **128**(7), 1956–1981 (2020)
13. Li, Q., Peng, Z., Zhou, B.: Efficient learning of safe driving policy via human-ai copilot optimization. *International Conference on Learning Representations* (2022)
14. Liu, S., Niles-Weed, J., Razavian, N., Fernandez-Granda, C.: Early-learning regularization prevents memorization of noisy labels. *Advances in neural information processing systems* **33**, 20331–20342 (2020)
15. Liu, Y., Zhou, L., Zhang, P., Bai, X., Gu, L., Yu, X., Zhou, J., Hancock, E.R.: Where to focus: Investigating hierarchical attention relationship for fine-grained visual classification. In: *European Conference on Computer Vision*. pp. 57–73. Springer (2022)
16. Papadopoulos, D.P., Uijlings, J.R., Keller, F., Ferrari, V.: Extreme clicking for efficient object annotation. In: *Proceedings of the IEEE international conference on computer vision*. pp. 4930–4939 (2017)
17. Pavlitskaya, S., Hubschneider, C., Weber, M., Moritz, R., Huger, F., Schlicht, P., Zollner, M.: Using mixture of expert models to gain insights into semantic segmentation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. pp. 342–343 (2020)
18. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III* 18. pp. 234–241. Springer (2015)
19. Saab, K., Hooper, S.M., Sohoni, N.S., Parmar, J., Pogatchnik, B., Wu, S., Dunnmon, J.A., Zhang, H.R., Rubin, D., Ré, C.: Observational supervision for medical image classification using gaze data. In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II* 24. pp. 603–614. Springer (2021)
20. Tian, Z., Shen, C., Wang, X., Chen, H.: Boxinst: High-performance instance segmentation with box annotations. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 5443–5452 (2021)
21. Valvano, G., Leo, A., Tsaftaris, S.A.: Learning to segment from scribbles using multi-scale adversarial attention gates. *IEEE Transactions on Medical Imaging* **40**(8), 1990–2001 (2021)
22. Wang, C., Zhang, D., Ge, R.: Eye-guided dual-path network for multi-organ segmentation of abdomen. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 23–32. Springer (2023)
23. Wang, S., Ouyang, X., Liu, T., Wang, Q., Shen, D.: Follow my eye: Using gaze to supervise computer-aided diagnosis. *IEEE Transactions on Medical Imaging* **41**(7), 1688–1698 (2022)
24. Wang, S., Zhao, Z., Ouyang, X., Wang, Q., Shen, D.: Chatcad: Interactive computer-aided diagnosis on medical image using large language models. *arXiv preprint arXiv:2302.07257* (2023)

25. Wu, L., Zhong, Z., Fang, L., He, X., Liu, Q., Ma, J., Chen, H.: Sparsely annotated semantic segmentation with adaptive gaussian mixtures. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15454–15464 (2023)
26. Wu, T., Gao, G., Huang, J., Wei, X., Wei, X., Liu, C.H.: Adaptive spatial-bce loss for weakly supervised semantic segmentation. In: European Conference on Computer Vision. pp. 199–216. Springer (2022)
27. Yun, K., Peng, Y., Samaras, D., Zelinsky, G.J., Berg, T.L.: Exploring the role of gaze behavior and object detection in scene understanding. *Frontiers in psychology* **4**, 917 (2013)
28. Zhang, Z., Sabuncu, M.: Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems* **31** (2018)
29. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2921–2929 (2016)