GazeRefine: Expert Gaze as a Test-Time Prompt for Training-Free Medical Image Segmentation

Mohammed Oussama Benyahia[1], Marouane Tliba[1], Mohamed Amine Kerkouri[1], Taifour Yousra[1], Bin Wang[2], Max Bengtsson[2], Gorkem Durak[2], Elif Keles[2], Zuheng Ming[1], Marek Penhaker[3], Azeddine Beghdadi[1], Ulas Bagci[2], Aladine Chetouani[1]

Université Sorbonne Paris Nord[1], Northwestern University[2], VSB-Technical University of Ostrava[3]

Abstract. Medical image segmentation remains difficult to scale because high-performing methods typically rely on dense expert annotations and task-specific training. We introduce **GazeRefine**, a training-free framework that uses gaze as an inference-time prompt for zero-shot medical image segmentation. Sparse, duration-weighted fixations are converted into foreground and background priors that initialize semantic prototypes in frozen DINOv3 feature space. These prototypes are iteratively refined through foreground–background discrimination, feature-space affinity propagation, and anchoring to the initial gaze guidance, allowing segmentation to extend beyond directly fixated regions while limiting semantic drift. GazeRefine requires no segmentation masks, fine-tuning, adapters, prompt encoders, or gradient updates. We evaluate the method on gaze-annotated polyp segmentation and prostate MRI segmentation. The results show strong performance on colonoscopy images and competitive performance on prostate MRI, supporting gaze-guided prototype refinement as a promising approach for segmentation-label-efficient, human-in-the-loop medical image segmentation. Our tools and code can be found in the following repository **GazeRefine**

Keywords: Medical image segmentation; zero-shot learning; gaze-guided segmentation; eye tracking; training-free inference

1 Introduction

Medical image segmentation is essential for computer-aided diagnosis, treatment planning, surgical guidance, and quantitative disease monitoring. Despite substantial progress from U-Net architectures to self-configuring pipelines such as nnU-Net, most high-performing methods still depend on dense pixel-level annotations and task-specific training [11,18,25,32]. This dependence remains a major limitation in medical imaging, where expert masks are costly, time-consuming, subject to inter-observer variability, and difficult to scale across organs, modalities, institutions, and rare pathologies [24]. Domain shifts, heterogeneous acquisition protocols, and privacy constraints further limit the repeated collection of annotations and retraining of specialized models for new clinical settings [9, 14].

Annotation-efficient segmentation has therefore become an important direction in medical image analysis. Weakly supervised approaches reduce mask requirements through points, scribbles, boxes, or pseudo-labels, but generally still require task-specific training or optimization [2,5,21,30]. Promptable foundation models such as SAM and MedSAM enable interactive segmentation through explicit visual prompts and segmentation-specific architectures [16,22]. In parallel, self-supervised vision transformers such as DINO, DINOv2, and DINOv3 provide dense patch representations with ***emergent localization and semantic grouping properties*** without pixel-level supervision [4,23,29]. These representations have motivated training-free segmentation approaches based on frozen features, similarity matching, and feature propagation [1, 7, 10, 20, 26, 35]. In the medical domain, Gaze2Segment [15] demonstrated that gaze-derived cues can support model-free segmentation by combining gaze heatmaps with image saliency. However, many recent interactive approaches still depend on explicit clicks or boxes, segmentation-specific prompt encoders, adapters, fine-tuning, or inference-time optimization [13].

In this context, clinician gaze provides a natural source of clinical guidance for medical image segmentation. During image interpretation, fixation locations and durations reflect task-relevant visual attention and can indicate suspicious structures or clinically meaningful regions. Recent methods have used gaze either as weak supervision for training segmentation models, as in GazeMedSeg and GradTrack [33,36], or as an input prompt to SAM-based segmentation models, as in GazeSAM and GazeMedSAMv2 [28, 31]. In contrast, we investigate whether clinician gaze can directly steer frozen self-supervised representations without task-specific training or a segmentation-specific architecture.

We introduce **GazeRefine**, a training-free framework that uses clinician gaze as an inference-time prompt for zero-shot medical image segmentation. GazeRefine converts sparse, duration-weighted fixations into foreground and background priors that initialize semantic prototypes in frozen DINOv3 feature space [29]. These prototypes are iteratively refined through foreground–background discrimination, feature-space affinity propagation, and anchoring to the initial gaze-derived representations. This formulation enables the predicted mask to extend beyond directly fixated regions while limiting drift from the original clinical guidance. The method requires no segmentation masks, task-specific fine-tuning, adapters, prompt encoders, or gradient updates. We evaluate GazeRefine on two publicly available gaze-annotated medical segmentation benchmarks with distinct imaging characteristics: polyp segmentation in colonoscopy images and prostate segmentation in magnetic resonance imaging. The results show that gaze can effectively guide frozen visual representations, particularly when target and background regions are well separated in feature space.

Our contributions are summarized as follows:

- We introduce **GazeRefine**, a training-free and segmentation-label-free framework that uses sparse clinician gaze to guide zero-shot medical image segmentation in frozen DINOv3 feature space.

– We combine gaze-derived foreground and background prototypes with recurrent refinement based on foreground–background discrimination and kNN affinity propagation, enabling segmentation beyond directly fixated regions.

2 Proposed Method

2.1 Overview

We introduce **GazeRefine**, a training-free and segmentation-label-free framework for medical image segmentation. GazeRefine combines clinician gaze with semantic representations extracted from a frozen DINOv3 model [29] to generate dense segmentation masks without task-specific model training. Given an image $I \in \mathbb{R}^{3\times H\times W}$ and a set of duration-weighted gaze fixations $\mathcal{G} = \{(x_m, y_m, d_m)\}_{m=1}^{M}$, where (x_m, y_m) denotes the fixation location and d_m its duration, the gaze information is projected onto the DINOv3 patch grid and used to construct foreground and background prototypes in the frozen feature space. These prototypes are iteratively refined through foreground–background discrimination, k-nearest-neighbor (kNN) affinity propagation, and gaze-anchored updates, producing a dense segmentation mask from sparse clinician guidance. The entire pipeline requires no segmentation masks, gradient updates, fine-tuning, adapters, or prompt encoders.

As illustrated in Figure 1, the framework consists of four stages: (i) feature extraction and gaze modeling, (ii) gaze-guided prototype initialization, (iii) recurrent gaze-anchored prototype refinement, and (iv) final mask generation.

2.2 Feature Representation and Clinician Gaze Modeling

We map the input image to a pretrained semantic space using a frozen DINOv3 vision transformer [29]. Let $P = \{p_i\}_{i=1}^{N}$, with $p_i \in \mathbb{R}^D$, denote the ℓ_2-normalized patch embeddings extracted from the final transformer layer.

In parallel, clinician gaze is converted into a spatial prior. Given the fixation set $\mathcal{G} = \{(x_m, y_m, d_m)\}_{m=1}^{M}$, where (x_m, y_m) is the fixation location and d_m its duration, we construct a duration-weighted gaze map:

$$H(u, v) = \sum_{m=1}^{M} \tilde{d}_m \exp\left(-\frac{(u - x_m)^2 + (v - y_m)^2}{2\sigma^2}\right), \tag{1}$$

where $\tilde{d}_m = d_m/(\sum_{r=1}^{M} d_r + \epsilon)$, σ controls the spatial spread, and ϵ ensures numerical stability.

The gaze map is normalized to $[0, 1]$ and mapped to the DINOv3 patch grid, yielding $h_i \in [0, 1]$. The initial foreground and complementary background weights are then defined as

$$W_{\mathrm{fg},i}^{(0)} = \frac{h_i + \epsilon}{\sum_{j=1}^{N}(h_j + \epsilon)}, \qquad W_{\mathrm{bg},i}^{(0)} = \frac{1 - h_i + \epsilon}{\sum_{j=1}^{N}(1 - h_j + \epsilon)}. \tag{2}$$

These weights initialize the foreground and background prototypes in the frozen feature space.

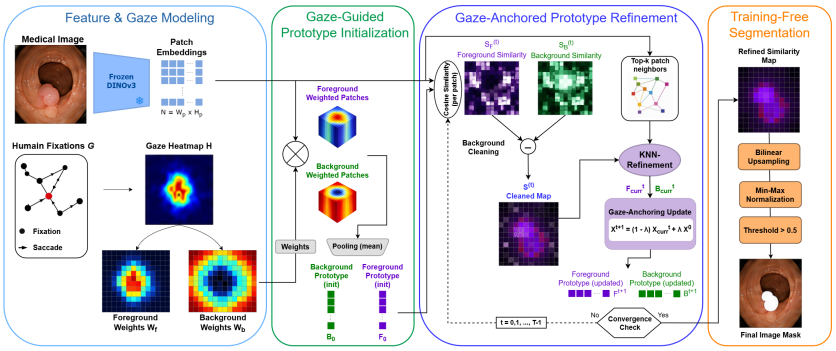

Fig. 1. Overview of **GazeRefine**. Clinician gaze and frozen DINOv3 features initialize foreground and background prototypes, which are iteratively refined to generate the final segmentation mask.

2.3 Gaze-Guided Prototype Initialization

Using the gaze-derived weights, we initialize foreground and background prototypes through weighted pooling in the frozen DINOv3 feature space. Let $\nu(z) = z/(\|z\|_2 + \epsilon)$. The initial prototypes are

$$F^{(0)} = \nu\left(\sum_{i=1}^{N} W_{\mathrm{fg},i}^{(0)} p_i\right), \qquad B^{(0)} = \nu\left(\sum_{i=1}^{N} W_{\mathrm{bg},i}^{(0)} p_i\right). \tag{3}$$

These prototypes serve as gaze-derived foreground and background anchors.

2.4 Gaze-Anchored Prototype Refinement

Starting from the initial gaze-derived prototypes $F^{(0)}$ and $B^{(0)}$, we refine them for at most T iterations. At iteration t, the foreground and background similarities of patch p_i are computed as $s_{\mathrm{fg},i}^{(t)} = p_i^\top F^{(t)}$ and $s_{\mathrm{bg},i}^{(t)} = p_i^\top B^{(t)}$, respectively. Since the patch embeddings and prototypes are ℓ_2-normalized, these inner products correspond to cosine similarities in the frozen DINOv3 feature space.

To suppress background responses, we define the non-negative foreground confidence score

$$S_i^{(t)} = \max\left(0,\, s_{\mathrm{fg},i}^{(t)} - s_{\mathrm{bg},i}^{(t)}\right). \tag{4}$$

The confidence scores are then refined using kNN affinity propagation in the frozen feature space. We compute the affinities $A_{ij} = p_i^\top p_j$, select the k nearest neighbors $\mathcal{N}_k(i)$, and obtain

$$\bar{S}_i^{(t)} = \sum_{j \in \mathcal{N}_k(i)} \frac{\exp(A_{ij}/\tau)}{\sum_{m \in \mathcal{N}_k(i)} \exp(A_{im}/\tau)} S_j^{(t)}, \tag{5}$$

where τ controls the affinity distribution among neighboring patches.

The propagated scores are converted into soft foreground confidence weights as $q_i^{(t)} = \mathrm{sigmoid}(\bar{S}_i^{(t)})$. The updated foreground and background weights are

$$W_{\mathrm{fg},i}^{(t+1)} = \frac{q_i^{(t)} + \epsilon}{\sum_{j=1}^{N}(q_j^{(t)} + \epsilon)}, \qquad W_{\mathrm{bg},i}^{(t+1)} = \frac{1 - q_i^{(t)} + \epsilon}{\sum_{j=1}^{N}(1 - q_j^{(t)} + \epsilon)}. \tag{6}$$

The current prototypes are recomputed as

$$F_{\mathrm{cur}}^{(t)} = \nu\left(\sum_{i=1}^{N} W_{\mathrm{fg},i}^{(t+1)} p_i\right), \qquad B_{\mathrm{cur}}^{(t)} = \nu\left(\sum_{i=1}^{N} W_{\mathrm{bg},i}^{(t+1)} p_i\right). \tag{7}$$

To limit semantic drift, the current prototypes are blended with the initial gaze-derived anchors:

$$F^{(t+1)} = \nu\Big((1-\lambda)F_{\mathrm{cur}}^{(t)} + \lambda F^{(0)}\Big), \qquad B^{(t+1)} = \nu\Big((1-\lambda)B_{\mathrm{cur}}^{(t)} + \lambda B^{(0)}\Big), \tag{8}$$

where $\lambda \in [0,1]$ controls the anchoring strength.

Refinement stops when the maximum number of iterations T is reached or when both prototypes converge, i.e., when $\|F^{(t+1)} - F^{(t)}\|_2 < \varepsilon$ and $\|B^{(t+1)} - B^{(t)}\|_2 < \varepsilon$. We denote the final iteration by $T^\star$.

2.5 Final Segmentation and Training-Free Inference

After refinement, the score map $\bar{S}^{(T^\star)}$ is reshaped to the $h \times w$ patch grid and upsampled to the original image resolution, producing a dense response map M. The map is normalized as

$$M_{\mathrm{norm}} = \frac{M - M_{\min}}{M_{\max} - M_{\min} + \epsilon}. \tag{9}$$

The final binary mask is obtained by thresholding the normalized response map, i.e., $\hat{Y}(u,v) = \mathbb{I}(M_{\mathrm{norm}}(u,v) > \theta)$, with $\theta = 0.5$ in all experiments.

3 Experimental Setup

Datasets. We evaluate GazeRefine on Kvasir-SEG for polyp segmentation [12] and NCI-ISBI for prostate segmentation [3]. We follow the gaze annotations and preprocessing protocol of GazeMedSeg [36], which provides fixation data for 1,000 Kvasir-SEG images and 789 NCI-ISBI slices. After removing fixations shorter than 50 ms and those outside the image boundaries, the sequences contain 5–99 fixations per Kvasir-SEG image and 6–54 per NCI-ISBI slice.

Implementation Details. GazeRefine is evaluated in a zero-shot setting with respect to model training: no segmentation labels, parameter updates, task-specific fine-tuning, adapters, or prompt encoders are used. Images are processed at 592×592 resolution using a frozen DINOv3 ViT-L/16 backbone [29], producing 37×37 patch embeddings with dimension $D = 1024$. Fixation locations and

durations are converted into gaze-derived foreground and background priors for prototype refinement. The same hyperparameters are used for both datasets: $\lambda = 0.5$, $T = 5$, $k = 5$, $\tau = 0.1$, $\varepsilon = 10^{-6}$, and $\theta = 0.5$. The Gaussian spread is set to $\sigma = 2.0$ for Kvasir-SEG and $\sigma = 0.5$ for NCI-ISBI.

Evaluation Protocol. For comparison with trainable baselines, we follow the established splits used for each benchmark: 90%/10% for Kvasir-SEG and 87%/13% for NCI-ISBI [3,12]. GazeRefine is evaluated only on the corresponding test subsets. Segmentation performance is measured using the Dice coefficient [8] with its standard deviation.

4 Experimental Results

4.1 GazeRefine vs. State-of-the-Art Methods

Table 1. Comparison with state-of-the-art methods on Kvasir-SEG and NCI-ISBI. Dice (%).

Method	Supervision	Kvasir-SEG Dice (%) ↑	NCI-ISBI Dice (%) ↑
U-Net [25]	Full	82.12 ± 1.11	80.58 ± 0.48
nnU-Net [11]	Full	**88.41 ± 0.47**	82.43 ± 0.18
PointSup [5]	Point	73.05 ± 1.64	73.46 ± 4.71
AGMM [34]	Point	75.57 ± 0.84	73.86 ± 1.26
AGMM [34]	Scribble	67.23 ± 1.02	72.70 ± 1.03
USTM [19]	Scribble	66.31 ± 0.93	56.89 ± 1.23
DMPLS [21]	Scribble	69.23 ± 0.32	57.44 ± 0.56
BoxInst [30]	Box	65.72 ± 2.97	73.78 ± 1.15
BoxTeacher [6]	Box	73.33 ± 1.30	75.60 ± 1.15
VAM [27]	Gaze	72.21 ± 0.71	78.31 ± 0.47
VAM_D-CRF [17]	Gaze	73.12 ± 0.60	73.86 ± 1.91
GazeMedSeg [36]	Gaze	77.80 ± 1.02	77.64 ± 0.57
GradTrack [33]	Gaze	81.01 ± 0.66	80.25 ± 0.40
GazeSAM [31]	**Zero-Shot** + Gaze	80.43 ± 0.22	82.60 ± 0.17
GazeMedSAMv2 [28]	**Zero-Shot** + Gaze	85.19 ± 0.13	**85.62 ± 0.10**
Ours	**Zero-Shot** + Gaze	**88.10 ± 0.14**	80.28 ± 0.11

Table 1 compares GazeRefine with fully supervised, weakly supervised, gaze-supervised [33,36], and zero-shot gaze-guided methods [28,31]. All methods are evaluated under the same protocol and on the same test set. On Kvasir-SEG, GazeRefine achieves a Dice score of $88.10\% \pm 0.14$, closely approaching the fully supervised nnU-Net result of $88.41\% \pm 0.47$, obtained using 90% of the annotated data for training, and outperforming the zero-shot gaze-guided GazeMedSAMv2 result of $85.19\% \pm 0.13$. GazeRefine also shows lower variability than nnU-Net under the same evaluation protocol.

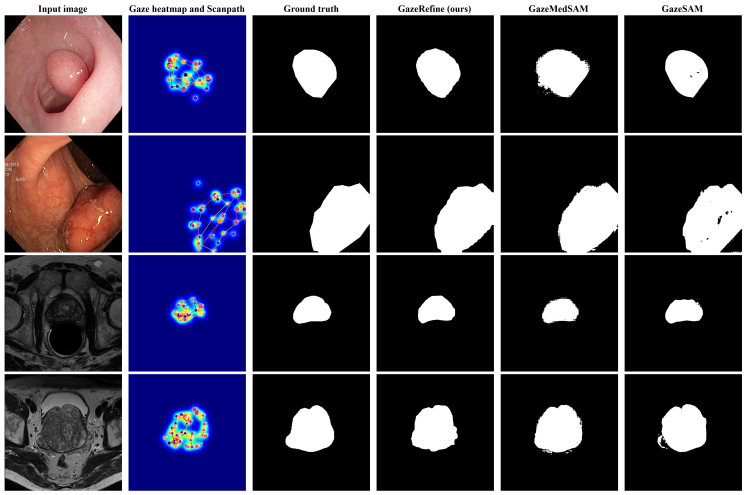

Fig. 2. Qualitative comparison of GazeRefine with gaze-guided zero-shot segmentation methods on representative Kvasir-SEG and NCI-ISBI samples.

On NCI-ISBI, GazeRefine achieves a mean Dice score of $80.28\% \pm 0.11$. Although this is below the best zero-shot gaze-guided result of $85.62\% \pm 0.10$, it remains competitive with several weakly supervised baselines while relying only on frozen DINOv3 features and gaze-derived priors. The lower performance on prostate MRI may reflect the limited separability of low-contrast anatomical boundaries in general-purpose feature spaces, whereas polyps in colonoscopy images often exhibit more distinctive texture and appearance cues. These results suggest that GazeRefine is most effective when foreground and background regions are well separated in the frozen representation space. Figure 2 presents qualitative segmentation results on representative samples from both Kvasir-SEG and NCI-ISBI. Compared with GazeSAM and GazeMedSAMv2, GazeRefine produces masks that more closely match the ground-truth boundaries while maintaining smoother and more coherent object contours. GazeSAM frequently exhibits under-segmentation and is sensitive to specular highlights and illumination artifacts in endoscopic images. Although GazeMedSAMv2 generally provides better object coverage, both SAM-based methods occasionally introduce spurious foreground regions and boundary artifacts. In contrast, GazeRefine yields cleaner masks with fewer false positives and improved boundary localization. These qualitative results demonstrate that the proposed recurrent prototype refinement effectively enhances foreground-background separation and produces more accurate segmentations across both endoscopic and MRI modalities.

4.2 Ablation Study

Table 2 evaluates kNN affinity propagation and background cleaning (**BG Clean**), following the abbreviations used in the table. The full model achieves 88.10%

Table 2. Ablation of kNN propagation and background cleaning. Dice (%).

kNN	BG Clean	Kvasir ↑	NCI-ISBI ↑
✓	✓	**88.10**	**80.28**
✓	✗	28.70	13.50
✗	✓	87.85	73.55
✗	✗	28.35	12.79

Dice on Kvasir-SEG and 80.28% on NCI-ISBI. Removing kNN propagation reduces performance to 87.85% and 73.55%, respectively, showing that affinity propagation is particularly beneficial for prostate MRI. Disabling BG Clean causes a much larger drop on both datasets, confirming its importance for limiting background contamination in the prototype space. When both components are removed, performance remains close to the setting without BG Clean, indicating that background cleaning is the dominant refinement component, while kNN propagation provides complementary gains. Gaze is not ablated because it defines the initial foreground and background prototypes and is therefore an intrinsic input to GazeRefine.

5 Discussion and Conclusion

GazeRefine performs medical image segmentation directly from frozen DINOv3 patch representations, without relying on SAM-like segmentation models, task-specific prompt encoders, fine-tuning, or gradient updates. Clinician gaze initializes foreground and background prototypes, which are refined through background suppression, affinity propagation, and anchoring to the original gaze guidance. The results show that clinician gaze can effectively steer frozen visual representations. GazeRefine achieves strong zero-shot performance on Kvasir-SEG, while its lower performance on NCI-ISBI highlights the challenge of applying general-purpose features to low-contrast prostate MRI. This suggests that the method is most effective when foreground and background regions are sufficiently separable in the pretrained feature space. The ablation study confirms that background cleaning is critical for limiting prototype contamination, while kNN affinity propagation improves spatial coherence. Overall, GazeRefine provides a training-free and segmentation-label-free approach that converts sparse clinician gaze into coherent segmentation masks, supporting clinician-guided prototype refinement as a promising direction for human-in-the-loop medical image segmentation.

Acknowledgment

1- This project was primarily funded by Université Sorbonne Paris Nord.
2- This article has been produced with the financial support of the European Union under the LERCO project number CZ.10.03.01/00/22_003/0000003 via the Operational Programme Just Transition.
3- This work was also partially supported by the National Institutes of Health (NIH) under grants R01-HL171376 and U01-CA268808.

References

1. Barsellotti, L., et al.: Talking to dino: Bridging self-supervised vision backbones with language for open-vocabulary segmentation (2025), https://arxiv.org/abs/2411.19331
2. Bearman, A., Russakovsky, O., Ferrari, V., Fei-Fei, L.: What's the point: Semantic segmentation with point supervision (2016), https://arxiv.org/abs/1506.02106
3. Bloch, B.N., et al.: Nci-isbi 2013 challenge: Automated segmentation of prostate structures. 10.7937/K9/TCIA.2015.zF0vlOPv (2015), the Cancer Imaging Archive (TCIA)
4. Caron, M., et al.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF ICCV. pp. 9650–9660 (2021)
5. Cheng, B., Parkhi, O., Kirillov, A.: Pointly-supervised instance segmentation (2022), https://arxiv.org/abs/2104.06404
6. Cheng, T., Wang, X., Chen, S., Zhang, Q., Liu, W.: Boxteacher: Exploring high-quality pseudo labels for weakly supervised instance segmentation (2023), https://arxiv.org/abs/2210.05174
7. Cohen, N., Newman, Y., Shamir, A.: Semantic segmentation in art paintings (2022), https://arxiv.org/abs/2203.03238
8. Dice, L.R.: Measures of the amount of ecologic association between species. Ecology **26**(3), 297–302 (1945). https://doi.org/10.2307/1932409, https://esajournals.onlinelibrary.wiley.com/doi/abs/10.2307/1932409
9. Gibson, E., et al.: Niftynet: A deep-learning platform for medical imaging. Computer Methods and Programs in Biomedicine **158**, 113–122 (2018). https://doi.org/10.1016/j.cmpb.2018.01.025
10. Hamilton, M., Zhang, Z., Hariharan, B., Snavely, N., Freeman, W.T.: Unsupervised semantic segmentation by distilling feature correspondences (2022), https://arxiv.org/abs/2203.08414
11. Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**(2), 203–211 (2021)
12. 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: MultiMedia Modeling. Lecture Notes in Computer Science, vol. 11962, pp. 451–462. Springer (2020). https://doi.org/10.1007/978-3-030-37734-2_37
13. Jiang, Y.: Prompt engineering in segment anything model: Methodologies, applications, and emerging challenges (2025), https://arxiv.org/abs/2507.09562
14. Karani, N., Erdil, E., Chaitanya, K., Konukoglu, E.: Test-time adaptable neural networks for robust medical image segmentation (2020)
15. Khosravan, N., et al.: Gaze2segment: A pilot study for integrating eye-tracking technology into medical image segmentation (2016), https://arxiv.org/abs/1608.03235
16. Kirillov, A., et al.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4015–4026 (2023)
17. Krähenbühl, P., Koltun, V.: Efficient inference in fully connected crfs with gaussian edge potentials (2012), https://arxiv.org/abs/1210.5644
18. Litjens, G., et al.: A survey on deep learning in medical image analysis. Medical Image Analysis **42**, 60–88 (2017)
19. Liu, X., et al.: Weakly supervised segmentation of covid19 infection with scribble annotation on ct images. Pattern Recognition **122**, 108341 (2022). https://doi.org/10.1016/j.patcog.2021.108341

20. Liu, Y., Zhu, M., Li, H., Chen, H., Wang, X., Shen, C.: Matcher: Segment anything with one shot using all-purpose feature matching (2024), https://arxiv.org/abs/2305.13310
21. Luo, X., et al.: Scribble-supervised medical image segmentation via dual-branch network and dynamically mixed pseudo labels supervision (2022), https://arxiv.org/abs/2203.02106
22. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**, 654 (2024). https://doi.org/10.1038/s41467-024-44824-z
23. Oquab, M., et al.: Dinov2: Learning robust visual features without supervision. Transactions on Machine Learning Research (2024), https://openreview.net/forum?id=a68SUt6zFt
24. others, N.T.: Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. Medical Image Analysis **63**, 101693 (2020). https://doi.org/10.1016/j.media.2020.101693
25. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Lecture Notes in Computer Science, vol. 9351, pp. 234–241. Springer (2015). https://doi.org/10.1007/978-3-319-24574-4_28
26. Sacha, M., Rymarczyk, D., Łukasz Struski, Tabor, J., Zieliński, B.: Protoseg: Interpretable semantic segmentation with prototypical parts (2025), https://arxiv.org/abs/2301.12276
27. Schneider, W.X.: Vam: A neuro-cognitive model for visual attention control of segmentation, object recognition, and space-based motor action. Visual Cognition **2**(2-3), 331–376 (1995). https://doi.org/10.1080/13506289508401737
28. Shmykova, T., Khaertdinova, L., Pershin, I.: Zero-shot gaze-based volumetric medical image segmentation (2025), https://arxiv.org/abs/2505.15256
29. Siméoni, O., et al.: Dinov3 (2025), https://arxiv.org/abs/2508.10104
30. Tian, Z., Shen, C., Wang, X., Chen, H.: Boxinst: High-performance instance segmentation with box annotations (2020), https://arxiv.org/abs/2012.02310
31. Wang, B., Aboah, A., Zhang, Z., Bagci, U.: Gazesam: What you see is what you segment (2023), https://arxiv.org/abs/2304.13844
32. Wang, R., others.: Medical image segmentation using deep learning: A survey. IET Image Processing **16**(5), 1243–1267 (Jan 2022)
33. Wang, Z., Ye, Y., Chen, Z., Xia, Y.: Enjoying information dividend: Gaze track-based medical weakly supervised segmentation. In: MICCAI 2025. Lecture Notes in Computer Science, vol. 15969. Springer, Cham (2026). https://doi.org/10.1007/978-3-032-05127-1_20
34. Wu, L., et al.: Sparsely annotated semantic segmentation with adaptive gaussian mixtures. 2023 IEEE/CVF (CVPR) pp. 15454–15464 (2023)
35. Zhang, R., Jiang, Z., Guo, Z., Yan, S., Pan, J., Ma, X., Dong, H., Gao, P., Li, H.: Personalize segment anything model with one shot (2023), https://arxiv.org/abs/2305.03048
36. Zhong, Y., et al.: Weakly-supervised Medical Image Segmentation with Gaze Annotations . In: proceedings of Medical Image Computing and Computer Assisted Intervention. vol. LNCS 15003. Springer Nature Switzerland (October 2024)