SAM-SPJunc: Self-Prompting for Junction Detection in Retinal Images via Radius-Based Representations

Minasadat Attari, Kannappan Palaniappan, Filiz Bunyak*
Department of Electrical Engineering and Computer Science
University of Missouri-Columbia, MO, USA
{ma8pz, pal, bunyak}@missouri.edu

Abstract

Detecting junctions in the retinal vasculature is vital to analyze topological structures relevant to disease diagnosis and progression. Although deep learning models have achieved high accuracy in medical image segmentation, their decision making remains opaque, limiting their adoption in sensitive clinical applications. In this work, we propose SAM-SPJunc, a SAM-based Self-Prompted Junction Detection architecture where a dedicated decoder first predicts a radius-aware soft mask that encodes potential junction regions. This coarse prediction is then used as a dense prompt to guide a second decoder that acts as a learnable refinement module generating the final junction predictions through regression to a distance transform. By embedding structural prior knowledge in the form of self-generated radius-based prompts, our model improves spatial focus, reduces false positives, and promotes interpretability. This modular design demonstrates that prompting can serve not only as a means of task control, but also as a foundation for more interpretable and structured medical AI systems.

1. Introduction and Background

The high-resolution and non-invasive characteristics of retinal imaging make it an ideal modality for computer vision-based approaches to early disease detection and large-scale screening. In particular, retinal vascular junctions are essential landmarks in biomedical image analysis, particularly in ophthalmology, where they support a variety of clinical and research applications. These junctions play a critical role in the extraction of the retinal vascular topology [2, 6], which is fundamental to the diagnosis of diseases such as diabetic retinopathy [3], cardiovascular diseases[5], Alzheimer's disease[15], and hypertensive retinopathy [9]. In addition, junctions can serve as retina-based biometric identification points for the registration of vascular structures [23] to support the monitoring of the progression of vascular lesions [10]. For these reasons, accurate and interpretable detection of vascular junctions remains a vital challenge in the analysis of retinal images.

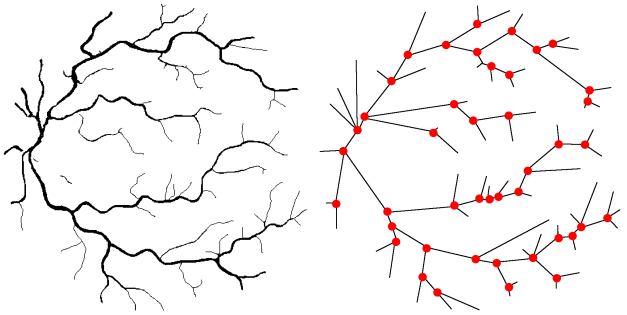


Figure 1. Visualization of a sample retinal vascular map and its corresponding graph representation, with junctions (red disks) defined as graph nodes.

Junction detection is inherently challenging due to the intricate structure of retinal vascular networks and the heterogeneous distribution of junction points. Earlier junction detection methods relied on hand-crafted features or morphological rules [1, 16], but such approaches were often limited in scalability, robustness, and interpretability, as they lacked the capacity to adaptively model vascular complexity across datasets. In recent years, deep learning methods have been applied to junction detection, similar to other image analysis tasks. For example, a multitask framework [18] was proposed that uses vessel features as supervision to identify candidate junctions but still relies on extensive post-processing for refinement. [11] proposed a heatmap regression-based approach, where a fully convolutional network is trained using supervision in the form of heatmaps automatically generated from annotated target pixel locations. In [22] a two-stage method was proposed to detect junctions in retinal images using a Regional Convolutional Neural Network (RCNN). While effective in identi-

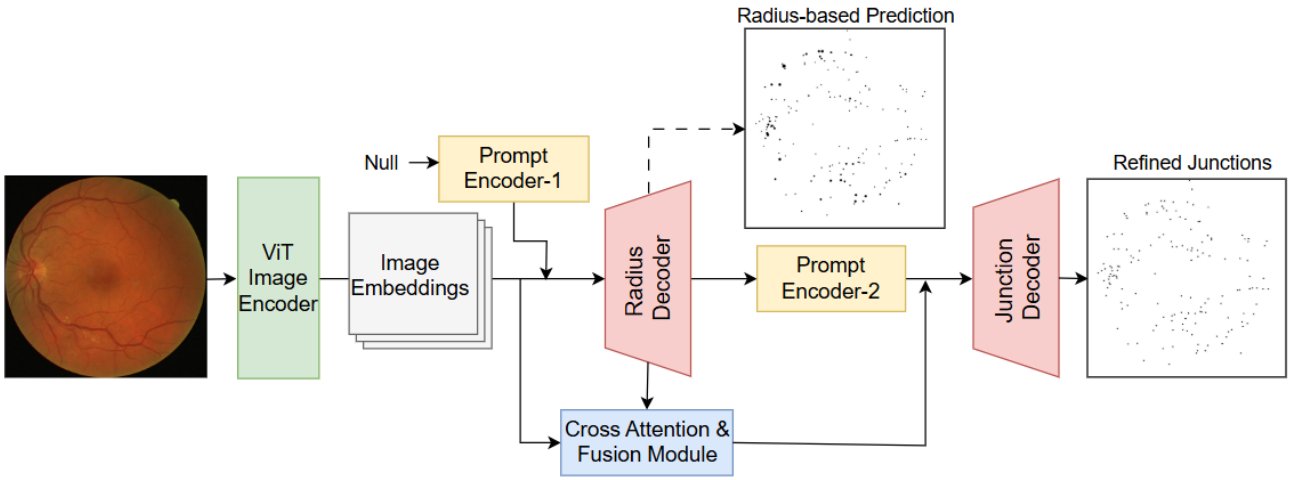


Figure 2. Proposed SAM-SPJunc Network Architecture. It consists of an image encoder adopted from the pre-trained SAM model [13], followed by a radius decoder, two prompt encoders, a junction decoder, and a cross-attention fusion module.

fying candidates, this approach requires bounding box calibration and incurs extra computational cost. Moreover, its bounding box-based design struggles to localize the small, densely clustered junctions typical of biomedical images.

In most recent works, [21] introduced the Attention O-Net, an O-shaped network architecture equipped with attention modules for junction detection in biomedical images without requiring segmentation. The network comprises two key components: the Junction Detection Branch (JDB), which regresses to junction heatmaps, and the Local Enhancement Branch (LEB), which employs a radius-adaptive labeling strategy to enhance thin branches and mitigate the imbalance in heatmap responses between thin and thick junction regions. An approach called Vessel-Guided Junction Detection Network (VGJD-Net) was proposed by [14] that leverages vessel guidance to detect retinal vascular junctions, VGJD-Net which comprises vessel segmentation and junction detection branches, with sharing the same structure, enhances both local and contextual perception of vascular structures through a vessel attention mechanism and a vessel feature perception module.

Recently, the rapid rise of foundational models has demonstrated strong generalization and semantic understanding. The Segment Anything Model (SAM) [13] is one such model for image segmentation, showing impressive zero-shot performance. SAM has been applied across tasks such as medical image segmentation [12], object detection [7], and remote sensing [4], either through fine-tuning or direct use. Prompting has recently gained traction in both vision-language models and image segmentation, including self-prompting strategies where models generate their own guidance signals. In SAM-based architectures, self-prompting has been implemented through simple pixel-wise classifiers that produce internal prompts such as points or boxes [19, 20]. However, such mechanisms have not yet been explored for the task of junction detection, nor have they been used to incorporate geometric priors such as vessel-aware radius masks. In contrast, our SAM-SPJunc integrates anatomical prior knowledge directly into the prompting mechanism. By employing a radius-aware intermediate decoder and guiding the final prediction using soft masks and cross-attention fusion, we propose a more structured and interpretable form of self-prompting that is explicitly grounded in the topological organization of vascular networks. Similarly, cross-attention has recently been injected into the SAM architecture for medical image segmentation[8], but we employ it here as a spatial refinement bridge between decoders.

Despite recent advances in deep learning for curvilinear structure analysis, existing junction detection methods often function as black boxes, lacking transparency and spatial interpretability. In high-precision tasks like junction detection, where decisions depend on subtle geometric cues, this opacity limits reliability and control. We wanted to answer this question: What if the network itself was structured in a way that reflected anatomical priors? To address this, we propose a self-prompting architecture built upon SAM, leveraging its strong image-prompt conditioning capabilities. In our SAM-SPJunc, an initial SAM-based decoder predicts a radius-based soft mask that implicitly encodes geometric priors related to vascular junctions. This self-generated mask is then used as a learned prompt to guide a second SAM decoder focused on refined junction detection. In the case of using a radius-based mask, our work shares some similarities with O-Net [21], which applied a radius-adaptive label to enhance junction visibility.

In contrast, we embed such anatomical priors into a self-prompting framework built upon the SAM architecture, enabling modularity, interpretability, and compatibility with foundation vision models. To our knowledge, this is the first work to utilize self-generated prompts within SAM for vascular junction analysis, offering a more interpretable and spatially grounded solution to this challenging task.

2. Method

2.1. Network Architecture

The overall architecture of the proposed SAM-SPJunc network is illustrated in Figure 2. It comprises an image encoder adapted from the pre-trained SAM model [13], a radius decoder, a junction decoder, two prompt encoders, and a cross-attention fusion module.

2.2. Vision Transformer-based Image Encoder

We adopt the image encoder from the pre-trained Segment Anything Model (SAM), specifically using the smallest variant, ViT-B. This encoder follows the Vision Transformer (ViT) architecture. The encoder transforms an input RGB image of size $(H, W, 3)$ into a dense feature map of shape $(H/16, W/16, D_{\text{feat}})$ and provides these rich visual features for the downstream decoders. To achieve this, the image is first partitioned into non-overlapping patches of size 16×16. Each patch is linearly projected into a token embedding, resulting in an initial tensor of shape $(H/16, W/16, D_{\text{feat}})$. This tensor is then processed through a stack of 12 Transformer blocks employing multi-head self-attention. The feature dimension D_{feat} remains constant throughout.

2.3. Radius and Junction Decoders

We employ two separate decoders: Radius Decoder and Junction Decoder: for radius-based and soft junction predictions, both following the architectural style of the SAM mask decoder [13]. Each decoder consists of a lightweight transformer module and an upsampling head that projects the transformer output back to image resolution.

At the core of each decoder is a two-way transformer, composed of two transformer blocks with 8 attention heads, an embedding dimension equal to that of the prompt encoder, and a feedforward MLP of dimension 2048. This module enables bidirectional interaction between prompt embeddings and image features, enhancing the contextual understanding of spatial cues.

The output of the transformer is passed through an upsampling module comprising two transposed convolution layers. The first layer upsamples the feature map using a kernel size of 2×2 and stride 2, reducing the channel dimension by a factor of 4. The second transposed convolution further doubles the spatial resolution while halving the channel depth. Each convolution is followed by a non-linear activation, and the first layer includes a 2D layer normalization. The final output is a probability map of shape $(H, W, 1)$, indicating the per-pixel likelihood of radius-based or final junction structures.

2.4. Prompt Encoders

Our pipeline employs two prompt encoder modules adapted from the SAM architecture, each serving distinct functions. The first prompt encoder (Prompt Encoder-1) that receives *null* input (without explicit input mask) generates prompt embeddings to guide the radius-based decoder, which outputs a soft radius-aware mask. This intermediate mask is then fed into the second prompt encoder (Prompt Encoder-2), which converts it into dense prompt embeddings. These serve as structured self-prompts for the second decoder, tasked with refining junction-level predictions.

Both encoders are configured with an embedding dimension of 256 and *mask_in_chans=16* , which sets the number of hidden channels used internally during mask embedding. This setup enables our model to integrate hierarchical spatial priors while maintaining modularity and interpretability across stages.

2.5. Cross Attention and Fusion Module

To integrate structural cues from the first decoder into the final junction prediction, we introduce a cross-attention and gated fusion module. Specifically, the intermediate features produced by the radius-based decoder are used to compute cross-attention with the original image embeddings. This guides the network to attend to spatial regions indicative of junction presence. The resulting attention map is fused back with the original image embeddings via a lightweight fusion gate, implemented as two 1×1 convolution layers with ReLU and Sigmoid activations. This gate modulates the contribution of the attended features, allowing the model to selectively enhance regions relevant to downstream junction decoding.

2.6. Loss Functions

To supervise the training of our two-stage junction detection model, we employed different loss functions tailored to each decoder's objective. The first decoder is trained to predict radius-based junction maps, which are sparse and highly imbalanced. To address this, we use the *focal loss*, which helps mitigate class imbalance by down-weighting easy negatives and focusing the learning on hard examples—improving the model's ability to detect small and infrequent junctions:

$$\text{FL}(p_t) = -\alpha(1 - p_t)^{\gamma} \log(p_t) \tag{1}$$

Here, p_t is the model's estimated probability for the true class, defined as p if the ground-truth label is 1

and $1 - p$ otherwise. The parameter $\alpha \in [0, 1]$ balances the importance of positive and negative examples, while $\gamma \geq 0$ modulates the focus on hard-to-classify instances. Following the default settings in `torchvision.ops.sigmoid_focal_loss`, we use $\alpha = 0.25$ and $\gamma = 2$.

For the second decoder, which regresses to the distance transform of ground truth junctions, we use the *mean squared error (MSE) loss*:

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \tag{2}$$

Here, y_i and $\hat{y}_i$ represent the ground truth and predicted values at pixel i, respectively, and N is the total number of pixels. MSE loss provides smooth and continuous supervision, encouraging the network to produce accurate spatial localization of junction centers by penalizing deviations from the target distance map.

3. Experimental Setup

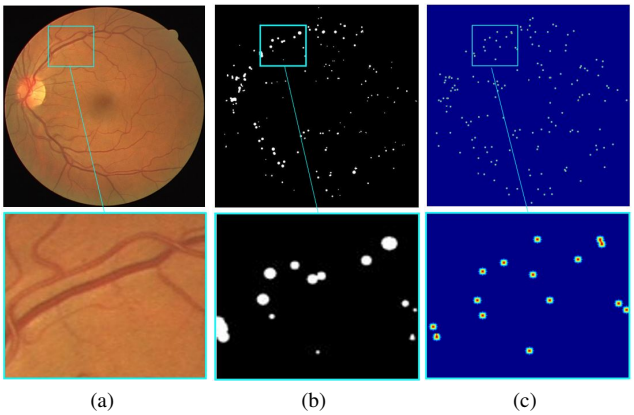
3.1. Dataset

The DRIVE [17] and IOSTAR [1] datasets were used for experimentation, containing 40 and 24 images with resolutions of 584×565 and 1204×1024, respectively. The DRIVE images were resized to 1024×1024 to match the input requirements of our ViT-based image encoder. The network was trained exclusively on the DRIVE training set and evaluated on the DRIVE and IOSTAR test sets. RGB images were used for training and testing. The junction ground truths for both datasets were provided by the authors of [1].

3.2. Mask Generation

To provide spatially informative supervision for junction detection, we construct two distinct ground truth masks, one for each decoder. The first decoder is trained using a radius-based junction mask, while the second decoder uses a soft supervision mask. In the following, we describe the construction process for each mask type in detail. A sample image and its corresponding masks are shown in Figure 3

Radius-based mask: this mask expands each junction point into a circular region whose size reflects the local vessel thickness. Given a binary vessel mask and a binary junction mask, we first compute a vessel thickness map by applying a distance transform to the vessel mask, which estimates the radius at each foreground pixel. For every junction point, we extract the corresponding vessel width and compute a radius by scaling this width, while constraining it within a predefined range $[R_{\min}, R_{\max}]$. A filled circle is then drawn around each junction location using the computed radius, resulting in a mask that contextualizes each junction within the geometry of the surrounding vessel. This radius-based mask is used to supervise the first decoder, encouraging it to learn topologically significant features that improve the accuracy and interpretability of junction prediction.



(a) (b) (c)

Figure 3. Example visualization of masks used for supervision: (a) Raw RGB image, (b) Radius-based junction mask, (c) Soft distance-based junction mask.

Soft mask: to generate the soft supervision mask used for the second decoder, we first dilate each junction point using a disk of radius 5. We then apply a distance transform to the dilated mask, resulting in a smooth, continuous map that provides gradient-based supervision for regression.

3.3. Evaluation Metrics

To evaluate the performance of the proposed SAM-SPJunc method, we follow the common approach used in related papers by computing precision, recall, and F1 Score. A detected junction is considered a True Positive if it falls within a 10-pixel radius of a ground-truth annotation. We adopted this exact evaluation strategy to ensure a fair comparison with previous works. However, since none of the previous papers specify whether their evaluation enforces a one-to-one matching between predictions and ground truths, we use a greedy matching strategy in our ablation study to better show the impact of each component in our network.

3.4. Implementation Details

We trained models for 20 epochs using the Adam optimizer. A base learning rate(LR) of 0.0001 was used for the newly initialized decoder modules, while the pretrained image encoder was fine-tuned with a learning rate of ($0.1\times$ the base LR). A *MultiStepLR* scheduler was applied with a milestone at epoch 9 and a decay factor of 0.1. To compensate for the limited number of training images, we adopted a patch-based training strategy with a stride of 64 to extract overlapping patches and employed simple data augmentations, including random rotations and horizontal/vertical flips.

Method	Threshold	DRIVE			IOSTAR		
		Precision	Recall	F1 Score	Precision	Recall	F1 Score
Pratt [16]	-	0.74	0.54	0.64	0.52	0.54	0.52
Abbasi [1]	-	0.40	0.74	0.52	0.47	0.66	0.62
Uslu [18]	-	0.65	0.69	0.67	0.52	0.67	0.59
Zhao [22]	-	0.71	0.70	0.70	0.62	0.57	0.60
Long [14]	-	0.83	0.76	0.79	0.62	0.66	0.64
Hervella [11]	-	0.80	0.73	0.76	0.74	0.69	0.71
Zhang [21]	-	0.85	0.80	**0.82**	0.72	**0.74**	**0.73**
SAM-SPJunc (Ours)	0.1	0.77	**0.83**	0.79	0.73	0.71	0.71
SAM-SPJunc (Ours)	0.2	**0.89**	0.70	0.78	**0.82**	0.56	0.66

Table 1. Quantitative comparison of SAM-SPJunc (Radius Prompt + Cross-Attention Fusion) against state-of-the-art approaches for junction detection on the DRIVE and IOSTAR datasets. The "Threshold" denotes the confidence level used during local maxima filtering of predicted junctions. The best-performing results are shown in **bold**, and the second-best results are underlined.

4. Results and Discussion

In this section, we present a comprehensive evaluation of our proposed SAM-SPJunc through quantitative metrics, qualitative visualizations, and ablation studies. We first report performance on standard benchmarks to demonstrate the effectiveness of our approach. The qualitative results then highlight the model's ability to locate junctions in challenging cases. Finally, ablation studies analyze the contribution of key architectural components, including radius-based prompting and cross-attention fusion, to assess their individual and combined impact on performance and interpretability.

4.1. Quantitative Results

Table 1 presents the quantitative performance of our proposed method and existing approaches on the DRIVE and IOSTAR datasets. Our SAM-SPJunc model, denoted as Ours (Radius Prompt + Cross-Attention Fusion), is evaluated at two confidence thresholds (0.1 and 0.2) to show representative operating points. On the DRIVE dataset, our method achieves an F1 score of 0.83 at threshold 0.1 and the highest precision of 0.89 at threshold 0.2, reflecting strong discriminative performance. While Zhang [21] reports a comparable F1 score of 0.82, our method exhibits a more structured design and precise localization through its geometry-informed architecture.

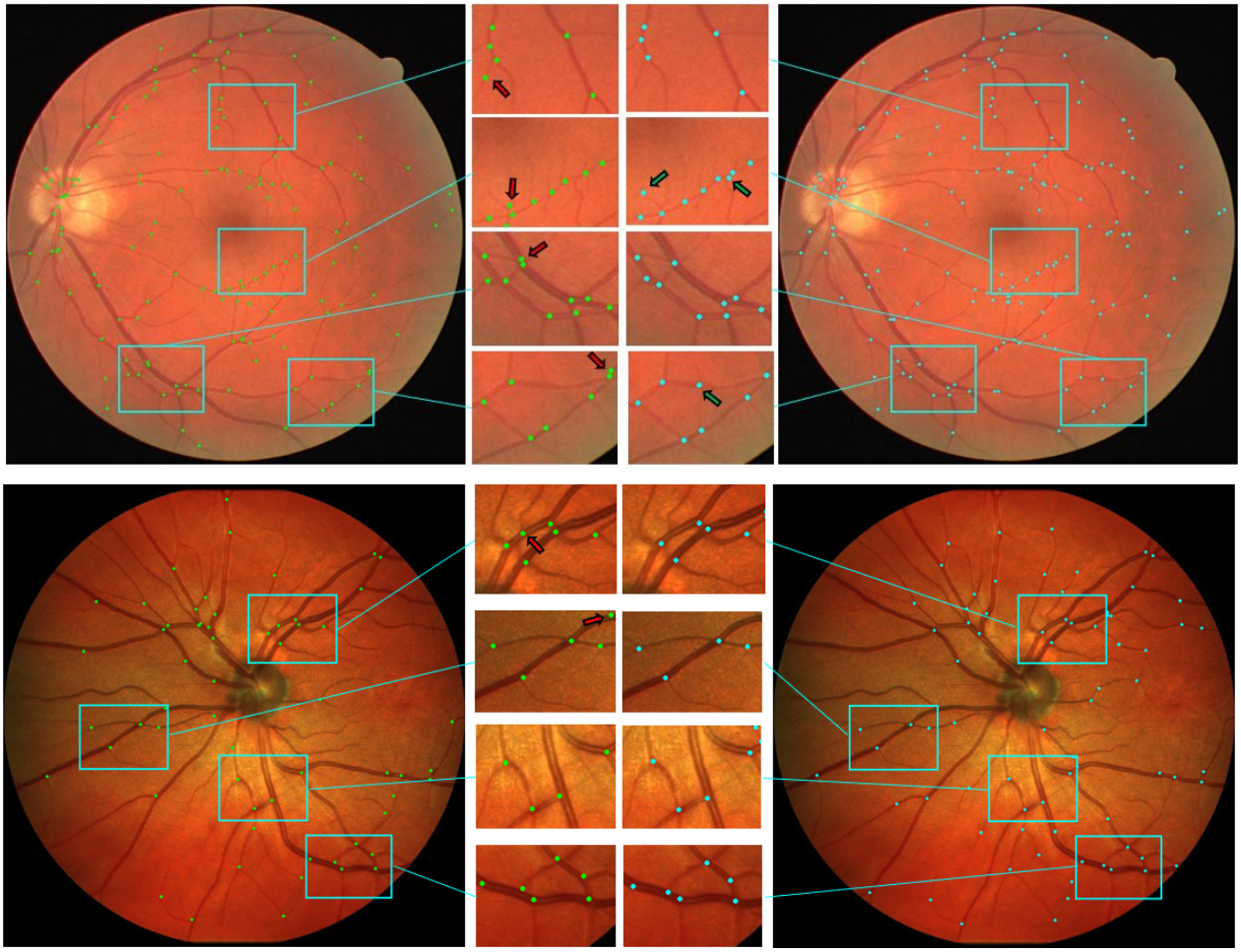
On the more visually challenging IOSTAR dataset, our method achieves the best reported precision and F1 score at threshold 0.1 (0.73 and 0.71, respectively), indicating strong generalization to unseen data. Unlike prior approaches that rely heavily on hand-crafted post-processing, our architecture performs internal soft refinement via its dual-decoder structure. The radius-aware prompt provides geometric guidance, while the second decoder refines spatial predictions—together forming a complementary refinement mechanism that remains fully end-to-end trainable. This design not only improves interpretability but also enhances robustness by incorporating geometric priors directly into the learning process.

4.2. Qualitative Results

Figure 4 presents qualitative results on sample images from the DRIVE and IOSTAR datasets. The predicted junctions (b) align well with the ground truth (a), even under challenging conditions such as central light reflex in the IOSTAR image. The zoomed-in patches demonstrate the model's ability to localize junctions accurately, including those near thin or low-contrast vessels.

Figure 5 highlights the model's robustness in recovering junctions that are entirely missing from both the vessel and junction ground truth masks. In all shown patches, the predicted junctions (c) fall on anatomically plausible locations, often on thin vessels omitted from the annotations. This illustrates the model's ability to generalize beyond incomplete supervision and recover semantically meaningful structures.

Figure 6 highlights the qualitative effect of radius-based self-prompting. Without prompting (Figure 6b), predicted junctions (green) are more scattered and less concentrated around high-confidence regions (blue), resulting in more false positives. With radius prompting (Figure 6c), predictions become more focused and anatomically aligned. Notably, the blue area (representing the activation map of the radius decoder) decreases from 66125 to 56783 pixels reflecting a relative reduction of 14.07%. This reduction suggests that the prompt helps the model suppress irrelevant activations and concentrate more effectively on true junction areas. For this specific example, precision increased from 59% to 69% without any drop in recall, resulting in an overall improvement in the F1 score. By guiding the decoder through anatomically meaningful priors, radius-based



(a) (b)

Figure 4. Visual results of the proposed SAM-SPJunc method on two sample images from the DRIVE and IOSTAR datasets. (a) Junction ground truth, and (b) Predicted junctions overlaid on the RGB image. Zoomed-in patches highlight key cases: junctions without arrows correspond to true detections, red arrows show missed junctions (present in ground truth but not predicted), green arrows show correctly predicted junctions not labeled in the ground truth.

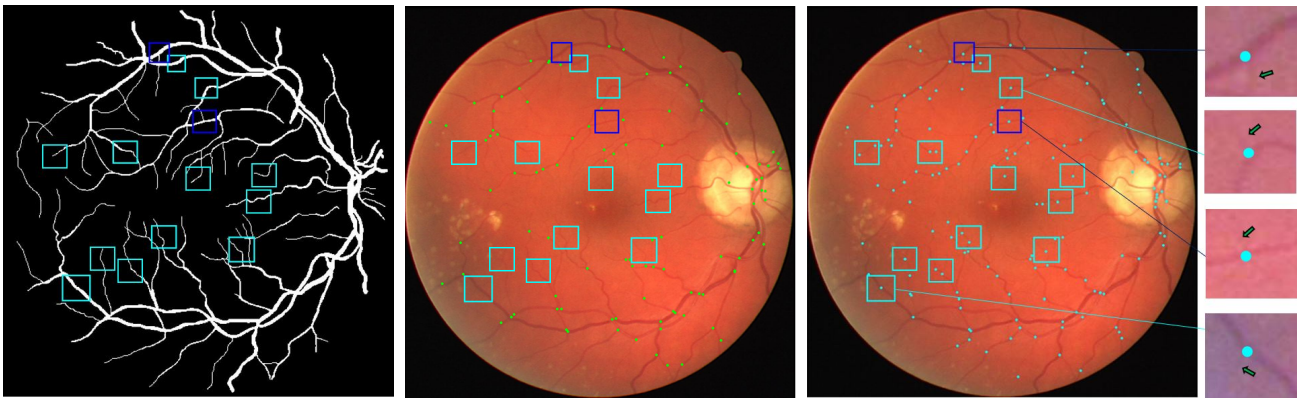
prompting enhances both precision and spatial specificity, contributing to improved interpretability. While this figure provides qualitative insight, the quantitative effect of prompting is further evaluated in the ablation study presented in the next section.

4.3. Ablation Study

Table 2 presents an ablation study evaluating different architectural configurations and prompting strategies for junction detection. For this ablation study, evaluation metrics were computed using a one-to-one (greedy) matching strategy, where each predicted junction is matched to the nearest ground truth within a fixed radius, and duplicate matches are not allowed. Since the models are trained only on the DRIVE dataset, the IOSTAR dataset serves as a more challenging cross-domain benchmark.

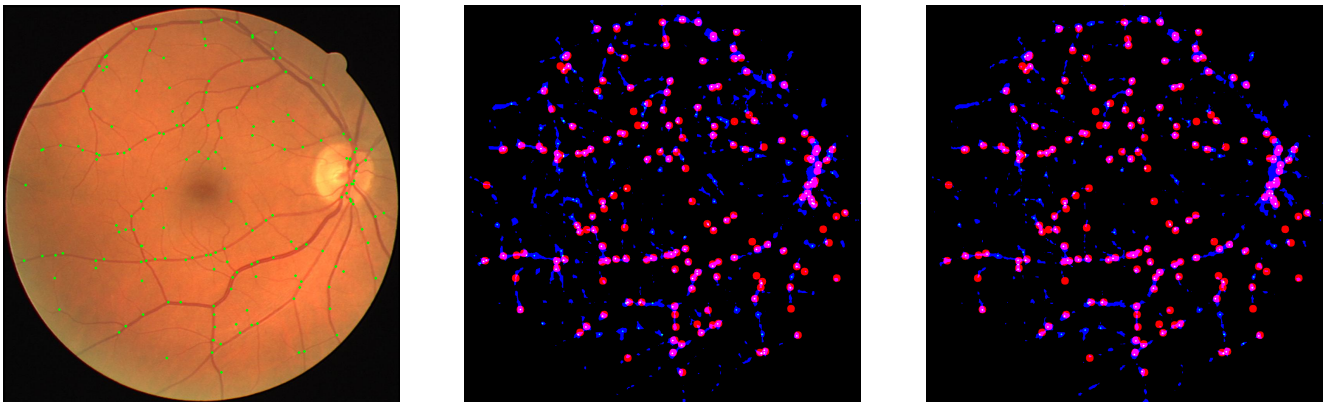
While the simplest configuration using a single decoder and a single radius-junction output achieves the highest recall (0.83 on DRIVE, 0.79 on IOSTAR), it suffers from extremely low precision (0.22 and 0.20, respectively), indicating a large number of false positives and the lack of an effective refinement mechanism. Introducing a two decoder design with vessel-prompt-based guidance improves precision significantly (e.g., 0.68 on DRIVE at threshold 0.2), yet the overall F1 score remains low. Qualitative inspection revealed that this vessel prompt often misleads the junction decoder by highlighting extended vessel regions rather than discrete junctions, resulting in widespread false activations.

Replacing the vessel prompt with a radius-based soft prompt led to a better balance between recall and precision, suggesting that radius-aware supervision helps local-



(a) (b) (c)

Figure 5. Examples of missing ground truth (GT) and abilities of SAM-SPJunc in detecting junctions on very thin and/or faint branching vessels. (a) Vessel GT, (b) Junction GT, (c) Predicted junctions. Cyan boxes indicate regions where both vessel and junction GT are missing; blue boxes indicate missing junction GT despite the presence of vessel GT. (c) and the contrast-enhanced zoomed patches show that the model successfully detects these cases, even for thin vessels (pointed by green arrows).



(a) (b) (c)

Figure 6. Effect of radius-based prompting on junction prediction. (a) Dilated ground-truth junctions overlaid on the raw image. (b) Final predictions using only cross-attention. (c) Predictions using both cross-attention and radius-based self-prompting. Red: dilated ground truth, green: predicted junctions, blue: radius decoder output. Prompting reduces the extent of the radius activation map (blue area: 66125 pixels$\rightarrow$ 56783 pixels: 14.13% decrease) and improves precision (59% $\rightarrow$ 69%: 10% increase) without affecting recall, leading to more focused and accurate junction predictions, try making BG white.

ize junction-relevant features more effectively. This design introduces an interpretable intermediate representation that encodes vessel context and mimics human intuition, widening attention around thicker junctions and constraining it around finer ones. Incorporating a cross-attention fusion module, used to selectively integrate encoder and decoder features, further stabilized training and improved robustness.

The best performance is achieved with the full model (radius prompt + attention fusion), reaching an F1 score of 0.66 on DRIVE and 0.60 on IOSTAR, with the highest precision across both datasets (0.76 on DRIVE and 0.71 on IOSTAR). These improvements highlight not only better generalization, but also the benefit of modular interpretability: The radius prompt encodes prior anatomical structure, while the attention mechanism offers transparent spatial refinement. Together, they make the model's reasoning more structured and explainable, aligning well with the goals of interpretable and controllable spatial computing.

5. Conclusion

In this paper, we introduced SAM-SPJunc: a self-prompted architecture for junction detection in retinal images, built upon the Segment Anything Model (SAM). Our method leverages anatomical priors by predicting a radius-aware soft mask, which serves as a dense prompt to guide a second decoder responsible for refining junction predictions.

Method	Threshold	DRIVE			IOSTAR		
		Precision	Recall	F1 Score	Precision	Recall	F1 Score
1 Decoder (Radius Junction)	0.1	0.22	**0.83**	0.34	0.20	**0.79**	0.31
1 Decoder (Radius Junction)	0.2	0.48	0.72	0.57	0.47	0.67	0.54
2 Decoders (Vessel Prompt)	0.1	0.45	0.71	0.55	0.44	0.66	0.50
2 Decoders (Vessel Prompt)	0.2	0.68	0.54	0.60	0.65	0.43	0.50
SAM-SPJunc (Radius Prompt)	0.1	0.60	0.70	0.64	0.56	0.63	0.59
SAM-SPJunc (Radius Prompt)	0.2	0.74	0.60	**0.66**	0.70	0.53	**0.60**
SAM-SPJunc (Cross-Attention Fusion)	0.1	0.61	0.69	0.64	0.64	0.61	0.58
SAM-SPJunc (Cross-Attention Fusion)	0.2	0.74	0.61	**0.66**	0.67	0.53	0.59
SAM-SPJunc (Radius Prompt + Cross-Attention Fusion)	0.1	0.63	0.68	0.65	0.61	0.60	**0.60**
SAM-SPJunc (Radius Prompt + Cross-Attention Fusion)	0.2	**0.76**	0.59	**0.66**	**0.71**	0.49	0.58

Table 2. Ablation study evaluating different model configurations and prompt strategies on the DRIVE and IOSTAR datasets. "Threshold" refers to the confidence level used for local maxima filtering. The best results are shown in **bold**, and the second-best results are underlined

This design not only reduces false positives, but also improves interpretability by aligning the model's internal attention with meaningful vascular structures.

Quantitative evaluations on the DRIVE and IOSTAR datasets demonstrate competitive or superior precision compared to existing methods, particularly under low confidence thresholds. The proposed framework acts as a self-refining pipeline, where the initial decoder provides coarse geometric guidance and the second decoder performs structured refinement, mimicking post-processing while remaining end-to-end trainable. Our results suggest that embedding task-specific priors into a prompting mechanism can improve both accuracy and transparency, offering a compelling direction for interpretable medical image analysis.

6. Acknowledgement

This work is partially supported by the U.S. National Institutes of Health (NIH), National Institute of Neurological Disorders and Stroke under Award Number R01NS110915. Computational resources for this research were supported by the National Science Foundation (NSF) under Award Number OAC-2322063 and the NSF National Research Platform, as part of the GP-ENGINE Award Number OAC-2322218. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the U.S. Government or its agencies.

References

[1] Samaneh Abbasi-Sureshjani, Iris Smit-Ockeloen, Erik Bekkers, Behdad Dashtbozorg, and Bart ter Haar Romeny. Automatic detection of vascular bifurcations and crossings in retinal images using orientation scores. In *IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 189–192, 2016. 1, 4, 5

[2] Minasadat Attari, Nguyen P Nguyen, Kannappan Palaniappan, and Filiz Bunyak. Multi-loss topology-aware deep learning network for segmentation of vessels in microscopy images. In *2023 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, pages 1–7. IEEE, 2023. 1

[3] Chandrakumar Balaratnasingam, Maiko Inoue, Seungjun Ahn, Jesse McCann, Elona Dhrami-Gavazi, Lawrence A Yannuzzi, and K Bailey Freund. Visual acuity is correlated with the area of the foveal avascular zone in diabetic retinopathy and retinal vein occlusion. *Ophthalmology*, 123 (11):2352–2367, 2016. 1

[4] Keyan Chen, Chenyang Liu, Hao Chen, Haotian Zhang, Wenyuan Li, Zhengxia Zou, and Zhenwei Shi. RSPrompter: Learning to prompt for remote sensing instance segmentation based on visual foundation model. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–17, 2024. 2

[5] Carol Y Cheung, Dejiang Xu, Ching-Yu Cheng, Charumathi Sabanayagam, Yih-Chung Tham, Marco Yu, Tyler Hyungtaek Rim, Chew Yian Chai, Bamini Gopinath, Paul Mitchell, et al. A deep-learning system for the assessment of cardiovascular disease risk via the measurement of retinal-vessel calibre. *Nature Biomedical Engineering*, 5(6):498–508, 2021. 1

[6] Behdad Dashtbozorg, Ana Maria Mendonça, and Aurélio Campilho. An automatic graph-based approach for artery/vein classification in retinal images. *IEEE Transactions on Image Processing*, 23(3):1073–1083, 2013. 1

[7] Shixuan Gao, Pingping Zhang, Tianyu Yan, and Huchuan Lu. Multi-scale and detail-enhanced segment anything model for salient object detection. In *Proc. ACM International Conference on Multimedia*, pages 9894–9903, 2024. 2

[8] Shreyank N Gowda and David A Clifton. CC-SAM: SAM with cross-feature attention and context for ultrasound image segmentation. In *European Conference on Computer Vision*, pages 108–124. Springer, 2024. 2

[9] Andrea Grosso, Franco Veglio, Massimo Porta, FM Grignolo, and TY Wong. Hypertensive retinopathy revisited: some answers, more questions. *British Journal of Ophthalmology*, 89(12):1646–1654, 2005. 1

[10] Carlos Hernandez-Matas, Xenophon Zabulis, and Antonis A Argyros. REMPE: Registration of retinal images through eye modelling and pose estimation. *IEEE Journal of Biomedical and Health Informatics*, 24(12):3362–3373, 2020. 1

[11] Álvaro S Hervella, José Rouco, Jorge Novo, Manuel G Penedo, and Marcos Ortega. Deep multi-instance heatmap regression for the detection of retinal vessel crossings and bifurcations in eye fundus images. *Computer Methods and Programs in Biomedicine*, 186:105201, 2020. 1, 5

[12] Yuhao Huang, Xin Yang, Lian Liu, Han Zhou, Ao Chang, Xinrui Zhou, Rusi Chen, Junxuan Yu, Jiongquan Chen, Chaoyu Chen, et al. Segment anything model for medical images? *Medical Image Analysis*, 92:103061, 2024. 2

[13] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *IEEE Int. Conf. on Computer Vision*, pages 4015–4026, 2023. 2, 3

[14] Jun Long, Yining Xie, Pinhan Yuan, and Yuhang Zhang. Vessel-guided and graph-based retinal vessel junction detection and classification. *Expert Systems with Applications*, 275:126927, 2025. 2, 5

[15] Yuhui Ma, Huaying Hao, Jianyang Xie, Huazhu Fu, Jiong Zhang, Jianlong Yang, Zhen Wang, Jiang Liu, Yalin Zheng, and Yitian Zhao. Rose: a retinal oct-angiography vessel segmentation dataset and new model. *IEEE Transactions on Medical Imaging*, 40(3):928–939, 2020. 1

[16] Harry Pratt, Bryan M Williams, Jae Ku, Frans Coenen, and Yalin Zheng. Automatic detection and identification of retinal vessel junctions in colour fundus photography. In *Medical Image Understanding and Analysis (MIUA)*, pages 27–37. Springer, 2017. 1, 5

[17] Joes Staal, Michael D Abràmoff, Meindert Niemeijer, Max A Viergever, and Bram Van Ginneken. Ridge-based vessel segmentation in color images of the retina. *IEEE Transactions on Medical Imaging*, 23(4):501–509, 2004. 4

[18] Fatmatülzehra Uslu and Anil Anthony Bharath. A multi-task network to detect junctions in retinal vasculature. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 92–100. Springer, 2018. 1, 5

[19] Assefa S Wahd, Jessica Küpper, Jacob L Jaremko, and Abhilash R Hareendranathan. Semantic AutoSAM: Self-prompting segment anything model for semantic segmentation of medical images. In *Int. Conf. IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1–4, 2024. 2

[20] Qi Wu, Yuyao Zhang, and Marawan Elbatel. Self-prompting large vision models for few-shot medical image segmentation. In *MICCAI Workshop on Domain Adaptation and Representation Transfer*, pages 156–167. Springer, 2023. 2

[21] Yuqiang Zhang, Min Liu, Fuhao Yu, Tieyong Zeng, and Yaonan Wang. An o-shape neural network with attention modules to detect junctions in biomedical images without segmentation. *IEEE Journal of Biomedical and Health Informatics*, 26(2):774–785, 2021. 2, 5

[22] He Zhao, Yun Sun, and Huiqi Li. Retinal vascular junction detection and classification via deep neural networks. *Computer Methods and Programs in Biomedicine*, 183:105096, 2020. 1, 5

[23] Yuanjie Zheng, Ebenezer Daniel, Allan A Hunter III, Rui Xiao, Jianbin Gao, Hongsheng Li, Maureen G Maguire, David H Brainard, and James C Gee. Landmark matching based retinal image alignment by enforcing sparsity in correspondence matrix. *Medical Image Analysis*, 18(6):903–913, 2014. 1