Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

TPW-YOLO for microaneurysm detection with tile-phase consistency and workload-aware ranking

chao xu

1275564616@qq.com

The Engineering & Technical College of Chengdu University of Technology

Research Article

Keywords: candidate ranking, diabetic retinopathy, false-positive budget, microaneurysm detection, tile-phase consistency

Posted Date: November 26th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11635050/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

TPW-YOLO for microaneurysm detection with tile-phase consistency and workload-aware ranking

Chao Xu

*Department of Electronic Information and Computer Engineering, The Engineering & Technical College of Chengdu University of Technology, Leshan, 614000, Sichuan, China.

Corresponding author(s). E-mail(s): 1275564616@qq.com;

Abstract

Detecting microaneurysms in retinal images requires accurate localization and careful candidate selection when the number of false positives is limited. TPW-YOLO combines visibility-gated tile-phase consistency learning with residual candidate ranking. During training, the detector aligns predictions for the same annotated lesion across shifted tiles. At inference, the ranker uses evidence from overlapping tiles in a single grid and leaves the candidate boxes unchanged. We evaluated the method on a common image-level partition of three public retinal image sources. On 67 internal test images, sensitivity reached 43.39% at four false positives per image on test-set score-threshold curves, exceeding the matched random-shift baseline by 1.45 percentage points. Across three training seeds at the smaller detector scale, the descriptive mean gain was 4.75 percentage points, with a sample standard deviation of 2.84 points. Candidate ranking added 0.34–0.52 ms per image. Cross-tile supervision and candidate ranking improved lesion recovery at a fixed false-positive budget in this image-level benchmark.

Keywords: candidate ranking, diabetic retinopathy, false-positive budget, microaneurysm detection, tile-phase consistency

1 Introduction

Diabetic retinopathy (DR) is a major cause of vision impairment, and microaneurysms (MAs) are among its earliest visible retinal signs [1, 2]. Locating individual MAs provides evidence that can be inspected in a fundus image. A lesion-level detector must identify these small findings while limiting the spurious regions presented for review.

Deep learning has advanced image-level diabetic retinopathy detection [3–5]. Locating individual microaneurysms is a separate task, for which learned local appearance and dense prediction have been used [6, 7].

An MA may occupy only a few pixels. Downsampling a high-resolution image can therefore remove much of its signal before feature extraction begins. Low contrast and resemblance to vessel fragments or other red lesions make recognition harder. Recent MA-specific detectors use multiscale attention, enhanced feature pyramids, or dilated networks to preserve and distinguish these signals [8–10].

Overlapping tiles preserve local detail without processing an entire image at native resolution [11, 12]. The grid origin, or tile phase, changes a lesion's position and context. Lesions near crop boundaries may be partly visible, yet conventional supervision treats their views independently.

Studies of shift sensitivity [13] and transformation-based consistency [14] motivate comparing paired predictions in a common coordinate frame. We identify pairs by annotation identity and select those with sufficient lesion visibility.

Overlapping tiles yield multiple observations of a candidate, whereas ranking by maximum confidence uses only one. Learning to rank [15] allows evidence from all observations to inform selection at a fixed number of false positives per image (FPI) [2].

We propose TPW-YOLO, which combines tile-phase consistency learning (TPCL) and image-level review-budget-aware ranking (IRBR) with a shared YOLO11 detector [16]. TPCL aligns predictions for the same annotated lesion in original-image coordinates. It selects sufficiently visible observations and uses the more central view as a detached teacher. IRBR learns a residual correction to detector confidence from candidate support and geometry, using image-disjoint out-of-fold (OOF) predictions for training. At inference, it reuses observations from one overlapping grid to change candidate priorities while keeping the boxes fixed.

Our experiments use matched controls to distinguish consistency learning and ranking from paired random-shift augmentation within each detector scale. Three n-scale training seeds test whether the gain from the full pipeline repeats across the evaluated fits. We examine fixed-FPI results alongside AP50 and spatial diagnostics to assess how changes in candidate selection relate to other measures of detection performance.

2 Related work

2.1 Microaneurysm detection and multiscale representation

Early microaneurysm detectors used ensembles [17] or rotating cross-section profiles [18] to characterize local appearance.

Subsequent approaches used two-step convolutional networks [6], early non-proliferative retinopathy analysis [19], or lesion-specific proposal and classification streams [20].

Dense and end-to-end approaches [7, 10] require conversion of responses to individual findings, so comparisons depend on the evaluation target.

Region proposals [21], feature pyramids [22], channel and spatial attention [23], and anchor-free level selection [24] address small-object representation. Lin et al. use scale selection, attention, and a Trident RPN [8]. Tahir et al. use image enhancement, a lightweight pyramid, and attention [9].

IDRiD [25], ROC [2], and e-ophtha [26] provide pixel annotations or center marks. We evaluate the local comparator implementations with the same internal partition and matcher.

2.2 High-resolution tiling and consistency learning

Slicing-aided detection [11] and the specialist detectors of PRISM-DR [12] preserve small lesions through tiling. Repeated predictions can be compared in original-image coordinates.

Ordinary augmentation leaves predictions from related crops independent. Shift sensitivity [13] and transformation-based detection consistency [14] motivate explicit alignment.

Mean Teacher [27], Unbiased Teacher [28], and Soft Teacher [29] select supervision from unlabeled examples. TPCL uses known lesion identities and selects paired targets by visibility.

The shared detector processes both labeled crops, with the more central eligible view supplying a detached target in original-image coordinates.

2.3 Candidate aggregation and lesion ranking

Soft-NMS [30] decays scores using overlap, while weighted boxes fusion [31] combines box coordinates. In overlapping-tile detection, aggregation can also retain the observations that support the chosen candidate.

IoU-Net [32] uses localization confidence to guide selection and refinement. IRBR uses pairwise ranking [15] with matched detections and hard negatives from the same image, while keeping boxes fixed.

IRBR uses image-disjoint out-of-fold predictions to learn from detector confidence, tile support, and geometry.

3 Proposed method

3.1 Architecture and tile construction

TPW-YOLO first trains a shared detector on paired tile views with TPCL, then fits IRBR to OOF candidates with the detectors frozen (Fig. 1). At inference, detections from one overlapping grid are restored to original-image coordinates, aggregated, and ranked. The YOLO11 backbone and feature-fusion neck feed P2–P5 branches at strides 4, 8, 16, and 32. The high-resolution branch preserves detail for small lesions.

We covered native-resolution images with 640 × 640 tiles at 20% overlap and stride 512, clamping border coordinates to cover the full image [11]. Tile phase is the grid-origin offset. Each training epoch paired a zero-offset grid with a shifted grid whose two offsets were sampled independently from integers 0–256. Each tile was paired with the tile having the nearest center in the other grid. Pairing in both directions covered both grids, and cyclic indexing kept the number of sampling slots per image fixed.

A view retained a detection target when at least 30% of the original box area was visible, and the box was clipped to the tile. We kept letterbox transforms and tile origins to recover coordinates. Inputs used RGB normalization, with no additional mosaic, mixup, or affine augmentation. The matched random-shift baseline used the same paired sampling with a consistency weight of zero. Fig. S1 (Online Resource 1) shows the annotation correspondence and tile geometry.

3.2 Tile phase consistency learning

TPCL links predictions for the same lesion across shifted tile views that conventional crop-wise training treats independently (Fig. 1c). Only sufficiently visible observations contribute cross-view targets. Let R_l denote the stored lesion region defined in Section 4.1, and let T_v^{inset} denote tile v after an eight-pixel inward margin. The visibility ratio and eligible lesion set are defined as

$$v_{l,v} = \frac{|R_l \cap T_v^{\text{inset}}|}{|R_l|}; \qquad Q = \{l : v_{l,a} \geq \eta \text{ and } v_{l,b} \geq \eta\}. \tag{1}$$

We set the visibility threshold η to 0.95. Eligible lesions also had to remain detection targets and receive positive assignments in both views. Each observation was represented by its mean assigned class probability and predicted center. We mapped centers to original-image coordinates by reversing the letterbox transforms and adding the tile origin. For each lesion, the view with the closer tile center supplied detached teacher targets, so teacher and student roles could differ between lesions in the same pair.

Fig. S1 shows the same annotated lesion across shifted tile views. Changing the tile origin changes the lesion's position and surrounding context. Annotation identity and coordinate recovery maintain correspondence in the original image. The component mask and enclosing boxes are from a training sample, and the visibility and positive-assignment conditions above determine eligibility for consistency supervision.

$$\begin{aligned} L_{\text{score}} = & -\operatorname{sg}(p_t)\log(p_s) \\ & -[1-\operatorname{sg}(p_t)]\log(1-p_s). \end{aligned} \tag{2}$$

In Equation (2), $\mathrm{p_t}$ and $\mathrm{p_s}$ denote the teacher and student probabilities, respectively, and $\operatorname{sg}(\cdot)$ denotes stop-gradient. Teacher predictions were detached from the computational graph. We bounded the student probability and its complement below by 10^{-6} inside the logarithm. The localization term was the L1 distance between the student and detached teacher centers in original-image coordinates, normalized by the ground-truth box diagonal $\mathrm{d_l}$ plus $\varepsilon = 1$ pixel. The consistency loss averaged the weighted score and localization terms over valid paired lesions:

$$L_{\text{TPCL}} = \frac{1}{|Q^*|}\sum_{l\in Q^*}\left[\alpha L_{\text{score},l} + \beta\frac{\|c_{s,l}-\operatorname{sg}(c_{t,l})\|_1}{d_l+\varepsilon}\right]. \tag{3}$$

Q* contains lesions with valid positive assignments in both views. Batches without valid pairs contributed zero consistency loss. Full TPCL used $\alpha = \beta = 1$. The detection loss L_{det} included box-regression, classification and distribution-focal losses for both views. The first-stage objective combined detection supervision with the consistency term:

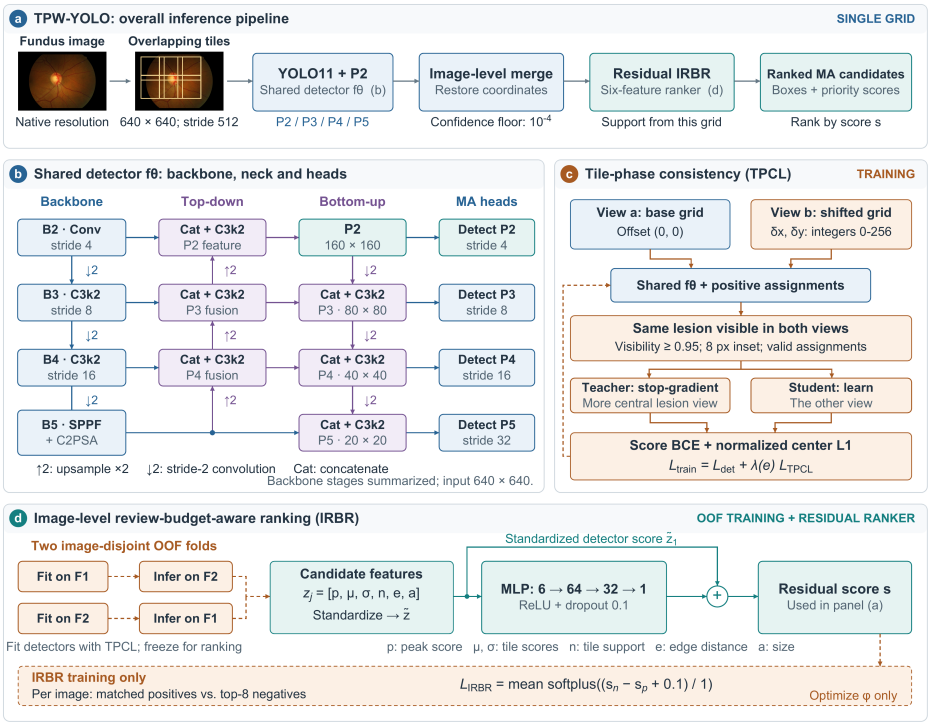

Fig. 1 TPW-YOLO architecture. (a) Single-grid inference, aggregation, and ranking. (b) YOLO11 detector with P2–P5 branches. (c) Visibility and assignment gating with a detached teacher. (d) Residual ranking trained on two image-disjoint out-of-fold pools. Cat denotes concatenation and BCE denotes binary cross-entropy

$$L_{\text{train}} = L_{\text{det}} + \lambda(e) L_{\text{TPCL}};$$
$$\lambda(e) = 0.3 \min(1, \max(0, (e - 15)/10)). \quad (4)$$

The consistency weight was zero for the first 15 epochs, increased linearly over the next 10 epochs, and remained at 0.3 thereafter. This warmup allows lesion assignments to develop before the cross-view constraint reaches full strength. TPCL is used only during detector training, so the paired-view branch adds no computation at inference.

3.3 Candidate aggregation and ranking

We aggregated tile detections before image-level ranking (Fig. 1a). At inference, predictions from a single overlapping grid were mapped to original-image coordinates. A confidence floor of 10^{-4} retained a broad candidate pool. Detections were processed in descending detector-score order and aggregated when their box IoU exceeded 0.5. For tiny boxes, aggregation was also allowed when their centers were less than one pixel apart and both relative width and height differences were below 0.5. The highest-scoring box represented each aggregate. Support statistics used only the maximum confidence from each distinct supporting tile.

Each aggregated candidate j was represented by six features (Fig. 1d): its highest detector confidence p_j, mean confidence μ_j across supporting tiles, confidence standard deviation σ_j, number of distinct supporting tiles n_j, normalized tile-edge distance e_j and normalized size a_j. The edge feature was the minimum distance from the representative center to the boundaries of its generating tile, divided by 320 and clipped to [0,1]. Size was the square root of the box area divided by the larger image dimension. Single-tile candidates had $n_j = 1$ and $\sigma_j = 0$ and were retained as ranking inputs. Support statistics came from overlapping tiles in the same inference grid; no additional phase was required.

$$z_j = [p_j, \mu_j, \sigma_j, n_j, e_j, a_j];$$
$$s_j = \tilde{z}_{j,1} + g_\varphi(\tilde{z}_j). \tag{5}$$

We standardized features using OOF training statistics, flooring standard deviations at 10−6. The network gφ was a 6 → 64 → 32 → 1 multilayer perceptron with ReLU activations and dropout 0.1. Its output corrects the standardized detector-score feature using support and geometry information, while confidence remains the reference score. The resulting real-valued scores determine candidate order, and the aggregated boxes remain fixed.

We divided the 212 training images into two fixed, image-disjoint OOF folds. For each seed, a detector fitted on one fold generated candidates for the other using the same aggregation procedure as at inference. Every ranker-training image thus received predictions from a detector fitted on other images. For evaluation, we paired the ranker with the corresponding A4 detector fitted on all training images. The OOF detectors were used only to generate training candidates.

We labeled candidates with the one-to-one center-matching rule in Section 4.3. For image i, P_i denotes matched lesion candidates and N_i denotes unmatched candidates. The eight negatives with the highest current ranking scores, or all negatives if fewer were available, formed H_i. The within-image ranking objective was

$$L_{\text{IRBR}} = \frac{1}{|P_i|\,|H_i|} \sum_{p \in P_i} \sum_{n \in H_i} \text{softplus}\left(\frac{s_n - s_p + m}{\tau}\right). \tag{6}$$

We used margin m = 0.1 and temperature τ = 1. Images containing both positive and negative candidates were sampled uniformly for 2,000 Adam updates at a learning rate of 10−3. The detector remained frozen, and hard-negative selection focused training on competitive false positives. At inference, IRBR scored all aggregated candidates before thresholding.

4 Experimental setup

4.1 Data and annotations

The dataset contained 279 images with 5,059 lesions from IDRiD [25], the publicly annotated ROC training set [2], and lesion-positive e-ophtha MA images [26]. We used a fixed image-level partition of 212 training images and 67 internal test images (Table 1). IDRiD retained its official 54/27 split. For ROC, a seed-0 NumPy default_rng permutation of sorted names gave a 40/10 split, excluding the hidden-label challenge images. The converted e-ophtha partition contained 118/30 images. All methods used the same partition, and images were split before tile generation.

The training and test partitions contained 12 and one lesion-free images, respectively, all from ROC. The 233 additional lesion-free e-ophtha images were excluded from the dataset.

We derived lesion boxes from eight-connected components. For ROC center marks, we used the annotated radii with a minimum half-width of three pixels. Boxes were clipped and stored in normalized YOLO format. Visibility maps used converted rectangular regions for IDRiD/ROC and component masks for e-ophtha. Images were kept at native resolution before tiling and normalized to [0,1]. All images were public; no new images were acquired.

4.2 Implementation

We ran the experiments with Python 3.11.17, PyTorch 2.9.1/CUDA 12.8, Ultralytics 8.4.174, and an RTX 4070 Ti with 12 GB of memory. Models were trained for 30 epochs using AdamW with a learning rate of 0.002, $\beta_1 = 0.9$, weight decay of $5 \times 10-4$, three warmup epochs, and a final learning-rate factor of 0.01. Each batch contained four pairs/eight views, and mixed precision was disabled. We evaluated the epoch-30 EMA weights. The primary s-scale comparison used seed 0, and the n-scale configurations used seeds 0, 1, and 2.

Both detector scales used depth 0.50. The s and n detectors had widths of 0.50/0.25 and 9,570,676/2,665,796 parameters, respectively. Initialization from YOLO11s transferred 297/593 state entries for s and 58/593 for n. The n transfer comprised 57 compatible scalar entries and a partial transfer of the first convolution. These

Table 1 Dataset composition after annotation conversion

Source	Training images	Training lesions	Test images	Test lesions
IDRiD	54	2412	27	1085
ROC	40	250	10	86
e-ophtha MA	118	952	30	274
Total	212	3614	67	1445

counts refer to state entries, not fractions of model parameters. We kept architecture and initialization fixed within each cohort and analyzed the scales separately.

Primary inference used offset (436, 326) for all configurations and training seeds. We evaluated a separate diagnostic grid at offset (242, 262). Offsets are defined modulo the 512-pixel stride. The fixed inference offsets fall outside the training grid-origin sampling range, and the phase diagnostics characterize the two evaluated grid placements. Within each seed, A4, A6, and A7 shared candidate boxes. IRBR used one grid, with an average of 27.54 tiles/image (P90 54; range 2–54).

4.3 Evaluation

We matched lesions in original-image coordinates, processing ground-truth lesions in their stored order and pairing each with the nearest unused prediction center within 12 pixels. A prediction could match at most one lesion. Matching was repeated at each score threshold, with unmatched and duplicate predictions counted as false positives. Sensitivity was TP/(TP + FN), and FPI was FP divided by the number of original images. The primary endpoint, Sn@4, was sensitivity at an attainable threshold that did not exceed 4 FPI. We also evaluated 0.125, 0.25, 0.5, 1, 2, and 8 FPI, using the mean across all seven operating points as the FROC score. AP50 used score-ordered one-to-one IoU matching at 0.5 with 101-point interpolation. To describe the algorithmic candidate burden, we measured count/image, image-level P90, and the FPI needed to reach at least 50% lesion sensitivity.

All operating points came from test score-threshold curves. Source-specific curves used separate thresholds, whereas pooled evaluation applied one threshold to all 67 images (Tables S13 and S17). Means and sample SDs summarize the three n-scale runs. Image-bootstrap intervals used 2,000 resamples at fixed full-sample thresholds; paired contrasts resampled the same images with seed 710 (Table S7). These pointwise intervals are conditional on the fitted models and test-derived thresholds and were not adjusted for multiple comparisons. We calculated differences from unrounded values. Supplementary Methods define the boundary and phase diagnostics.

4.4 Comparators and component controls

We evaluated four paper-derived implementations: FPN–FSAF [9, 24], the MA stream of Unified [20], SSPNET with Trident RPN [8], and partial PMNPDR [19]. All used the same data and evaluation protocol, a 30-epoch schedule, AdamW learning rate of 0.002, weight decay of $5 \times 10{-}4$, three-epoch warmup, seed 0, and the final EMA checkpoint, with mixed precision disabled. We retained method-specific preprocessing, geometry, losses, and initialization. Supplementary Methods describe the adaptations and sampling exposure. These shared nominal settings did not equalize training cost, and TPW-YOLO required two additional OOF detector fits for ranking. The reported comparisons apply to these implementations.

A3 used matched paired random-shift sampling without consistency. A4 added full TPCL, and A6 then added full-feature IRBR. A7 removed the support features μ, σ, and n while retaining confidence, edge distance, and size. The n cohort also included no-gate, score-only, and localization-only variants, giving seven configurations across three seeds. No-gate TPCL used $\eta = 0$ and zero inset, with the 30% detection rule retained. Rankers used the same OOF pools and optimization. Table S5 includes additional DRN + ROI and PRISM-DR adaptations.

5 Results and Discussion

5.1 Comparative lesion detection

In the primary s-scale experiment, TPW-YOLO reached 43.39% Sn@4, exceeding FPN–FSAF by 7.96 percentage points (paired 95% image-bootstrap interval 5.27–10.74; Table 2). FROC increased from 0.2067 to 0.2635 and AP50 from 25.82% to 39.67% under the same evaluation protocol.

All comparisons use seed 0 and the same internal partition. A3/A4 are matched s-scale controls. Intervals use fixed full-sample cutoffs, and pp denotes percentage points. SSPNET sampled one patch per image per epoch and recorded 119 EMA updates; the partial PMNPDR adaptation omitted curvelet enhancement and differential-evolution optimization (Supplementary Methods).

The full pipeline exceeded all four paper-derived comparators at every evaluated budget from 0.125 to 8 FPI (Fig. 2; Table S14). Sensitivity gains over FPN–FSAF were 3.94 points at 0.125 FPI, 7.96 at 4 FPI, and 10.10 at 8 FPI. FROC was higher for each source, with gains of 0.0244 on IDRiD, 0.0249 on ROC, and 0.1345 on e-ophtha (Tables S6 and S15).

The identical test regions in Fig. S2 show that candidate selection varied by image and that lesions were still missed. In the IDRiD crop, TPW-YOLO selected 10 lesions and FPN-FSAF seven, each with three false positives. All three methods selected seven lesions in the e-ophtha crop, with zero false positives for TPW-YOLO, four for FPN-FSAF, and 13 for SSPNET-Trident. In the ROC crop, TPW-YOLO selected one lesion and FPN-FSAF four, both without false positives.

5.2 Candidate yield and ranking cost

Precision is the proportion of retained candidates matched to lesions at the pooled 4-FPI cutoff. P90 is the image-level 90th percentile. Not reached means that the exported candidate pool did not attain 50% sensitivity.

At 4 FPI, TPW-YOLO selected 627 true lesions and 268 false positives, compared with 512 and 268 for FPN–FSAF, recovering 115 additional lesions at the same false-positive allowance. Selected-candidate precision increased from 65.64% to 70.06%, and the FPI needed for 50% sensitivity fell from 93.69 to 6.60 in the retained pools (Table 3). These results support generating candidates for assisted review. IRBR reused the A4 boxes and one grid, adding 0.52 ms/image in the s run and 0.34–0.36 ms/image in the n runs for CPU scoring (Table S16).

IRBR scoring took less than a millisecond per image, small relative to candidate extraction; TPCL adds computation only during training. The recorded s-scale total was 1,362.12 ms/image, excluding file reading. Tables S3 and S10 report detector and memory measurements.

5.3 Consistency learning and ranking

The seed columns report individual Sn@4 values, and means and sample SDs summarize the three n-scale runs. A6 and A7 use the corresponding A4 boxes. Tables S1 and S2 give per-run intervals and results by source.

The matched A3 control uses the same paired augmentation, so these comparisons measure the added contribution of consistency learning and ranking. In the s cohort, Sn@4 increased from 41.94% for A3 to 42.77% for A4 and 43.39% for A6. The full pipeline gained 1.45 points (paired 95% interval 0.14–2.82). In the n cohort, the descriptive means increased from 39.72% to 43.09% and 44.48%. A6 exceeded A3 by 3.74, 7.96, and 2.56 points across the three evaluated seeds, giving a descriptive mean gain of 4.75 points (across-seed sample SD 2.84 points). Each within-seed paired image-bootstrap interval excluded zero; these intervals are conditional on the fitted models and test-derived thresholds and do not estimate uncertainty in the across-seed mean. The corresponding A6–A4 gains were 1.25, 1.11, and 1.80 points (Fig. 3; Tables S1 and S7).

In Fig. 3A, localization-only consistency reached a mean Sn@4 of 43.41%, compared with 43.09% for full TPCL, 39.56% for score-only consistency, and 39.72% for A3. At this endpoint, the ablations indicate that alignment in original-image coordinates was the main effective consistency component. Full TPCL also exceeded the no-gate variant by 2.35, 1.04, and 0.55 points across the n seeds, supporting the use of visibility to select cross-view targets. TPCL links predictions for the same annotated lesion in a common

Table 2 Comparative detection performance on 67 images and 1,445 lesions

Method	Sn@4 FPI (%) [95% CI]	FROC score	AP50 (%)	A6 minus method (pp) [paired 95% CI]
TPW-YOLO (A6)	43.39 [38.47, 47.41]	0.2635	39.67	Reference
A3 Random shift (s, seed 0)	41.94 [36.86, 46.05]	0.2521	39.57	+1.45 [+0.14, +2.82]
A4 Full TPCL (s, seed 0)	42.77 [38.16, 46.58]	0.2570	41.13	+0.62 [-0.34, +1.51]
FPN + FSAF adaptation [9, 24]	35.43 [32.08, 38.54]	0.2067	25.82	+7.96 [+5.27, +10.74]
Unified MA stream [20]	30.38 [26.62, 34.17]	0.1659	14.36	+13.01 [+9.27, +16.34]
SSPNET + Trident adaptation [8]	7.06 [4.19, 11.13]	0.0464	1.16	+36.33 [+30.01, +41.52]
PMNPDR partial adaptation [19]	2.70 [1.77, 3.89]	0.0165	0.03	+40.69 [+35.93, +44.87]

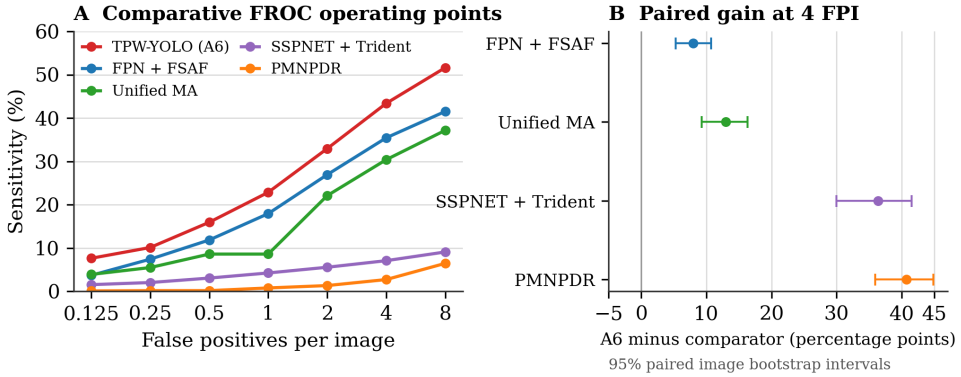

Fig. 2 Lesion sensitivity at seven false-positive budgets (A) and A6 minus each comparator at 4 FPI with paired 95% image-bootstrap intervals at fixed cutoffs (B)

Table 3 Candidate yield at the pooled 4-FPI cutoff and false positives required for 50% sensitivity

Method	Candidates/image at 4 FPI	P90 candidates	Selected candidate precision (%)	FP/image at Sn50
TPW-YOLO (A6)	13.36	33.4	70.06	6.60
FPN + FSAF adaptation [9, 24]	11.64	29.0	65.64	93.69
Unified MA stream [20]	10.55	22.0	62.09	Not reached
SSPNET + Trident adaptation [8]	5.42	12.4	28.10	Not reached
PMNPDR partial adaptation [19]	4.55	13.4	12.79	353.96

coordinate frame, with visibility selection limiting truncated observations and the more central view providing the teacher target. For ranking, the full-feature model exceeded the reduced-feature model by 2.42, 0.83, and 1.04 points in the n seeds and by 0.69 points in s. Both rankers used the same boxes, so these differences isolate the contribution of overlapping-tile support.

At the pooled s cutoffs, IRBR increased true detections from 618 to 627, with false positives changing from 267 to 268 (Table S17). Fig. 3B gives the paired contrasts and intervals within each seed.

Ranking was optimized for the evaluated lesion-selection endpoint. In the s cohort, Sn@4 increased from A4 to A6, but AP50 fell from 41.13% to 39.67%. At source-specific thresholds,

Table 4 Component Sn@4 sensitivity (%) in the separate s and n cohorts

Scale	Configuration	Seed 0	Seed 1	Seed 2	n mean ± SD
s	A3 Random shift	41.94	-	-	-
s	A4 Full TPCL	42.77	-	-	-
s	A6 TPW-YOLO	43.39	-	-	-
s	A7 Reduced features	42.70	-	-	-
n	A3 Random shift	41.45	37.09	40.62	39.72 ± 2.31
n	A4 Full TPCL	43.94	43.94	41.38	43.09 ± 1.48
n	A6 TPW-YOLO	45.19	45.05	43.18	44.48 ± 1.12
n	A7 Reduced features	42.77	44.22	42.15	43.04 ± 1.07
n	TPCL without gate	41.59	42.91	40.83	41.78 ± 1.05
n	Score consistency only	41.04	36.26	41.38	39.56 ± 2.86
n	Location consistency only	44.01	43.94	42.28	43.41 ± 0.98

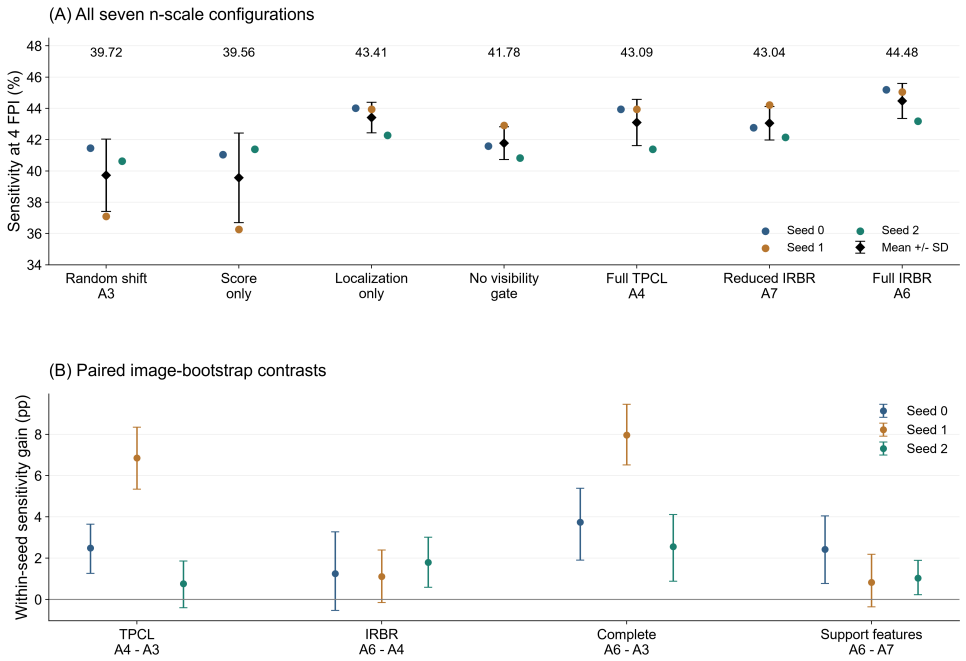

Fig. 3 n-scale results. (A) Sn@4 across three seeds, with mean and sample standard deviation. (B) Within-seed paired 95% image-bootstrap intervals at fixed test-derived cutoffs. A7 excludes support features. A4, A6, and A7 use the same boxes within each seed

e-ophtha Sn@4 decreased after IRBR in the s cohort (56.93% to 55.47%) and all three n seeds (Table S2). Performance also varied at higher recall (Tables S11 and S15), and the s A6–A4 paired interval included zero (Table S7).

5.4 Candidate selection before and after ranking

Fig. S3 compares lesion selection before and after IRBR using the same candidate boxes. In the IDRiD crop, true detections increased from seven for A4 to 10 for A6, and false positives increased from two to three. In e-ophtha, true detections increased from six to seven without false positives. The ROC crop had one fewer selected lesion after ranking. The circled candidates have identical geometry, so their scores and the corresponding operating thresholds determine selection. Across all 67 test images, ranking added nine true detections overall, with false positives changing from 267 to 268 (Table S17).

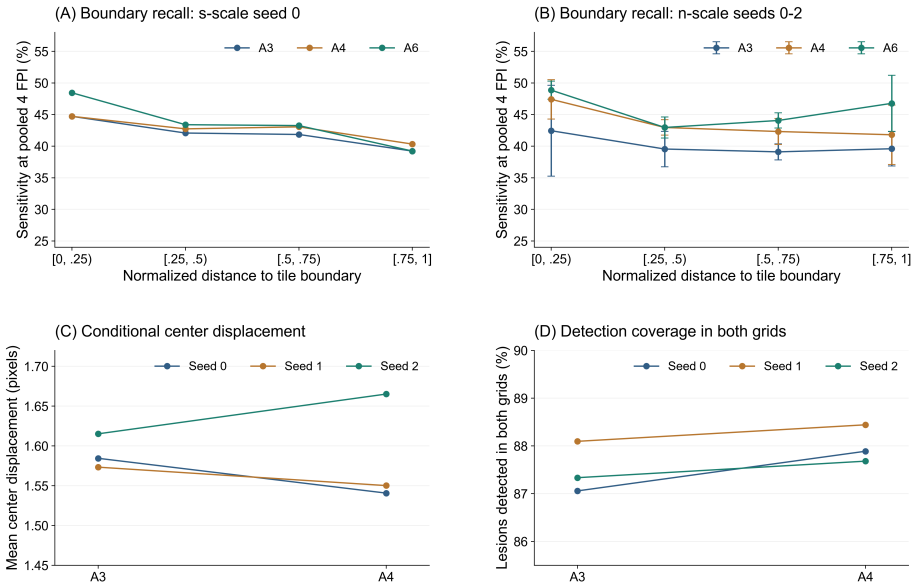

Fig. 4 Boundary sensitivity at pooled 4 FPI for the s run (A) and n runs (B). Boundary distance is the largest center-to-border distance among covering tiles, divided by 320. (C) Mean center displacement for jointly detected lesions. (D) Fraction detected in both grids. The diagnostic grids are evaluated separately

5.5 Boundary and phase diagnostics

Fig. 4A,B groups lesions by their center-to-border distance within covering tiles. In the least-central bin, s-scale sensitivities were 44.72%, 44.72%, and 48.45% for A3, A4, and A6. The corresponding descriptive n-scale means were 42.44%, 47.41%, and 48.86%. Fig. 4C,D reports localization and detection coverage across two independently evaluated grids. Compared with A3, A4 reduced conditional mean center displacement in seeds 0 and 1 and increased it in seed 2. Detection coverage in both grids increased in all three seeds. A6 uses the A4 boxes and preserves their geometry. Tables S4 and S18 give the individual-run measurements in Online Resource 1.

6 Conclusion and Future Work

In this internal benchmark, TPW-YOLO reached 43.39% Sn@4, 1.45 percentage points above the matched s-scale baseline. The three n-scale runs improved by a descriptive mean of 4.75 points (sample SD 2.84 points). Residual ranking reused one grid, kept boxes fixed, and added 0.34–0.52 ms/image.

Future work will examine localization-only TPCL followed by IRBR and evaluate the pipeline using visit-grouped validation, broader negative sampling, and images from unseen sources. These experiments will use independently calibrated operating points and a common setting for runtime comparisons.

Statements and Declarations

Funding. This research received no external funding.

Competing interests. The author declares no conflicts of interest.

Ethics statement. This study used publicly available retinal images from IDRiD, ROC, and e-ophtha. No participants were recruited and no new clinical data were collected.

Data availability. Source images are available from the original dataset providers: IDRiD (https://idrid.grand-challenge.org/), ROC (https://doi.org/10.1109/TMI.2009.2033909), and e-ophtha (https://www.adcis.net/en/third-party/e-ophtha/). Access is subject to the providers' terms.

Author contributions. Chao Xu performed the study and prepared the manuscript as the sole author.

References

[1] Wong TY, Cheung CMG, Larsen M, et al. Diabetic retinopathy. Nat Rev Dis Primers 2016; 2: 16012. https://doi.org/10.1038/nrdp.2016.12

[2] Niemeijer M, van Ginneken B, Cree MJ, et al. Retinopathy Online Challenge: automatic detection of microaneurysms in digital color fundus photographs. IEEE Trans Med Imaging 2010; 29: 185–195. https://doi.org/10.1109/TMI.2009.2033909

[3] Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 2016; 316: 2402–2410. https://doi.org/10.1001/jama.2016.17216

[4] Ting DSW, Cheung CY, Lim G, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. JAMA 2017; 318: 2211–2223. https://doi.org/10.1001/jama.2017.18152

[5] Abràmoff MD, Lavin PT, Birch M, et al. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. npj Digit Med 2018; 1: 39. https://doi.org/10.1038/s41746-018-0040-6

[6] Eftekhari N, Pourreza HR, Masoudi M, et al. Microaneurysm detection in fundus images using a two-step convolutional neural network. Biomed Eng Online 2019; 18: 67. https://doi.org/10.1186/s12938-019-0675-9

[7] Liao Y, Xia H, Song S, et al. Microaneurysm detection in fundus images based on a novel end-to-end convolutional neural network. Biocybern Biomed Eng 2021; 41: 589–604. https://doi.org/10.1016/j.bbe.2021.04.005

[8] Lin J, Liu S, Mao M, et al. Microaneurysm detection with multiscale attention and Trident RPN. Int J Imaging Syst Technol 2025; 35: e70015. https://doi.org/10.1002/ima.70015

[9] Tahir MZ, Lyu X, Nasir M, et al. Advanced image enhancement and a lightweight feature pyramid network for detecting microaneurysms in diabetic retinopathy screening. Int J Imaging Syst Technol 2025; 35: e70004. https://doi.org/10.1002/ima.70004

[10] Asiri N, Artoli AM, Al-Razgan M, et al. Dilated neural networks for improving microaneurysm detection in fundus images. IEEE Access 2025; 13: 104037–104046. https://doi.org/10.1109/ACCESS.2025.3579342

[11] Akyon FC, Altinuc SO and Temizel A. Slicing aided hyper inference and fine-tuning for small object detection. In: Proceedings of the IEEE international conference on image processing, 2022, pp.966–970. https://doi.org/10.1109/ICIP46576.2022.9897990

[12] Özeren Z and Uyar T. PRISM-DR: per-lesion retinal inference with specialist models for diabetic retinopathy. arXiv 2026; 2607.19864. https://doi.org/10.48550/arXiv.2607.19864

[13] Zhang R. Making convolutional networks shift-invariant again. In: Proceedings of the 36th international conference on machine learning. Proc Mach Learn Res 2019; 97: 7324–7334.

[14] Jeong J, Lee S, Kim J, et al. Consistency-based semi-supervised learning for object detection. In: Advances in neural information processing systems, vol.32, 2019, pp.10759–10768.

[15] Burges C, Shaked T, Renshaw E, et al. Learning to rank using gradient descent. In: Proceedings of the 22nd international conference on machine learning, 2005, pp.89–96. https://doi.org/10.1145/1102351.1102363

[16] Ultralytics. Ultralytics YOLO11. Ultralytics package version 8.4.174. https://docs.ultralytics.com/models/yolo11/ (accessed 28 October 2026).

[17] Antal B and Hajdu A. An ensemble-based system for microaneurysm detection and diabetic retinopathy grading. IEEE Trans Biomed Eng 2012; 59: 1720–1726. https://doi.org/10.1109/TBME.2012.2193126

[18] Lázár I and Hajdu A. Retinal microaneurysm detection through local rotating cross-section profile analysis. IEEE Trans Med Imaging 2013; 32: 400–407. https://doi.org/10.1109/TMI.2012.2228665

[19] Qiao L, Zhu Y and Zhou H. Diabetic retinopathy detection using prognosis of microaneurysm and early diagnosis system for non-proliferative diabetic retinopathy based on deep learning algorithms. IEEE Access 2020; 8: 104292–104302. https://doi.org/10.1109/ACCESS.2020.2993937

[20] Asiri N, Hussain M and Al Adel F. A deep learning-based unified framework for red lesions detection on retinal fundus images. arXiv 2021; 2109.05021. https://doi.org/10.48550/arXiv.2109.05021

[21] Ren S, He K, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in neural information processing systems, vol.28, 2015, pp.91–99.

[22] Lin TY, Dollár P, Girshick R, et al. Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp.936–944. https://doi.org/10.1109/CVPR.2017.106

[23] Woo S, Park J, Lee JY, et al. CBAM: convolutional block attention module. In: Computer vision – ECCV 2018. Cham: Springer, 2018, pp.3–19. https://doi.org/10.1007/978-3-030-01234-2_1

[24] Zhu C, He Y and Savvides M. Feature selective anchor-free module for single-shot object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp.840–849. https://doi.org/10.1109/CVPR.2019.00093

[25] Porwal P, Pachade S, Kamble R, et al. Indian diabetic retinopathy image dataset (IDRiD): a database for diabetic retinopathy screening research. Data 2018; 3: 25. https://doi.org/10.3390/data3030025

[26] Decencière E, Cazuguel G, Zhang X, et al. TeleOphta: machine learning and image processing methods for teleophthalmology. IRBM 2013; 34: 196–203. https://doi.org/10.1016/j.irbm.2013.01.010

[27] Tarvainen A and Valpola H. Mean teachers are better role models: weight-averaged consistency targets improve semi-supervised deep learning results. In: Advances in neural information processing systems, vol.30, 2017, pp.1195–1204.

[28] Liu YC, Ma CY, He Z, et al. Unbiased teacher for semi-supervised object detection. In: Proceedings of the international conference on learning representations, 2021. https://openreview.net/forum?id=MJIve1zgR_.

[29] Xu M, Zhang Z, Hu H, et al. End-to-end semi-supervised object detection with soft teacher. In: Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp.3040–3049. https://doi.org/10.1109/ICCV48922.2021.00305

[30] Bodla N, Singh B, Chellappa R, et al. Soft-NMS: improving object detection with one line of code. In: Proceedings of the IEEE international conference on computer vision, 2017, pp.5562–5570. https://doi.org/10.1109/ICCV.2017.593

[31] Solovyev R, Wang W and Gabruseva T. Weighted boxes fusion: ensembling boxes from different object detection models. Image Vis Comput 2021; 107: 104117. https://doi.org/10.1016/j.imavis.2021.104117

[32] Jiang B, Luo R, Mao K, et al. Acquisition of localization confidence for accurate object detection. In: Computer vision – ECCV 2018. Cham: Springer, 2018, pp.816–832. https://doi.org/10.1007/978-3-030-01264-9_48

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ESM1.pdf