AI-Driven Detection of Retinal Structural Changes Through Semi-Supervised OCT Image Analysis

Ekapad Salvi
Department of Computer Science, University of New Hampshire, Durham, NH, USA.
ekapad971@unh.edu

Abstract

Retinal optical coherence tomography (OCT) provides cross-sectional views of the neurosensory retina and retinal pigment epithelium, making it central to the detection of fluid, drusen, atrophy, layer disruption, and other structural abnormalities. A persistent difficulty is that clinically useful detection must be spatially precise and longitudinally sensitive while annotated OCT volumes remain scarce, heterogeneous across devices, and expensive to grade. This paper develops a semi-supervised framework for detecting retinal structural changes that treats segmentation as a spatial prior rather than as an isolated endpoint. A U-Net-style segmentation model is trained on labeled scans and regularized on unlabeled scans through class-specific pseudo-labeling, entropy-based reliability weighting, and consistency between augmented student and teacher predictions. A registered longitudinal branch converts layer and lesion probability maps into structural change scores, which are combined with predictive reliability in a detection rule. The composite objective balances supervised segmentation, unlabeled-data consistency, and temporal change consistency. The analysis specifies failure modes that matter in OCT, including confirmation bias from pseudo-labels, false change caused by registration error, speckle and vessel-shadow artifacts, and calibration drift under device shift. It also defines an evaluation logic based on patient-level splits, external device testing, lesion-wise detection metrics, calibration assessment, and ablation of uncertainty and longitudinal components. The contribution is a coherent formulation that links semi-supervised representation learning to clinically meaningful structural change endpoints and provides a technically explicit basis for label-efficient OCT analysis.

Keywords

semi-supervised learning; optical coherence tomography; retinal structural change; uncertainty calibration; pseudo-labeling; longitudinal image analysis

1. Introduction

Deep learning has become a practical route for retinal OCT interpretation because OCT captures layer morphology and lesion texture at a resolution that supports quantitative reading [1]. Convolutional models can detect macular fluid and quantify its extent [2], distinguish normal from age-related macular degeneration scans [3], and provide layer boundaries that support thickness measurement [4]. Encoder-decoder segmentation architectures, particularly U-Net-style networks, are widely used because they combine local texture evidence with multi-scale spatial context [5]. These methods are effective when large, consistently annotated training sets are available. In retinal practice, however, voxel-level labels are limited, lesion boundaries are ambiguous, and acquisition characteristics vary across manufacturers, scan protocols, and patient populations.

Semi-supervised learning addresses this label bottleneck by using unlabeled images to constrain model behavior. Mean-teacher training stabilizes predictions by enforcing

agreement between a student network and an exponential moving average of its parameters [6], while temporal ensembling aggregates earlier predictions as consistency targets [7]. Such approaches are attractive for OCT because unlabeled volumes are abundant relative to expert segmentation. Yet consistency alone does not ensure anatomical validity: a model may agree with itself on an erroneous boundary, especially for low-contrast fluid, drusen, or layer disruption. Segmentation guidance can improve detection-oriented semi-supervised learning by supplying structured spatial targets, an idea illustrated in three-dimensional intraoral scans where segmentation guides semi-supervised landmark detection [8]. The present work adapts that logic to retinal OCT but shifts the objective from static landmark or lesion localization to the detection of structural change over time.

The central contribution is a semi-supervised OCT framework that jointly learns retinal layer and lesion segmentation, constructs reliable pseudo-labels for unlabeled scans, and converts segmentation-derived structural maps into longitudinal change scores. The framework is designed around three claims that follow from the imaging problem rather than from a particular dataset. First, class-specific pseudo-label thresholds are necessary because large retinal layers and small lesions have very different boundary uncertainty. Second, predictive entropy should modulate consistency loss and decision thresholds because unreliable regions are precisely where pseudo-label confirmation bias and false change detection arise. Third, longitudinal change detection requires explicit registration and temporal differencing; static segmentation accuracy alone does not establish that a structural change has occurred.

2. Clinical and Imaging Problem Definition

OCT structural changes relevant to retinal disease include intraretinal fluid, subretinal fluid, pigment epithelial detachment, drusen, hyperreflective foci, epiretinal membrane, and thinning or disruption of specific retinal layers. These abnormalities are not equivalent as detection targets. Fluid is often regionally extensive but can be faint and boundary-ambiguous. Drusen are focal and irregular. Layer thinning is a quantitative deviation that may be gradual and bilateral. Hyperreflective foci are small and can be confused with noise or vessel shadows. A detection system must therefore represent both focal lesions and layer-level deviations without collapsing them into a single foreground class.

The imaging problem is also longitudinal. A single scan can show a lesion, but a structural change requires comparison with a prior scan or with a normative expectation. Clinically meaningful change includes the appearance or enlargement of fluid, increase in drusen burden, development of atrophy, and progressive thinning of the ganglion cell or retinal nerve fiber layers. The detection variable is not merely the presence of an abnormality at time t; it is the spatially localized difference between structural maps at t and a registered earlier scan. This distinction matters because segmentation errors that are stable across time cancel in a difference map, whereas segmentation jitter that varies between visits can create false change.

Several confounders are intrinsic to OCT. Speckle noise, posterior shadowing from vessels, signal attenuation, motion artifact, and segmentation ambiguity at the outer retinal boundary can all produce apparent structural alteration. Clinical endpoints such as fluid volume, central subfield thickness, and lesion area are meaningful only when the underlying segmentation is stable and calibrated. Public benchmarks for OCT fluid detection and segmentation have improved comparability, but they also reveal how strongly performance depends on lesion type and boundary convention [16]. The framework below is therefore

designed to keep detection claims tied to defined anatomical targets, explicit change thresholds, and reliability estimates rather than to a single aggregate score.

3. Semi-Supervised Detection Framework

3.1 Network and Segmentation Objective

Let D_L = {(x_i, y_i)}_{i∈L} denote labeled OCT scans with pixel-level annotations and D_U = {x_j}_{j∈U} denote unlabeled scans. The segmentation network f_θ maps an input scan x to a probability map p = f_θ(x), where p_i^c(v) is the probability of class c at location v. Classes include retinal layers, intraretinal fluid, subretinal fluid, pigment epithelial detachment, drusen, and background. The supervised objective combines cross-entropy and Dice loss to handle class imbalance and boundary sensitivity. The supervised term is

$$L_sup = \sum_\{i \in L\} [CE(y_i, p_i) + \lambda_d \, Dice(y_i, p_i)]$$

where CE is the pixel-wise cross-entropy, Dice is the soft Dice overlap, y_i is the one-hot reference annotation, p_i is the predicted probability map, and λ_d controls the contribution of overlap loss. The Dice term improves gradient behavior for small structures such as hyperreflective foci and thin fluid pockets, whereas cross-entropy stabilizes the classification of large retinal layers. The network can use a U-Net encoder-decoder with skip connections and dilated convolution in deeper layers to enlarge the receptive field without excessive downsampling; this design is consistent with established segmentation practice [5], [20].

3.2 Pseudo-Label Construction and Consistency Regularization

Unlabeled scans are used by generating pseudo-labels from the current model and retaining only predictions that exceed class-specific confidence thresholds. For unlabeled scan j, class c, and location v, the soft pseudo-label is

$$\tilde{y}_j\hat{}c(v) = 1[p_j\hat{}c(v) \geq \tau_c] \, p_j\hat{}c(v)$$

where 1[·] is the indicator function and τ_c is a threshold chosen separately for each class. Large layers can tolerate higher thresholds because their boundaries are supported by broad regional evidence. Small lesions require lower thresholds to avoid discarding true positives, but lower thresholds increase false-positive risk. The threshold should therefore be linked to an uncertainty gate rather than used alone. Segmentation guidance is useful here because it forces the pseudo-label to respect anatomical ordering, for example that subretinal fluid lies between the neurosensory retina and the retinal pigment epithelium.

Consistency regularization then encourages the model to produce stable predictions under augmentations. FixMatch-style training uses a weakly augmented view to generate a target and a strongly augmented view to compute the loss [9]. Unsupervised data augmentation extends this principle by matching predictions across augmentation distributions [10]. MixMatch combines consistency with entropy minimization and data mixing [11]. A broader treatment of semi-supervised assumptions is provided by the survey literature [12]. In OCT, augmentation must preserve anatomical plausibility: horizontal flipping, mild intensity shifts, and local elastic deformation are acceptable, whereas aggressive vertical flipping or arbitrary rotation can violate layer ordering and create anatomically impossible pseudo-labels.

3.3 Uncertainty-Weighted Reliability

Pixel-wise predictive entropy provides a simple reliability signal. For C classes and prediction p_j^c(v), the entropy is

$$H_j(v) = -\sum_{c=1}^{C} p_j^c(v) \log p_j^c(v)$$

where H_j(v) is high when the model distributes probability mass across several classes. Entropy is not a complete uncertainty decomposition, but it is useful as a gating variable because it is available at inference without additional labels. Bayesian approximation through dropout provides one route to epistemic uncertainty [13], [14], while deep ensembles offer a complementary and often better-calibrated signal through prediction disagreement [15]. The reliability weight used in consistency training is

$$u_j(v) = \exp\left(-H_j \frac{(v)}{\log} C\right)$$

where u_j(v) is close to one for confident predictions and approaches zero for highly uncertain locations. The normalization by log C keeps the weight in a comparable range across class configurations. The consistency objective is

$$L_con = \sum_\{j \in U\} \sum_v u_j(v) \, ||f_\theta(T(x_j))(v) - f_\theta'(T'(x_j))(v)||\text{^}2$$

where T and T' are different augmentations, θ is the student parameters, θ' is an exponential moving average teacher, and f_θ'(T'(x_j))(v) is the teacher prediction at location v. Weighting by u_j(v) reduces the influence of ambiguous boundaries and prevents the model from learning agreement on noise.

3.4 Longitudinal Structural Change Modeling

For paired scans acquired at times t and t−1, a structural map s_θ(x) is formed by concatenating layer thickness estimates and lesion probability maps. The previous map is registered to the current scan using an affine transform followed by deformable registration. The change score at location v is

$$\Delta_j(v) = |s_\theta(x_j\text{^}t)(v) - W(s_\theta(x_j\text{^}\{t-1\}))(v)|$$

where x_j^t is the current scan, s_θ(x_j^t)(v) is the current structural map at v, W is the registration operator that maps the previous map into current coordinates, and Δ_j(v) is the absolute structural difference. The change objective L_chg penalizes disagreement between Δ_j and the difference between current and registered segmentation probabilities and adds a total-variation term on the change map to suppress isolated pixel-level fluctuations. This choice reflects a simple clinical principle: true fluid enlargement or layer thinning is spatially coherent, whereas registration error and speckle noise tend to be locally fragmented.

3.5 Composite Training Objective

The full training objective combines supervised segmentation, unlabeled consistency, and longitudinal change consistency:

$$L = L_sup + \alpha L_con + \beta L_chg + \gamma ||\theta||_2\text{^}2$$

where α and β balance the unlabeled and longitudinal terms, γ is weight decay, and L_chg is the temporal consistency term defined above. The three terms have different roles. L_sup anchors the model to expert anatomy and lesion definitions. L_con exposes the model to unlabeled scan variation and prevents overfitting to the labeled subset. L_chg encourages the structural map to change only when the registered anatomy or lesion probability changes. Training should use patient-level sampling so that scans from one eye do not appear in both labeled and unlabeled partitions within the same optimization batch.

4. Detection Logic and Error Modes

The structural change decision at location v can be expressed as

$$D_j(v) = 1[\Delta_j(v) \geq \delta] \cdot 1[u_j(v) \geq \kappa]$$

where δ is a change-magnitude threshold, κ is a reliability threshold, and D_j(v) is the binary change indicator. The two thresholds serve different purposes. δ suppresses differences that are within measurement and registration error. κ suppresses differences in regions where the segmentation model is uncertain. Both thresholds should be selected on a labeled validation set with patient-level separation and should be reported explicitly, because a change detector can appear sensitive merely by lowering δ.

Several failure modes follow directly from the formulation. Confirmation bias can arise when a pseudo-label is wrong but consistent across augmented views. Class-specific thresholds, entropy gating, and teacher-student disagreement reduce this risk but do not eliminate it. Registration error can create false change at layer boundaries and at the optic nerve head. A deformable registration model that is too flexible may warp prior pathology into the current scan, whereas a rigid model may leave genuine anatomical differences. This trade-off must be evaluated on landmarks such as the foveal center and Bruch membrane opening. Signal attenuation and vessel shadowing can mimic fluid or layer loss, so quality-control variables should be recorded and used as covariates in error analysis. Finally, calibration drift across OCT devices can change the meaning of a fixed probability threshold; temperature scaling or ensemble calibration should be checked on device-stratified validation data.

Clinical endpoints should be defined before model evaluation. For fluid, lesion-wise detection and volume agreement are more informative than pixel Dice alone because a small but clinically important pocket can be a small fraction of the image. For layer change, thickness deviation and boundary displacement are appropriate. For drusen and hyperreflective foci, object-level counts and area are appropriate. These endpoints should be linked to recognized disease severity frameworks where relevant [19]. A model that improves pixel overlap without improving lesion-level detection or longitudinal consistency has not necessarily improved clinical utility.

5. Analytical Assessment and Validation Design

A defensible evaluation should compare the proposed framework with supervised-only segmentation, mean-teacher consistency, FixMatch-style pseudo-labeling, and a fully supervised upper bound trained on all available labels. Ablations should remove uncertainty weighting, segmentation-guided pseudo-labeling, longitudinal change consistency, and registration-based differencing one at a time. This design separates the contribution of unlabeled data from the contribution of anatomical structure and temporal modeling. External validation should include a different OCT device or acquisition protocol from the training set; otherwise domain shift is confounded with model capacity. Patient-level splitting is mandatory because multiple B-scans from one eye create leakage if split at the slice level.

Segmentation performance can be reported with Dice and 95th-percentile Hausdorff distance, but detection should be reported with lesion-wise precision, recall, F1, and free-response operating characteristics where multiple lesions per scan are possible. Longitudinal change detection should report sensitivity and specificity at prespecified operating points, together with the area under the receiver operating characteristic curve and the area under the

precision-recall curve. Calibration should be assessed with expected calibration error, Brier score, and reliability diagrams; a model with good discrimination but poor calibration is unsuitable for thresholded clinical alerts. Longitudinal consistency can be measured by test-retest variability on scans without true change and by agreement between predicted change maps and expert adjudication.

Statistical analysis should use bootstrap confidence intervals clustered by patient and paired tests for competing models on the same cases. Reader studies can clarify whether model-assisted review reduces missed lesions or false referrals, but such studies require prespecified reader training and case mix. The framework also invites a mechanism-oriented error analysis: false positives caused by shadowing, false negatives caused by low contrast, and false change caused by registration should be tabulated separately. This decomposition is more informative than a single aggregate error rate because the corrective actions differ. Quality-control monitoring should track input intensity distributions, segmentation confidence, change-score distributions, and device metadata after deployment. These checks are especially important in retinal services where OCT devices and scanning protocols change over time [17], [18].

6. Technical Discussion and Clinical Translation

The proposed framework is appropriate for OCT because it aligns the learning objective with the structure of the clinical problem. A purely supervised segmentation model can achieve high overlap on a curated benchmark but remain vulnerable to small labeled subsets and device shift. A purely unsupervised anomaly detector can flag unusual texture but cannot name the anatomical structure or distinguish fluid from shadow. Image-level weak supervision is also insufficient for structural change because it does not localize the relevant region. Semi-supervised segmentation occupies a useful middle ground: it uses unlabeled scans to learn invariance while retaining voxel-level anatomical targets from the labeled subset. Segmentation guidance then provides a spatial prior for detection, which is more specific than consistency alone.

The cost of this design is additional complexity. The model requires paired scans for the longitudinal branch, a registration method, and an uncertainty estimator. Deep ensembles increase memory and inference cost; Monte Carlo dropout is cheaper but may underrepresent model disagreement. The composite loss introduces three weighting coefficients that can interact. A practical strategy is to pretrain the segmentation branch, freeze the registration operator, and then introduce the longitudinal term after the pseudo-label distribution has stabilized. Thresholds δ and κ should not be tuned on the test set, and any calibration procedure should be reported with the model.

Clinical translation depends less on architectural novelty than on failure transparency. A change alert should be accompanied by the segmented structure, the change score, and a reliability indicator. Clinicians can then distinguish a high-confidence fluid increase from a low-confidence boundary shift. The system should be evaluated as a triage or review aid rather than as an autonomous diagnostic device unless prospective outcome evidence supports otherwise. Because the framework separates segmentation, reliability, and change magnitude, it also supports targeted quality improvement: if errors concentrate in low-reliability regions, additional annotation should focus there; if errors concentrate at registration boundaries, the registration method should be revised.

7. Conclusion

This paper has formulated a semi-supervised approach for detecting retinal structural changes in OCT that integrates segmentation, uncertainty-aware pseudo-labeling, and registered longitudinal differencing. The method uses labeled scans to anchor anatomical and lesion boundaries, unlabeled scans to improve invariance, and reliability weights to limit learning from ambiguous predictions. The change score and decision rule convert segmentation-derived structural maps into localized change indicators with explicit magnitude and confidence thresholds. The framework's main analytical value is that it treats detection as a spatial and temporal inference problem rather than as a byproduct of static segmentation. It also identifies the principal failure modes that must be measured: pseudo-label confirmation bias, registration-induced false change, artifact-driven false positives, and calibration drift across devices. These components provide a technically coherent basis for label-efficient OCT analysis and for evaluating whether apparent structural change reflects retinal biology or imaging and model uncertainty.

References

1. Schmidt-Erfurth, U., Sadeghipour, A., Gerendas, B. S., Waldstein, S. M., & Bogunović, H. (2018). Artificial intelligence in retina. Progress in Retinal and Eye Research, 67, 1–29. https://doi.org/10.1016/j.preteyeres.2018.07.001

2. Schlegl, T., Waldstein, S. M., Bogunović, H., Endstrasser, F., Sadeghipour, A., Philip, A. M., Podkowinski, D., Gerendas, B. S., Langs, G., & Schmidt-Erfurth, U. (2018). Fully automated detection and quantification of macular fluid in OCT using deep learning. Ophthalmology, 125(4), 549–558. https://doi.org/10.1016/j.ophtha.2017.10.031

3. Lee, C. S., Baughman, D. M., & Lee, A. Y. (2017). Deep learning is effective for classifying normal versus age-related macular degeneration OCT images. Ophthalmology Retina, 1(4), 322–327. https://doi.org/10.1016/j.oret.2016.12.009

4. Chiu, S. J., Li, X. T., Nicholas, P., Toth, C. A., Izatt, J. A., & Farsiu, S. (2010). Automatic segmentation of seven retinal layers in SDOCT images congruent with expert manual segmentation. Optics Express, 18(18), 19413–19428. https://doi.org/10.1364/OE.18.019413

5. Ronneberger, O., Fischer, P., & Brox, T. (2015). 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. https://doi.org/10.1007/978-3-319-24574-4_28

6. Tarvainen, A., & Valpola, H. (2017). Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. Advances in Neural Information Processing Systems, 30, 1195–1204.

7. Laine, S., & Aila, T. (2017). Temporal ensembling for semi-supervised learning. International Conference on Learning Representations.

8. Liu, S., Jin, J., Li, W., & Yan, X. (2026). Segmentation-Guided Semi-Supervised Landmark Detection on Intraoral 3D Scans. Preprints. https://doi.org/10.20944/preprints202609.1722.v1

9. Sohn, K., Berthelot, D., Li, C.-L., Zhang, Z., Carlini, N., Cubuk, E. D., Kurakin, A., Zhang, H., & Raffel, C. (2020). FixMatch: Simplifying semi-supervised learning with

consistency and confidence. Advances in Neural Information Processing Systems, 33, 596–608.

10. Xie, Q., Dai, Z., Hovy, E., Luong, M.-T., & Le, Q. V. (2020). Unsupervised data augmentation for consistency training. Advances in Neural Information Processing Systems, 33, 6256–6268.

11. Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., & Raffel, C. (2019). MixMatch: A holistic approach to semi-supervised learning. Advances in Neural Information Processing Systems, 32.

12. van Engelen, J. E., & Hoos, H. H. (2020). A survey on semi-supervised learning. Machine Learning, 109(2), 373–440. https://doi.org/10.1007/s10994-019-05855-6

13. Kendall, A., & Gal, Y. (2017). What uncertainties do we need in Bayesian deep learning for computer vision? Advances in Neural Information Processing Systems, 30, 5574–5584.

14. Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In Proceedings of the 33rd International Conference on Machine Learning (PMLR, Vol. 48, pp. 1050–1059).

15. Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. Advances in Neural Information Processing Systems, 30, 6402–6413.

16. Bogunović, H., Venhuizen, F., Klimscha, S., Apostolopoulos, S., Bab-Hadiashar, A., Bagci, U., Beg, M. F., Bekkers, E., Bhandary, S., Bhavna, R., et al. (2019). RETOUCH: The retinal OCT fluid detection and segmentation benchmark and challenge. IEEE Transactions on Medical Imaging, 38(8), 1858–1874. https://doi.org/10.1109/TMI.2019.2901398

17. De Fauw, J., Ledsam, J. R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., Askham, H., Glocker, B., O'Donoghue, K., Visentin, D., et al. (2018). Clinically applicable deep learning for diagnosis and referral in retinal disease. Nature Medicine, 24(9), 1342–1350. https://doi.org/10.1038/s41591-018-0107-6

18. Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., et al. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA, 316(22), 2402–2410. https://doi.org/10.1001/jama.2016.17216

19. Wilkinson, C. P., Ferris, F. L., Klein, R. E., Lee, P. P., Agardh, C. D., Davis, M., Dills, D., Kampik, A., Pararajasegaram, R., & Verdaguer, J. T. (2003). Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. Ophthalmology, 110(9), 1677–1682. https://doi.org/10.1016/S0161-6420(03)00475-5

20. Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2018). DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(4), 834–848. https://doi.org/10.1109/TPAMI.2017.2699184