Shape-Aware Deep Segmentation in Retinal Vessel Imaging from Sparse and Weak Labels

Oscar Alonso[1,2], Philippe Valmaggia[3,4], Simone Bonato[1], Arzu Cöltekin[2], Peter Maloca[3,4], and Susanne Suter[1]

[1] FHNW School of Computer Science, Institute of Data Science (I4DS), Brugg-Windisch, Switzerland
[2] FHNW School of Computer Science, Institute of Interactive Technologies (IIT), Brugg-Windisch, Switzerland
[3] University of Basel, Department of Biomedical Engineering, Allschwil, Switzerland
[4] Department of Ophthalmology, Universitätsspital Basel, Basel, Switzerland

Abstract. Retinal vessel segmentation in optical coherence tomography (OCT) is challenging because dense annotations require expert knowledge and are time-consuming. We investigated shape-aware deep segmentation from sparse vessel-center annotations covering only a subset of visible vessels. Using UNet++ as baseline, we evaluated three shape priors: SPM, SPGNet, and an HBS-inspired loss. All methods were compared on a generated proxy dataset with dense masks. SPM achieved the highest validation scores and was subsequently evaluated on real OCT using blob weak labels derived from sparse annotations. Although Dice, Generalized Dice, and mean Intersection over Union favored the conservative UNet++ baseline against blob labels, visual and quantitative foreground analyses showed that SPM predicted more vessel-like foreground and connected components and achieved higher weak-label recall. These results indicate that SPM can produce plausible vessel-like predictions beyond weak-label coverage, while overlap metrics alone cannot establish anatomical correctness from sparse annotations.

Keywords: Retinal vessel segmentation, Shape priors, Weak supervision, Medical image segmentation, UNet++

1 Introduction

Dense image annotations require costly, time-consuming expert delineation, limiting supervised deep-learning image segmentation. Retinal optical coherence tomography (OCT) sequences contain hundreds of noisy cross-sectional scans, where structures of interest may occupy only a small fraction of each image.

We segmented retinal vessels in individual B-scans repeatedly acquired at the same retinal location over several seconds. Only centers of some visible vessels were annotated, without masks describing width, boundary, continuity, orientation, or topology. Segmentation rather than keypoint detection was chosen to recover vessel extent and shape. Incomplete supervision encourages agreement with marked centers rather than recovery of all visible vessels. Similarly, overlap metrics count predictions outside annotated regions as false positives, even if they correspond to unannotated vessels.

Retinal vessels vary in diameter and appearance but share compact, elongated, and locally connected cross-sectional shapes. We investigated shape-aware segmentation to exploit these regularities and potentially recover regions associated with unannotated

vessels. However, center annotations lack full shapes, preventing direct assessment of anatomical correctness.

We used a two-stage strategy: three shape-prior approaches were compared with UNet++ on a controlled, densely labeled proxy dataset, followed by comparison of the best-performing approach with UNet++ on real OCT using blob weak labels derived from vessel centers. Dice-style metrics were suitable for the proxy data but less informative under sparse OCT annotations. Real-data evaluation therefore also measured foreground area, connected components, weak-label recall, and foreground outside annotated regions.

The main contributions of this work are:

- a controlled comparison of shape-prior strategies for image segmentation, including architecture-level, statistical, and loss-level approaches, using a common proxy dataset;
- an adaptation of sparse vessel-center annotations into weak segmentation targets for slice-wise OCT training;
- an analysis illustrating how standard overlap metrics and qualitative inspection can lead to different interpretations under incomplete labels; and
- a discussion of how shape-aware predictions could support future expert-in-the-loop annotation refinement.

2 Related Work

2.1 UNet-Based Medical Image Segmentation

U-Net encoder-decoder architectures combine contextual features with high-resolution information through skip connections [1]. UNet++ uses nested dense pathways to reduce the semantic gap between encoder and decoder representations [2]. We selected it as the baseline because dense feature reuse helps preserve fine details during downsampling and reconstruction, particularly for small OCT vessel structures.

Conventional segmentation losses optimize pixel-level agreement with reference annotations. Complete masks may support implicit shape learning, whereas explicit priors encode geometry, topology, or boundaries beyond local pixel-wise objectives. Such information may be useful with incomplete annotations.

2.2 Explicit Shape-Prior Strategies

Explicit shape priors include learnable modules, statistical models, and geometry-aware objectives [3]. Before experimentation, we selected SPM, SPGNet, and an HBS-inspired regularizer to represent these mechanisms. They respectively learn priors jointly, construct them from contours, and use a reference descriptor, a relevant contrast because OCT annotations lack complete shapes.

The Shape Prior Module (SPM) of You et al. [4] stores coarse shape information in learnable prior maps. Self-update models interactions among priors; cross-update combines them with decoder features before the segmentation head. SPM can be integrated into U-Net-like models without replacing the encoder-decoder architecture.

SPGNet is a dedicated dual-path network with segmentation and shape-prior branches that jointly predict masks, shape classes, and landmarks [5]. It derives interpretable statistical priors from aligned training contours using principal component analysis and clustering, requiring consistent landmarks and offline preprocessing.

The Harmonic Beltrami Signature (HBS) represents two-dimensional shapes through the Beltrami coefficient of a harmonic map and is invariant to translation, rotation, and scaling [6]. We adapted its descriptor principle into a differentiable regularization loss that compares predicted contours with a training-mask reference while preserving the UNet++ architecture.

3 Data and Methods

The two-stage experimental workflow is summarized in Fig. 1.

3.1 Real OCT Dataset and Sparse Annotations

The dataset comprises 113 annotated dynamic OCT sequences acquired using Valmaggia et al.'s protocol [7]. Repeated video-rate B-scans at one retinal location capture temporal vessel signals associated with fringe washout. Each sequence contains 512 B-scans of 496×1024 pixels at a nominal A-scan rate of 85 kHz, with speckle noise and a paired scanning laser ophthalmoscopy (SLO) image. The 113,664 vessel-center annotations distinguish arteries from veins. We retained the existing split of 90 training and 23 validation sequences. It included no separate OCT test subset, so the reported OCT scores use validation data also used for checkpoint selection.

The sequence-level split from the existing OCT pipeline did not enforce patient-level separation, limiting claims about generalization to unseen patients. Slice-wise 2D processing matched vessel-center labels and the UNet++ pipeline, yielding 46,080 training and 11,776 validation B-scans but ignoring temporal continuity.

Each vessel-center annotation was expanded into a local binary blob to obtain a weak segmentation target (Fig. 2). These targets remained incomplete regarding vessel boundaries and unannotated structures. The primary UNet++ versus SPM comparison used blob-only supervision; auxiliary variants added Gaussian heatmaps with $\sigma = 9$.

3.2 Controlled Proxy Dataset

A controlled proxy dataset was generated to evaluate the methods under complete shape supervision (Fig. 3). It contains 100 grayscale images with dense binary masks and random ellipses varying in size, eccentricity, rotation, overlap, and occlusion. Low contrast,

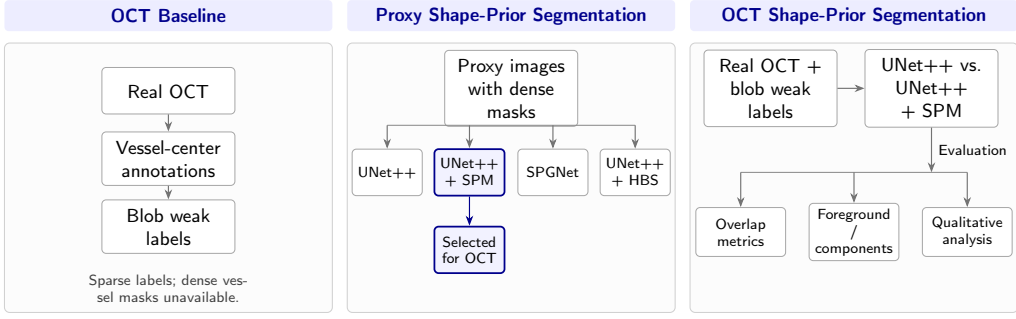


Fig. 1. Two-stage experimental strategy: shape-prior benchmarking on proxy data under dense supervision, followed by evaluation of the selected SPM approach on real OCT under weak supervision

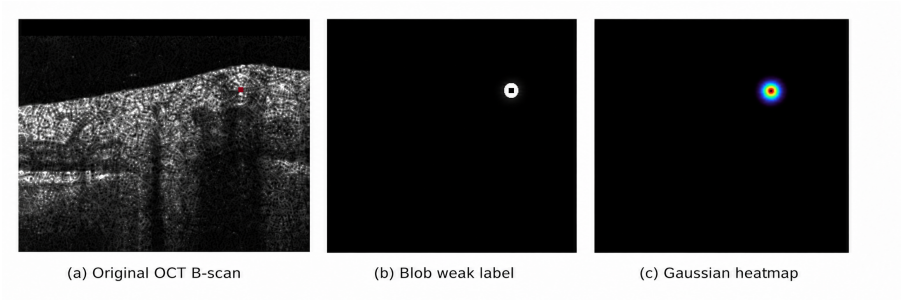


Fig. 2. Weak supervision from vessel-center annotations on a real OCT example, showing the original vessel-center label (red square), binary blob, and Gaussian heatmap

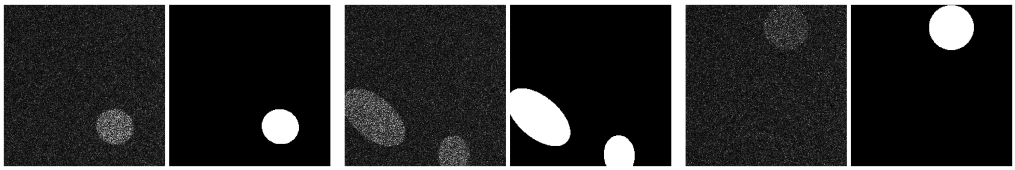


Fig. 3. Three proxy-image/dense-mask pairs illustrating variation in object number, size, eccentricity, orientation, position, boundary truncation, and contrast

speckle-like noise, blur, and additive noise provided challenging appearance without simulating OCT physics. A common 70/15/15 split allocated most samples to training while retaining equal validation and test subsets.

3.3 Segmentation Methods

UNet++ Baseline UNet++ [2] was configured for binary segmentation. Its implementation also served as the backbone for SPM and HBS; only the shape-prior component or regularization objective was added, enabling comparison against a common architecture.

UNet++ with SPM SPM [4] was inserted after the UNet++ decoder (Fig. 4). The selected proxy configuration used one learnable prior; real OCT used four. In both settings, the 16×16 priors were initialized from a zero-centered normal distribution with standard deviation 0.02. The module used four attention heads, hidden size 256, and two update layers. Self-update models prior interactions; cross-update combines prior information with decoder features before the segmentation head.

SPGNet Each dense proxy mask's foreground boundary (contour) was represented by 128 landmarks. Alignment, principal component analysis, and clustering produced statistical shape classes. SPGNet jointly predicted masks, landmarks, and classes: landmark prediction encouraged learned boundary geometry, while classification associated predictions with statistical shape groups. Transfer to OCT was less direct because blob weak labels provide artificial local boundaries rather than complete vessel contours.

HBS-Inspired Regularization Proxy-mask contours were normalized and converted into HBS-like descriptors; their mean served as a global prior. The predicted-mask descriptor

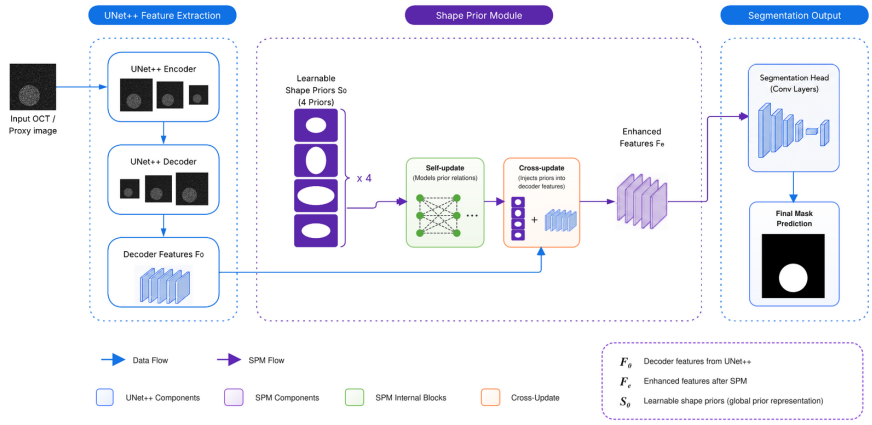


Fig. 4. SPM integration into UNet++: self-update refines learnable priors, and cross-update combines them with decoder features before segmentation

discrepancy augmented the segmentation objective:

$$\mathcal{L} = \mathcal{L}_{\text{seg}} + \lambda_{\text{HBS}} \mathcal{L}_{\text{HBS}}. \tag{1}$$

Here, $\mathcal{L}_{\text{seg}}$ is the segmentation loss, $\mathcal{L}_{\text{HBS}}$ the descriptor discrepancy, and λ_{HBS} its weight. This is an adaptation, not the complete published framework.

3.4 UNet++ Baseline and Shape-Prior Segmentation on Proxy Data

UNet++, UNet++ with SPM, SPGNet, and UNet++ with HBS-inspired regularization were trained independently for 50 epochs with batch size 20 on the same split. Each retained its required components and objectives. Checkpoints were selected by validation Generalized Dice Score (GDS), a class-weighted overlap measure for imbalanced segmentation [8]; Dice and mean Intersection over Union (IoU) provided familiar complementary measures. Proxy performance and integration requirements informed selection for OCT.

3.5 UNet++ Baseline and Shape-Prior Segmentation on OCT

UNet++ and the selected SPM variant shared the sequence-level partition, blob weak labels, learning rate of 10^{-3}, and evaluation protocol, but were trained independently. Both were evaluated with blob-only supervision and with auxiliary Gaussian heatmaps. For each configuration, we selected the checkpoint with the highest validation GDS and also reported Dice and mean IoU. Fixed-crop evaluation was used to reduce crop-sampling variability.

Complementary Foreground Analysis Foreground denotes pixels classified as vessel. Overlap metrics count predictions outside the incomplete weak labels as false positives. We complemented them by comparing the selected blob-only UNet++ and SPM checkpoints on the same random batch of 100 weak-label-positive validation slices. Foreground area measured total positive output, connected-component count the number of separate predicted regions, recall the coverage of weak labels, and extra foreground the predictions outside them. These measures characterize prediction behavior without establishing anatomical correctness.

Table 1. Validation performance on the controlled proxy dataset

Method	GDS	Dice	Mean IoU
UNet++	0.9812	0.9812	0.9633
UNet++ + SPM	**0.9826**	**0.9826**	**0.9659**
SPGNet	0.9789	0.9789	0.9589
UNet++ + HBS loss	0.9753	0.9753	0.9521

Table 2. Best validation performance on real OCT weak labels

Model and supervision	GDS	Dice	Mean IoU
UNet++ blobs	**0.4106**	**0.4252**	**0.3221**
SPM blobs	0.3906	0.4033	0.2954
UNet++ blobs + heatmap	0.3701	0.3855	0.2916
SPM blobs + heatmap	0.3899	0.4031	0.2978

4 Results

4.1 Baseline and Shape-Prior Comparison on the Proxy Dataset

All methods achieved high proxy validation overlap scores, indicating a relatively easy task (Table 1). SPM ranked first, with small improvements across all three metrics. SPGNet and the HBS-inspired variant did not improve over UNet++. SPM's validation performance and direct integration into UNet++ led to its selection for real OCT.

4.2 Real OCT Overlap Performance

Table 2 compares UNet++ and SPM under blob-only and auxiliary heatmap supervision. UNet++ achieved higher GDS with blobs alone; SPM achieved higher GDS with heatmaps. Neither heatmap configuration exceeded blob-only UNet++.

In all four configurations, validation GDS peaked during the early training epochs and declined thereafter. Evaluation using fixed validation crops produced more stable scores but preserved the early peak and subsequent decline.

4.3 Foreground and Connected-Component Analysis

On the 100 weak-label-positive validation slices, SPM predicted approximately three times as much foreground and more than twice as many connected components as UNet++ (Table 3). Weak-label recall increased from 0.404 to 0.801, alongside substantially more foreground outside the labels, counted as false positives by overlap metrics.

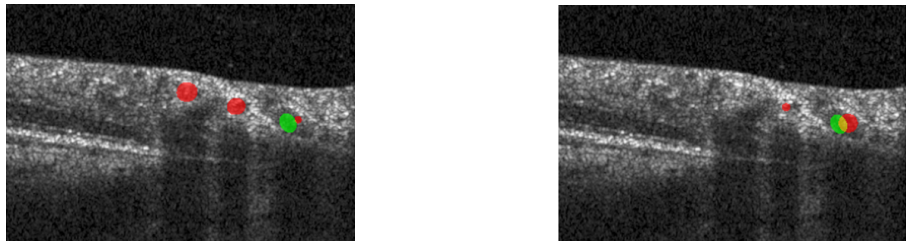
Additional validation analyses showed the same tendency beyond the images saved during training, with less pronounced numerical differences. This supports interpreting SPM as less conservative, not every additional component as a true vessel.

4.4 Qualitative Comparison

Figure 5 compares predictions on the same OCT slice. UNet++ concentrates output near annotated blobs. SPM covers the weak labels and adds compact predictions supported by visible image cues. Across inspected samples, these tended to follow plausible vessel positions and shapes rather than appearing uniformly or randomly throughout the image.

Table 3. Foreground analysis on a random batch of 100 weak-label-positive validation slices

Metric	UNet++ blobs	SPM blobs
Predicted foreground area	229.8	691.8
Connected components	1.02	2.13
Recall against weak labels	0.404	0.801
Extra foreground outside labels	90.6	416.7



(a) SPM (b) UNet++

Fig. 5. Representative real OCT validation slice. Green indicates weak-label foreground, yellow prediction-label overlap, and red prediction outside the weak label. SPM produces more vessel-like foreground, whereas UNet++ remains more conservative

5 Discussion

5.1 Interpretation of the Proxy Experiment

The small gain over UNet++ did not establish whether SPM specifically learned or exploited explicit shape information. The proxy provided a common setting for comparing implementations and practical requirements: SPGNet needed fixed landmarks and offline preprocessing, HBS added regularization to UNet++, and SPM integrated directly after its decoder. Together with validation performance, these requirements informed the OCT selection.

The proxy screened candidate approaches rather than assessing anatomical or clinical performance. Its ellipses simplified vessel morphology, and its degradation model did not reproduce OCT physics.

5.2 Interpreting Real OCT Results Under Weak Labels

UNet++ better reproduced blob weak labels, whereas SPM produced more foreground and components and higher weak-label recall. Some additional predictions appeared spatially plausible. Increased foreground alone does not establish improved segmentation: without dense expert annotations, predictions outside the weak labels cannot reliably be classified as vessels or false positives. Foreground measures describe model behavior, not anatomical accuracy. Agreement with incomplete labels and recovery of plausible vessels are distinct objectives; Dice-style metrics primarily measure the former.

SPM's advantage over UNet++ with auxiliary heatmaps suggests possible benefit from denser spatial guidance. Neither heatmap variant exceeded blob-only UNet++, and further experiments are needed to support this interpretation.

5.3 Limitations

Without dense expert masks or an independent OCT test set, improved anatomical accuracy and generalization cannot be established. Evaluation used validation sequences, with checkpoints selected early in training. Slice-wise processing ignored continuity across B-scans. The small, geometrically simple proxy did not capture the full complexity of retinal vessel appearance and topology. HBS results concern the adapted regularizer, not the complete published framework.

5.4 Toward Human-in-the-Loop Annotation Refinement

A future workflow could use SPM to propose candidates outside weak labels for an ophthalmic expert to accept, reject, or adjust. Foreground and component measures could prioritize slices with large model disagreement, reducing exhaustive annotation effort. Only expert-confirmed candidates would enrich training through semi-supervised or iterative annotation; the model would not define ground truth autonomously. This would address the annotation bottleneck while preserving human oversight.

6 Conclusion

Among three shape-prior strategies, SPM achieved the highest proxy overlap scores and integrated directly into UNet++. On real OCT, UNet++ achieved higher blob-label overlap, while SPM predicted more foreground and connected components and achieved higher weak-label recall. Shape priors altered prediction behavior, but incomplete labels prevented anatomical assessment through overlap metrics alone. Dense expert validation or expert-in-the-loop annotation is needed to assess and use these additional vessel-like candidates.

References

1. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention, Springer (2015) 234–241
2. Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J.: UNet++: A nested U-Net architecture for medical image segmentation. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, Springer (2018) 3–11
3. Bohlender, S., Öksüz, İ., Mukhopadhyay, A.: A survey on shape-constraint deep learning for medical image segmentation. IEEE Reviews in Biomedical Engineering **16** (2023) 225–240
4. You, X., He, J., Yang, J., Gu, Y.: Learning with explicit shape priors for medical image segmentation. arXiv preprint arXiv:2303.17967 (2023)
5. Song, Z., Liu, X., Zhang, W., Gong, Y., Hao, T., Zeng, K.: SPGNet: A shape-prior guided network for medical image segmentation. In: Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence. (2024) 1263–1271
6. Lin, C., Lui, L.M.: Shape prior segmentation guided by harmonic beltrami signature. SIAM Journal on Imaging Sciences **19**(1) (2026) 519–554
7. Valmaggia, P., Cattin, P.C., Sandkühler, R., Inglin, N., Otto, T.P., Aumann, S., Teussink, M.M., Spaide, R.F., Scholl, H.P.N., Maloca, P.M.: Time-resolved dynamic optical coherence tomography for retinal blood flow analysis. Investigative Ophthalmology & Visual Science **65**(6) (2024) 9
8. Sudre, C.H., Li, W., Vercauteren, T., Ourselin, S., Cardoso, M.J.: Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, Springer (2017) 240–248