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.

Input resolution and decoder design shape Transformer-based retinal lesion segmentation: a controlled multi-seed study with zero-shot external evaluation

Abdulrazak Oladeji Adekunle

abdulrazak_25005993@utp.edu.my

Universiti Teknologi Petronas

Hitham Alhussain

Universiti Teknologi Petronas

Maged Nasser

Universiti Teknologi Petronas

W. Eltayeb Ahmed

Imam Mohammad ibn Saud Islamic University

Research Article

Keywords: Retinal lesion segmentation, Diabetic retinopathy, Input resolution, Vision transformers, External validation, Calibration, Equivalence testing

Posted Date: November 20th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10928310/v1

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

Additional Declarations: No competing interests reported.

Input resolution and decoder design shape Transformer-based retinal lesion segmentation: a controlled multi-seed study with zero-shot external evaluation

Abdulrazak Oladeji Adekunle[1*], Hitham Alhussain[1], Maged Nasser[1], W. Eltayeb Ahmed[2]

[1*]Department of Computing, Universiti Teknologi PETRONAS, Seri Iskandar, 32610, Perak, Malaysia.
[2]Department of Mathematics and Statistics, College of Science, Imam Mohammad Ibn Saud Islamic University, Riyadh, 11432, Saudi Arabia.

*Corresponding author(s). E-mail(s): abdulrazak_25005993@utp.edu.my;
Contributing authors: seddig.alhussain@utp.edu.my;
maged.nasser@utp.edu.my; waahmed@imamu.edu.sa;

Abstract

Background: Input resolution is usually reported as a preprocessing setting in retinal lesion segmentation, yet downsampling compresses small lesion annotations far more than large ones. We asked how resolution, encoder family and decoder design jointly affect segmentation of microaneurysms, haemorrhages, hard exudates and soft exudates, and how much of an apparent architecture ranking survives testing. Because input size also changes token count, positional-embedding interpolation, window configuration and the memory plan, we estimate resolution-conditioned system performance, not an isolated effect of spatial resolution.

Methods: Four ImageNet-pretrained Transformer systems (ViT-Small/16, CaiT-XS24, Swin-Tiny, SwinV2-CR-Tiny) were trained on the FGADR Seg-set at 224, 384, 512 and 768 pixels with five seeds each, under one frozen image-disjoint partition, schedule and scoring pipeline. A shared-final-decoder control and a fixed-budget GroupNorm U-Net reference completed a frozen base of 120 models, evaluated on a held-out FGADR cohort and by zero-shot transfer to IDRiD. Outcomes were pooled pixel-level AUPR and mean average precision over lesion-positive images. Systems were compared under seed-paired t-tests and a

hierarchical seed-and-image bootstrap with Holm correction, and by equivalence testing at margins of 0.01, 0.02 and 0.03.
Results: Between 224 and 768 pixels macro pooled AUPR rose for all four systems on FGADR (0.0794–0.1614) and IDRiD (0.1198–0.2150), but not monotonically: CaiT-XS24 peaked at 512 pixels, a decline absent on FGADR under positive-image aggregation. Most multiplicity-adjusted evidence lay in the first step of the range. Pairwise testing separated systems far more readily under evaluation-cohort than seed resampling; externally at 768 pixels the two agreed on none of ten pairs. The hierarchical systems were equivalent internally at 512 pixels and above at margins of 0.02 and 0.03, but not 0.01. A shared decoder reversed the ViT-Small/16 versus SwinV2-CR-Tiny ordering in all four conditions ($\boldsymbol{p_{\mathrm{Holm}} = 0.004}$), an encoder–decoder interaction. Macro calibration error fell in all eight Transformer conditions (sign test $\boldsymbol{p = 0.008}$), and training cost per epoch rose 11- to 38-fold.
Conclusions: The benefit of resolution is neither universal, nor decoder-independent, nor free, and part of it depends on how performance is aggregated over lesion-negative images. Resolution, encoder and decoder should be chosen together and reported with calibration, aggregation and measured compute.

Keywords: Retinal lesion segmentation, Diabetic retinopathy, Input resolution, Vision transformers, External validation, Calibration, Equivalence testing

1 Background

Retinal lesions are the visible record of vascular and neural injury in diabetic retinopathy (DR). A global meta-analysis estimated that 22.27% of people with diabetes had DR in 2020 and projected that 160.50 million adults could be affected by 2045 [1]. Automated image-level grading may support screening, but a grade alone does not show where the pathology lies. Pixel-level segmentation supplies that spatial evidence, and can therefore support lesion quantification, quality review and more interpretable computer-assisted assessment.

Multi-lesion segmentation is hard because a single fundus photograph can contain abnormalities with very different geometry and appearance. Microaneurysms (MA) are compact and often occupy only a few pixels after resizing; haemorrhages (HE) range from small dots to broad irregular regions; hard exudates (EX) tend to be brighter and more sharply bounded, whereas soft exudates (SE) often have indistinct margins. MA and HE are commonly grouped as red lesions, and weak contrast, small size and confusion with normal vessels make them particularly demanding to detect [2]. Earlier convolutional work showed that the four classes can be segmented jointly [3], and expert-annotated datasets such as FGADR and IDRiD have since made controlled multi-lesion evaluation possible [4–6].

Resolution is usually reported as an implementation setting, although it changes the learning problem itself. When a fundus image and its mask are downsampled, a small connected component may survive as only a handful of cells, thin boundaries can vanish, and neighbouring structures can merge. This loss is not evenly distributed

across lesion classes: it should matter most for compact annotations, while larger regions retain their coarse shape. Patch-based encoders add a second sampling step, because image evidence is summarised inside tokens before any decoding happens. Raising the resolution preserves more spatial detail, but the number of image locations grows quadratically with side length, which increases memory demand and can force changes in batch size or optimisation. A recent multi-resolution study compared 512- and 1024-pixel inputs across convolutional and Transformer-based systems, found lesion-dependent effects rather than a uniform benefit, and proposed a multi-resolution feature stem to exploit them [7]. That work is the closest antecedent to ours, and the difference in aim is worth stating plainly: it compares two resolutions in order to motivate an architecture, whereas we hold the architecture space fixed and treat resolution itself as the object of study, across four scales, four pretrained encoder designs, five seeds per cell and a decoder control.

Transformer encoders do not respond to resolution in the same way. The Vision Transformer (ViT) divides an image into fixed-size patches and processes the resulting token sequence globally [8]; CaiT extends that design with deeper token processing and class attention [9]; Swin Transformer instead builds a hierarchical representation through shifted local windows [10], and Swin Transformer V2 modifies the attention and scaling formulation to improve stability at larger scales [11]. Changing input size therefore alters token counts, feature-map dimensions and window configurations in architecture-specific ways. The encoder is also only part of a segmentation system, because predictions appear after a decoder combines, transforms and upsamples encoder features [12, 13], so a ranking obtained with each Transformer's native decoder cannot be read as an encoder-only comparison. A standardised decoder provides a useful control: if an advantage survives it, the advantage depends less on the original head; if the ordering flips, the result belongs at the system level. External evaluation introduces a second dependence, because cameras, fields of view, preprocessing, disease mix and annotation practice all shift between retinal datasets, so a strict test of transfer requires every model-selection and threshold decision to be fixed on source-domain information before the external labels are examined.

This study examines how input resolution and decoder design shape multi-lesion retinal segmentation within a controlled family of pretrained Transformer systems. Three linked questions are addressed: whether lesion-wise performance changes systematically from 224 to 768 pixels; whether those changes survive zero-shot transfer from FGADR to IDRiD; and whether conclusions about architecture or resolution depend on the decoder paired with the encoder.

The study makes four contributions. First, it measures the resolution-conditioned response of four pretrained Transformer systems under one cohort, one frozen partition, one schedule and five seeds per cell, reporting lesion-specific internal and zero-shot external performance under two aggregation schemes and relating the pattern to annotation geometry at the level of the individual image; that geometric measurement is released as *maskscale*, an open-source diagnostic reporting footprint compression, component fusion and component loss for any segmentation dataset without a model, training or a GPU. Second, it moves beyond a descriptive ranking: every pair of systems is compared under two complementary paired tests, and the pairs that cannot

be separated are identified by equivalence testing across three margins rather than by a failure to reject. Third, it establishes an encoder–decoder interaction through a shared-final-decoder control. Fourth, it treats probability calibration and measured training cost as axes separate from ranking; to our knowledge calibration has not previously been reported as a function of input resolution for retinal lesion segmentation. Together these analyses replace a one-size-fits-all resolution recommendation with evidence specific to lesion class, architecture, decoder, aggregation scheme, evaluation domain and compute budget.

2 Methods

2.1 Study design and frozen experimental matrix

We designed a controlled, multi-seed study in which resolution was treated as a property of the complete segmentation system rather than a neutral preprocessing setting. Changing the input side length does not change one quantity: token count, positional-embedding interpolation, window configuration, feature-map size, the frozen micro-batch and the use of gradient checkpointing all move with it (Table 1). The estimand throughout is therefore *resolution-conditioned system performance* – what a complete pretrained system does when operated at a given input size – rather than the isolated causal effect of spatial resolution, and we use that term in place of any phrasing that would attribute the whole response to preserved lesion detail. Figure 1 summarises the design and the order in which decisions were frozen.

Table 1 What co-varies with input resolution in this design

Quantity	Moves with resolution	Held fixed
Token or patch count	Yes, quadratically	—
Positional embedding	Yes, interpolated to the active grid	—
Window configuration (Swin, SwinV2)	Yes, windows per feature map	Window size
Feature-map dimensions	Yes, at every stage	Channel widths
Micro-batch size	Yes, 1–8 under the memory plan	Effective batch (8)
Gradient checkpointing	Yes, where memory required it	—
Peak device memory	Yes	Device class (one Tesla T4)
Annotation footprint	Yes, $M_R = AR^2/HW$	Native annotation
Encoder weights, initialisation	No	Identical across resolutions
Optimiser, schedule, loss, seeds	No	Identical across resolutions
Partition, scoring pipeline	No	Frozen once

The study estimates the joint effect of the quantities in the first block. It cannot attribute the response to spatial resolution alone, and no such attribution is made.

The principal experiment crossed four Transformer-based segmentation systems with four square input resolutions (224, 384, 512 and 768 pixels) and five seeds, giving 80 models: two columnar encoders, ViT-Small/16 and CaiT-XS24, and two hierarchical encoders, Swin-Tiny and SwinV2-CR-Tiny [8–11], each retaining its native decoder.

Two controls were added. ViT-Small/16 and SwinV2-CR-Tiny were retrained at 224 and 768 pixels behind the same final-feature decoder, giving 20 shared-final runs; this matches the decoder layer pattern but not encoder capacity, feature stride, patch size or pretraining, so it tests decoder dependence rather than isolating an effect of encoder hierarchy. A GroupNorm U-Net trained from scratch at all four resolutions and five seeds added 20 convolutional runs, for 120 trained source models in total. All runs shared the same audited cohort, frozen partition, supervision policy, effective batch size, loss, validation schedule, checkpoint rule and native-resolution scoring pipeline. The partition was generated once and was not reselected after inspecting performance.

The U-Net arm was deliberately interpreted as a fixed-budget reference. A run was flagged as potentially budget-limited when it completed at least 90% of the 40-epoch budget and either selected its best checkpoint within the final two tuning intervals or improved by more than 10^{-4} AUPR per epoch across its final three tuning assessments. Eleven of 20 U-Net runs met this frozen rule and no Transformer run did, so U-Net results describe resolution response and do not support a claim of absolute Transformer superiority.

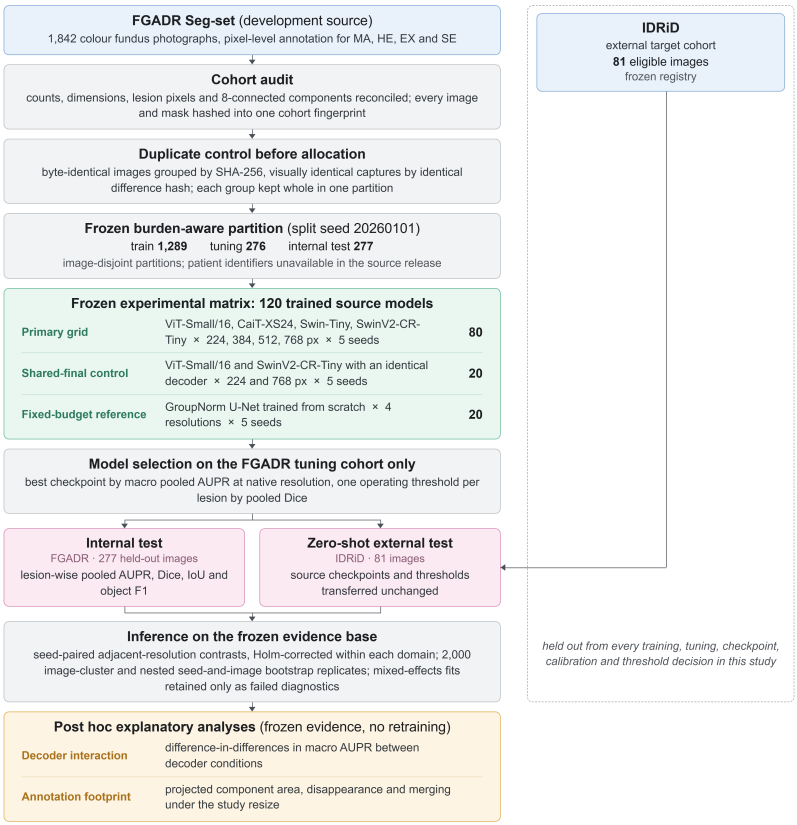


Fig. 1 Study design.

2.2 Datasets, cohort construction and leakage control

The Fine-Grained Annotated Diabetic Retinopathy (FGADR) Seg-set was used for model development and internal testing [4]. Its 1,842 colour fundus photographs were acquired predominantly at hospitals in the United Arab Emirates and carry pixel-level annotation for six lesion types: microaneurysms, haemorrhages, hard exudates, soft exudates, intraretinal microvascular abnormalities and neovascularisation. This study used the four classes shared with the external dataset (MA, HE, EX, SE), and a single network predicted all four probability maps simultaneously; intraretinal microvascular abnormalities and neovascularisation were outside its scope. The extraction was audited against the expected cohort and cryptographically hashed into a fingerprint carried into every downstream run record before partitioning.

Duplicate control ran before allocation: byte-identical images were grouped by SHA-256 and visually identical captures through connected components of identical 64-bit difference hashes at Hamming distance zero, with every group assigned to one partition in full. This establishes an exact-duplicate-disjoint and identical-dHash-disjoint partition, but not patient-level separation, because patient identifiers were unavailable; we therefore use the term *image-disjoint* throughout. The split was generated once with a fixed seed at fractions of 70%, 15% and 15%, giving 1,289 training, 276 tuning and 277 internal-test images, and allocation was group-complete and balanced on per-lesion presence and burden. The manifest was frozen and hashed before training began.

The Indian Diabetic Retinopathy Image Dataset (IDRiD), released for the ISBI-2018 segmentation and grading challenge and the first such resource representative of an Indian population, was reserved for external evaluation [5, 6]. All 81 eligible images from the official segmentation partitions were placed in a frozen target registry mapped to the same four lesion classes; the optic-disc annotations distributed with that subset were not used. The two cohorts therefore differ in source population and acquisition setting, so the external stage tests transfer across a real domain shift rather than across a partition boundary. No IDRiD image contributed to training, tuning, checkpoint selection, threshold selection or hyperparameter adaptation; the same source checkpoint and FGADR-selected thresholds were applied without target feedback, so we describe this stage as *zero-shot external evaluation* rather than external validation.

2.3 Preprocessing and segmentation systems

Fundus photographs were converted to CIELAB, contrast-limited adaptive histogram equalisation (CLAHE) was applied to the luminance channel with a clip limit of 2.0 and an 8×8 tile grid [14], and the image was converted back to RGB. Each image was resized directly to $R \times R$ pixels, $R \in \{224, 384, 512, 768\}$, by area interpolation with no letterboxing, padding or field-of-view mask, and normalised with ImageNet channel statistics identically for every system. Binary lesion masks followed the same pathway and retained fractional lesion coverage rather than a nearest-neighbour label, which reduces uncontrolled deletion of small annotations while preserving the footprint loss that downsampling causes. Training augmentation was geometric and paired

between image and masks: independent flips and rotation by a multiple of 90 degrees. Evaluation images were not augmented.

The columnar systems used patch-16 ViT-Small and CaiT-XS24 encoders, collecting features at four depths and fusing them through a common convolutional decoder that projects each same-stride map to 128 channels, concatenates and fuses them, then upsamples in four 3×3 convolution–GroupNorm–GELU stages to a 1×1 four-lesion head. Swin-Tiny and SwinV2-CR-Tiny emitted hierarchical maps at nominal strides 4 to 32, decoded by a feature-pyramid path into the same fusion and head. The shared-final control accepted only the encoder's final spatial feature map and applied an identical layer sequence for both encoders. The conventional reference was a four-level U-Net [12]. GroupNorm was used throughout because it does not depend on batch-level running statistics [15]. Layer-level specifications are in Additional file 1.

Transformer encoders were initialised from public ImageNet-pretrained weights distributed through `timm` [16] and fine-tuned end to end, with initialisation-state hashes invariant across resolutions, seeds and decoder conditions. The nominal pretrained geometries were 224 pixels for ViT-Small, Swin-Tiny and SwinV2-CR-Tiny and 384 pixels for CaiT-XS24, so the experiment estimates the response of each complete pretrained system to input size and does not isolate resolution from positional-embedding adaptation or window configuration. The U-Net was initialised from scratch, with one initial state per seed shared across all four resolutions.

2.4 Training, model selection and operating thresholds

All models were optimised with AdamW at an initial learning rate of 1×10^{-4} and decoupled weight decay of 0.05 [17], with a cosine schedule stepped once per epoch, $T_{\max} = 40$ and $\eta_{\min} = 0$ [18], and no warm-up, restart, layer-wise rate or gradient clipping. The objective was the unweighted sum of binary cross-entropy with logits and a per-sample, per-lesion Focal–Tversky loss [19],

$$\mathrm{TI} = \frac{\mathrm{TP} + \varepsilon}{\mathrm{TP} + \alpha\mathrm{FN} + \beta\mathrm{FP} + \varepsilon}, \qquad \mathcal{L}_{\mathrm{FT}} = (1 - \mathrm{TI})^{\gamma}, \tag{1}$$

with $\alpha = 0.7$, $\beta = 0.3$, $\gamma = 0.75$ and $\varepsilon = 1$, averaged over sample–lesion pairs so that the objective is invariant to the micro-batch partition required at higher resolutions. The effective batch size was eight in every run, formed by sample-weighted gradient accumulation over frozen, resolution-dependent micro-batches; mixed precision was enabled and gradient checkpointing used only where the frozen memory plan required it.

Models were trained for up to 40 epochs, with the tuning cohort evaluated every two epochs on macro pooled AUPR after restoring predictions to native resolution; training stopped after four consecutive checks without an improvement exceeding 10^{-4}, subject to a 10-epoch minimum, and the state with the highest tuning macro pooled AUPR was restored. One operating threshold per lesion was then chosen on the FGADR tuning cohort as the value maximising pooled Dice over 201 equally spaced thresholds, and transferred unchanged to IDRiD. These thresholds affected Dice, intersection-over-union, object measures and false-positive counts, but not AUPR.

2.5 Outcome measures

The primary measure was lesion-specific pooled pixel-level AUPR. Predictions and labels from every cohort image, including lesion-negative images, contributed to one cohort-level estimator per lesion, computed from 2,048-bin score histograms as an approximation to average precision. Macro AUPR is the unweighted mean over MA, HE, EX and SE.

Pooled AUPR is influenced by lesion prevalence and by false positives on lesion-negative images, so a second aggregation is reported alongside it: mean average precision computed per image over images containing at least one annotated component of that lesion, then averaged over images, seeds and lesions. The two answer different questions. Pooled AUPR describes behaviour over a whole screening cohort, negatives included; positive-image mAP describes behaviour where the lesion is present. Both are reported wherever they diverge.

Score histograms used 2,048 bins. Re-binning the retained histograms at 256, 512, 1,024 and 2,048 bins shows that coarser binning biases pooled AUPR downward and that the estimator has not fully converged at 2,048, so the absolute values reported here are slightly conservative. What the study reports is stable: across an eightfold change in bin count the endpoint macro gains move by at most 0.012, the ordering of systems at 768 pixels never changes, and the difference between the two hierarchical systems moves by at most 0.002, well inside the equivalence margin.

No field-of-view mask was applied at any stage, so all pixel-level measures – the calibration terms and the whole-cohort false-positive counts included – are computed over the full image, retinal background included.

Before scoring, output logits were bilinearly restored to the native dimensions of the reference mask, so that lower-input models were evaluated against the original annotations rather than against an equally degraded ground truth.

Threshold-dependent measures were Dice and intersection-over-union averaged over lesion-positive images, with whole-cohort false positives reported alongside them. For object-level evaluation, masks were decomposed into 8-connected components and matched by maximum-cardinality, maximum-score one-to-one assignment solved with the Hungarian algorithm [20], a primary match requiring reference and predicted coverage each to be at least 0.10. Free-response receiver-operating-characteristic (FROC) analysis was run for the 20 endpoint models formed by ViT-Small/16 and SwinV2-CR-Tiny at 224 and 768 pixels in both cohorts, summarised by the competition performance metric (CPM). Matching sensitivity analyses, the FROC protocol and the full curves are in Additional file 1.

2.6 Statistical analysis

The completed design was audited before inference: the result tables had to contain exactly 320, 80 and 80 lesion-level rows for the primary, shared-final and U-Net arms, and any missing, duplicate, non-finite or truncated record caused the audit to fail. Cohort, manifest, checkpoint and per-image evidence hashes were reconciled between source and external records.

Across seeds, results were summarised by the mean, standard deviation and t-based 95% confidence interval, and adjacent-resolution comparisons were paired by seed. Within each architecture family and lesion, values from the two representative backbones were first averaged within seed to give a family-level estimand, with model-specific trajectories retained and reported separately. Holm's procedure controlled multiplicity [21], with FGADR and IDRiD corrected separately and each principal-domain family containing 24 tests defined by two architecture families, four lesions and three adjacent steps. Image-sampling uncertainty was estimated by resampling whole images, never individual pixels [22], summing per-image sufficient statistics and recomputing the relevant pooled AUPR. Condition contrasts used 2,000 nested paired bootstrap replicates in which seeds were resampled as training-process replicates and, within a selected seed, the same ordered image draw was applied to both conditions. This is a hierarchical bootstrap over the two levels at which the design varies, seeds and images, and it is the primary uncertainty statement wherever the five-seed t-test has little resolution.

Adjacent-step contrasts answer whether a system improves with resolution, not whether one system is better than another. We therefore compared every pair of the five systems at every resolution in each cohort under two complementary paired tests that isolate different sources of uncertainty: a seed-paired t-test on macro pooled AUPR, which quantifies variability of the training process, and a paired image-cluster bootstrap in which whole images were resampled with a single weight matrix of 2,000 multinomial draws shared by every condition, which quantifies variability of the finite evaluation cohort. Each cohort formed one Holm family of 40 tests per test type – all ten pairs of the five systems at each of four resolutions – and a pair was called separated only where both tests agreed. Because failing to reject a null hypothesis is not evidence of equivalence, particularly with five seeds, every pair was also submitted to two one-sided tests against a pre-specified margin of $\delta = 0.02$ macro pooled AUPR. That value was chosen before testing on a substantive basis: it is smaller than the macro gain of one resolution step for the hierarchical systems between 512 and 768 pixels, the smallest difference this design would act on, and it is of the same order as the seed-to-seed standard deviation of macro AUPR within a condition, so a smaller difference is not reliably reproducible by retraining. Because the conclusion depends on that choice, every equivalence test was run at $\delta = 0.01$ and $\delta = 0.03$ as well, and all three margins are reported. Equivalence was declared where the 90% confidence interval of the paired difference fell entirely inside $(-\delta, +\delta)$ on both the seed-paired differences and the image-level bootstrap distribution.

Pooled AUPR is invariant to any monotone transformation of the predicted scores, so it cannot say whether the predicted probabilities mean what they claim. Using the same 2,048-bin histograms we computed the Brier score over all pixels and decomposed it in the standard way,

$$\mathrm{BS} = \underbrace{\sum_b \frac{n_b}{N} \left(p_b - \bar{y}_b\right)^2}_{\text{calibration}} - \underbrace{\sum_b \frac{n_b}{N} \left(\bar{y}_b - \bar{y}\right)^2}_{\text{discrimination}} + \bar{y}\left(1 - \bar{y}\right), \tag{2}$$

where n_b is the number of pixels in score bin b, p_b the bin centre, $\bar{y}_b$ the observed positive frequency in that bin and $\bar{y}$ the lesion prevalence; we also report the expected calibration error $\sum_b (n_b/N)|\bar{y}_b - p_b|$. Smaller calibration and larger discrimination terms are better, and to avoid confusion with input resolution we use *discrimination* for the second term throughout. Mixed-effects models were examined only as diagnostics. A fitted model counted as trustworthy only if it met all four pre-specified criteria: the optimiser converged; no random-effect variance was estimated at or near the boundary and no random-effects covariance matrix was singular; the Hessian was positive definite and yielded finite standard errors; and any clustered fallback had at least ten seed clusters. All six fitted diagnostics failed at least one criterion, in every case a boundary or singular random-effect estimate, which is the expected behaviour with five seed clusters, so their p-values were excluded; inferential emphasis rests on effect magnitude, multiplicity-adjusted paired contrasts and the bootstrap intervals.

2.7 Post hoc explanatory analyses

Three analyses were performed after the primary results had been inspected. None belonged to the confirmatory multiplicity families and none changed the frozen 120-run evidence base.

The descriptive decoder rank reversal motivated a direct interaction test. For ViT-Small/16 and SwinV2-CR-Tiny at 224 and 768 pixels in both cohorts, the estimand was

$$D = (\mathrm{AUPR}_{\mathrm{ViT}} - \mathrm{AUPR}_{\mathrm{SwinV2}})_{\mathrm{shared}} - (\mathrm{AUPR}_{\mathrm{ViT}} - \mathrm{AUPR}_{\mathrm{SwinV2}})_{\mathrm{native}}\,, \tag{3}$$

with positive D indicating that the shared-final decoder shifts relative performance toward ViT. All four conditions were resampled jointly with the nested seed-and-image procedure; the four macro tests formed one Holm family and the 16 lesion-specific effects a separate secondary family.

The second analysis asked whether the lesion-scale interpretation is geometrically plausible without training another model. Every native FGADR mask was passed through the exact area-interpolation pathway used for training, with no checkpoint loaded and no inference performed. For an 8-connected annotation component of area A pixels in an $H \times W$ mask, its projected target-pixel-equivalent footprint at resolution R is

$$M_R = \frac{AR^2}{HW}. \tag{4}$$

For each lesion and resolution we estimated the proportion of components with M_R below 1, 4 and 9 pixels, and examined the resized soft masks at occupancy thresholds from 0.01 to 0.50; where several native components mapped to one target component the event was recorded as a merger, which measures loss of instance separability. These quantities describe annotation-footprint retention, not retinal contrast, biological visibility or a causal network mechanism. They were packaged as *maskscale*, an open-source command-line diagnostic that reports projected-footprint

fractions, component-fusion rates, component-loss rates and retained mask mass for any directory of binary masks, with no model or GPU required.

The third analysis repeated the geometric comparison at the level of the individual image, where the proposed explanation can fail. For every internal-test image with at least one annotated component of a given lesion we computed average precision at 224 and 768 pixels from that image's own score histogram, averaged over seeds, and joined the difference to that image's annotation geometry at 224 pixels, summarising associations by Spearman rank correlation. We also asked whether the footprint statistics predict which classes a resolution increase will help. Letting $\varphi_c(R)$ be the fraction of class c's components projecting below four target pixels and $\Delta\varphi_c$ the reduction in that fraction across a step, classes were ranked by $\Delta\varphi$ within each step and the ranking compared with realised gain on an absolute and a relative scale. Because $\Delta\varphi$ depends only on lesion and step, its ordering is identical across systems and cohorts and the rankings are not independent replicates; we therefore used an exact permutation test that relabels classes within each step and recomputes every ranking.

2.8 Reproducibility and software

Each run used a single NVIDIA Tesla T4 GPU with DataParallel disabled, so the wall-clock times reported below are directly comparable across conditions. All generators were seeded and cuDNN benchmarking was disabled; strict deterministic algorithms were not forced, so the study is seeded rather than bitwise deterministic. Every training record embedded the same source-tree and code-commit identity, final collation verified all evidence hashes, and the post hoc package reproduced the frozen decoder point estimates to within 1.86×10^{-9}. Software-library versions, pretrained-weight identifiers and initialisation hashes, the code commit and source-tree hash, the split seed, the audited-cohort fingerprint, and the resolution-dependent memory plan are reported in Additional file 1

3 Results

3.1 Lesion targets, annotation geometry and completeness of the evidence base

The four target classes differ in prevalence and in scale by more than an order of magnitude (Table 2). Hard exudates contributed the largest number of annotation components, 51,824 across 1,279 positive images, yet their median native component area, 79 pixels, was the smallest of the four classes. Microaneurysms were similar in scale, with a median of 81 pixels, but an order of magnitude less numerous. Soft exudates sat at the opposite extreme: 2,080 components with a median area of 1,340 pixels, present in only 627 images. This matters for what follows, because the two classes that respond most strongly to resolution are the two with the smallest median component area, and microaneurysms are not uniquely small among them.

All 120 planned models were completed and retained, and the lesion-level result tables contained exactly 320, 80 and 80 rows for the primary, shared-final and U-Net arms, with all 160 expected run-by-cohort evidence records and no missing, duplicate,

non-finite or unexpected design cell. All 216 nested seed-and-image contrasts reconciled with their paired run-level estimates to within 1.126×10^{-9}. None of the 100 Transformer runs met the frozen convergence-flag rule; 11 of the 20 U-Net runs did.

Table 2 Lesion targets in the FGADR development cohort and their native annotation geometry

Lesion	Positive images (of 1,842)	Annotation components	Median native area (pixels)	Median projected area at 224 px (target pixels)	Components below 4 px at 224 px (%)
Microaneurysms (MA)	1,424	11,221	81	2.48	69.7
Haemorrhages (HE)	1,456	37,812	182	5.57	40.3
Hard exudates (EX)	1,279	51,824	79	2.42	70.6
Soft exudates (SE)	627	2,080	1,340	41.04	7.9

Components are 8-connected regions of the native annotation masks and are analytical units rather than independently adjudicated clinical lesions. Projected area is the target-pixel-equivalent footprint of Eq. (4).

3.2 Resolution response on the internal and zero-shot external cohorts

Higher input resolution improved macro pooled AUPR for all four native Transformer systems across the full 224 → 768 comparison in both cohorts (Fig. 2, Table 3). Internally the gain ranged from 0.0794 to 0.1614 and was largest for ViT-Small/16, which rose from 0.1699 to 0.3313; Swin-Tiny rose from 0.3166 to 0.3960 and SwinV2-CR-Tiny from 0.3050 to 0.4033. The response was not uniformly monotonic: CaiT-XS24 reached 0.3004 at 512 pixels and then fell to 0.2875, although the 768-pixel value remained 0.1043 above its 224-pixel value.

The pattern transferred to IDRiD without any target-dataset tuning. External macro pooled AUPR increased by 0.2150 for ViT-Small/16, 0.1198 for CaiT-XS24, 0.1338 for Swin-Tiny and 0.1577 for SwinV2-CR-Tiny, and each endpoint gain exceeded its FGADR counterpart. CaiT-XS24 again peaked at 512 pixels while the other three reached their highest means at 768 pixels, where the two native hierarchical systems held the highest means in both cohorts. These values describe complete encoder–decoder systems and do not, by themselves, isolate an encoder-hierarchy effect.

Aggregating over lesion-positive images only changes two of these readings (Table 4). The resolution response itself is robust: every system still improves from 224 to 768 pixels, and by more under positive-image aggregation than under pooled AUPR. But CaiT-XS24's decline at 768 pixels disappears on FGADR, where positive-image mAP rises throughout (0.1892, 0.3002, 0.3290, 0.3440), so the internal non-monotonicity is a pooled-AUPR phenomenon driven by false positives on lesion-negative images; the decline does persist externally. More consequentially, the fixed-budget U-Net moves from third place under pooled AUPR at 768 pixels to first

under positive-image mAP, ahead of both hierarchical Transformers. Across all conditions the two aggregations correlate at Spearman $\rho = +0.81$, so they are related but not interchangeable. The divergence has a coherent explanation, developed in the calibration and discussion sections: the U-Net is the worst-calibrated system in the study and over-predicts, which costs it most on the negative images that pooled AUPR counts and positive-image mAP does not.

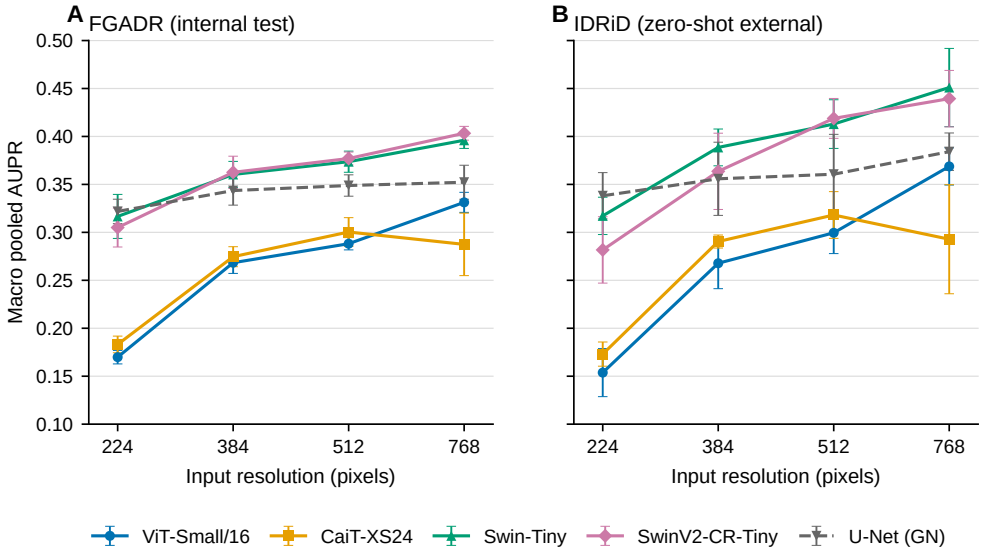


Fig. 2 Macro pooled AUPR across input resolutions. a FGADR internal-test results. **b** Zero-shot external results on IDRiD. Points are means across five training seeds and error bars are t-based 95% confidence intervals. Macro AUPR is the unweighted mean of the four lesion-specific pooled AUPRs. The GroupNorm U-Net (grey, dashed) is a fixed-budget reference: 11 of its 20 runs met the frozen convergence flag, so it is shown to describe resolution response and not to rank architectures.

3.3 Where the gain accrues, and which lesions respond

Endpoint differences hide where the gain occurs. Testing the three adjacent-resolution steps separately places most of the effect at the low end of the range. Of the 24 Holm-corrected family-level contrasts in each domain, 8 reached significance internally and 12 externally, and the first step accounted for most of them: 6 of 8 lesion-by-family contrasts at 224 $\rightarrow$ 384 were significant on FGADR and 7 of 8 on IDRiD. Beyond 512 pixels only one FGADR contrast survived correction (microaneurysms, hierarchical family, $p_{\text{Holm}} = 0.046$) and only two did so externally (microaneurysms, $p_{\text{Holm}} = 0.011$; hard exudates, $p_{\text{Holm}} = 0.021$). Moving from 224 to 384 pixels therefore changed lesion-wise performance almost everywhere; moving from 512 to 768 pixels changed it in specific lesion-by-architecture combinations and nowhere else. Full step-wise effects are in Additional file 1, Fig. S1.

Table 3 Macro pooled AUPR by dataset, native Transformer system and input resolution

Dataset	System	224 px	384 px	512 px	768 px
FGADR	ViT-Small/16	0.1699	0.2683	0.2882	**0.3313**
FGADR	CaiT-XS24	0.1831	0.2746	**0.3004**	0.2875
FGADR	Swin-Tiny	0.3166	0.3604	0.3737	**0.3960**
FGADR	SwinV2-CR-Tiny	0.3050	0.3626	0.3770	**0.4033**
IDRiD	ViT-Small/16	0.1537	0.2678	0.2995	**0.3687**
IDRiD	CaiT-XS24	0.1731	0.2904	**0.3181**	0.2928
IDRiD	Swin-Tiny	0.3172	0.3886	0.4129	**0.4510**
IDRiD	SwinV2-CR-Tiny	0.2817	0.3636	0.4188	**0.4394**

Values are five-seed means. Bold type marks the highest mean within each dataset–system row and is descriptive only. Complete lesion-wise values with standard deviations are in Additional file 1.

Table 4 Macro performance under two aggregation schemes, FGADR internal test

System	Pooled AUPR		Positive-image mAP	
	224 px	768 px	224 px	768 px
ViT-Small/16	0.1699	0.3313	0.1725	0.3717
CaiT-XS24	0.1831	0.2875	0.1892	0.3440
Swin-Tiny	0.3166	0.3960	0.3240	0.4625
SwinV2-CR-Tiny	0.3050	**0.4033**	0.3264	0.4697
U-Net (GN)	0.3218	0.3522	0.4075	**0.4804**

Five-seed means. Bold marks the highest value in each 768-pixel column. Pooled AUPR includes lesion-negative images; positive-image mAP averages per-image average precision over images containing the lesion. The two orderings differ at the top, and the U-Net is the system most sensitive to the choice. Both cohorts at all four resolutions are in Additional file 1.

The response was most consistent for microaneurysms and hard exudates (Fig. 3). Across the 12 adjacent comparisons per lesion, nested seed-and-image intervals excluded zero in 10 comparisons for MA and 11 for EX on FGADR, against 9 for HE and 3 for SE. On FGADR, MA pooled AUPR rose from 0.0081 to 0.0665 for ViT-Small/16, from 0.0458 to 0.1232 for Swin-Tiny and from 0.0481 to 0.1243 for SwinV2-CR-Tiny, while CaiT-XS24 peaked at 0.0576 at 512 pixels before falling to 0.0468. Soft exudates moved at the first step and then plateaued, and for Swin-Tiny hardly moved at all (0.2859, 0.3082, 0.3062 and 0.3010 across the four resolutions), which is consistent with their much larger native footprint. Externally, microaneurysms showed the most sustained adjacent-resolution support, with Holm-adjusted p-values of 0.0007, 0.0050 and 0.0107 across the three steps for the hierarchical family, where internally only the last step met the same corrected threshold.

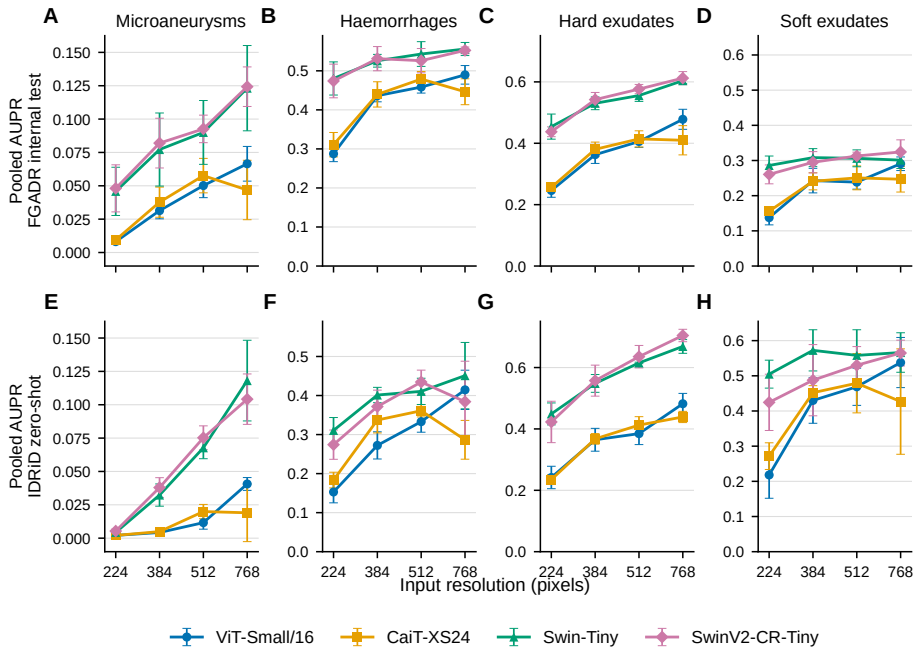


Fig. 3 Lesion-specific resolution response for the four native Transformer systems. a–d FGADR internal test. **e–h** IDRiD zero-shot external evaluation. Points are five-seed means with *t*-based 95% confidence intervals.

3.4 Annotation geometry accounts for part of the response, but does not predict it

The post hoc footprint analysis supplies a geometric boundary for these results (Fig. 4). At 224 pixels, 69.7% of MA components and 70.6% of EX components projected to fewer than four target pixels, against 40.3% for HE and 7.9% for SE; at 768 pixels the proportions fell to 2.8%, 4.8%, 1.5% and 1.0%. Two components of that loss behave differently. Aggregate soft-target mass was retained with high numerical accuracy even at the lowest resolution, with detectable EX soft-mass retention rising only from 0.9922 to 0.9989 between 224 and 768 pixels. Instance separability changed far more: at an occupancy threshold of 0.10 the proportion of EX components involved in a merger fell from 58.2% at 224 pixels to 22.7% at 768 pixels, whereas the equivalent MA proportion was already low at 6.2% and fell to 1.8%. Downsampling did not simply erase supervision for fine annotations; it compressed their footprint and, for the densely clustered exudate class, fused neighbouring components.

Testing the same account image by image gives it a chance to fail. Between 224 and 768 pixels, per-image average precision rose for 90.0% of hard-exudate, 89.1% of microaneurysm, 85.4% of haemorrhage and 74.5% of soft-exudate images, and the gain was concentrated: the most-improved fifth of images contributed 42% to 59% of the total positive gain depending on class. Geometry predicts the gain for two of the four classes. For haemorrhages and hard exudates, images whose components were

most compressed at 224 pixels gained the most (Spearman ρ 0.53 and 0.54 for Swin-Tiny, 0.47 and 0.57 for SwinV2-CR-Tiny), with correlations against median projected component area negative as the same explanation requires. For microaneurysms the correlation is near zero because the covariate saturates: 31.3% of MA images have every component below four target pixels at 224 pixels, against 5.7% for EX and 0.5% for HE. The aggregate microaneurysm effect is well supported, but this per-image explanation cannot be tested for it with this covariate.

Because the footprint statistics need no model, it is worth asking whether they predict which classes a resolution increase will help. They do not. On the relative scale the association looks strong: across all 120 combinations of cohort, system, lesion and step, $\Delta\varphi$ correlates with relative gain at $\rho = +0.57$, and within the two steps below 512 pixels the class with the largest $\Delta\varphi$ was the class with the largest relative gain in 19 of 20 rankings. Neither observation survives scrutiny. Relative gain is strongly negatively correlated with baseline performance ($\rho = -0.61$) and the four classes differ in baseline by an order of magnitude; on the absolute scale, which is uncorrelated with baseline ($\rho = +0.01$), the same ranking succeeds in 2 of 20. The inferential claim fails independently of the scale: under the exact permutation test that respects the dependence structure, the 19-of-20 result gives $p = 0.0625$, against $p = 5.6\times10^{-11}$ from a naive binomial test. What survives is a partial rank correlation of +0.45 between $\Delta\varphi$ and absolute gain after controlling for baseline, so annotation compression is not merely a proxy for baseline difficulty. That is an association, not a rule, and we make no predictive claim for these statistics. Supporting analyses are in Additional file 1, Figs. S2 and S3.

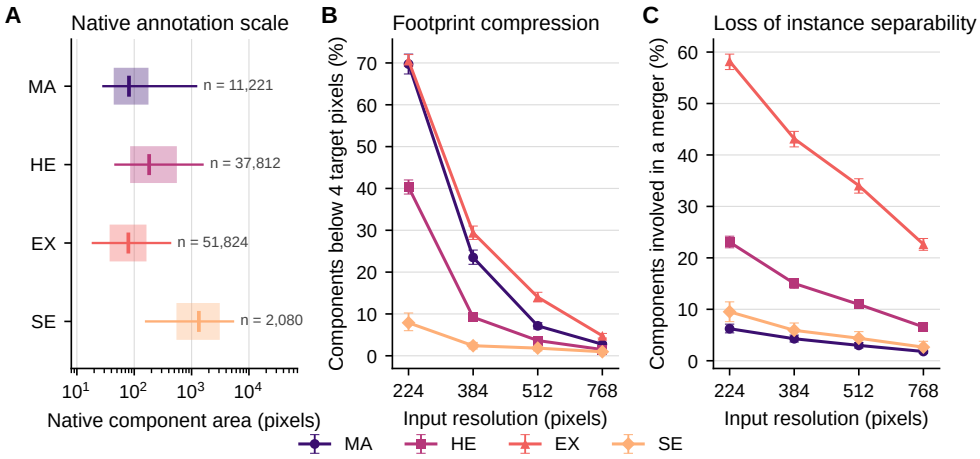


Fig. 4 Native annotation scale and what the study resize pathway does to it. a Distribution of native 8-connected component areas by lesion; the bar spans the interquartile range, the whisker the 10th to 90th percentiles, the tick marks the median, and n gives the number of components. **b** Percentage of components projecting to fewer than four target pixels at each input resolution, from Eq. (4). **c** Percentage of components fused with at least one neighbour after resizing, at an occupancy threshold of 0.10. Error bars in **b** and **c** are 95% image-cluster bootstrap intervals from 2,000 replicates. This analysis was post hoc, used no model predictions, and describes annotation geometry rather than lesion visibility.

3.5 Which systems are distinguishable, and which are equivalent

Ranking five systems by their mean is easy; separating them is not. Across the 40 pairwise comparisons in each cohort, the image-level bootstrap separated 30 pairs internally and 35 externally after Holm correction, whereas the seed-paired t-test separated 23 and 11 (Fig. 5). The two tests never disagreed in direction; they differ in power, and the gap widens externally. At 768 pixels on IDRiD the image-level test separated 8 of 10 pairs while the seed-level test separated none, because five seeds evaluated on 81 images leave large training-process variability. A difference of 0.158 macro AUPR between Swin-Tiny and CaiT-XS24 was unambiguous on the image-level test yet did not survive Holm correction on the seed-level test.

How much structure survives depends on which rule is applied, and we report both. Requiring both tests to agree, as our decision rule states, separates 5 of 10 pairs at 768 pixels on FGADR and none of 10 on IDRiD, so the conjunction yields no externally replicated ordering at all. Under the image-level test alone, 8 of 10 pairs separate at 768 pixels in each cohort, with the same two exceptions in both: Swin-Tiny against SwinV2-CR-Tiny, and the fixed-budget U-Net against ViT-Small/16. We do not claim a tier structure that survives both tests, because it does not.

Pairs that failed to separate were tested for equivalence against a margin of $\delta = 0.02$ macro pooled AUPR. Swin-Tiny and SwinV2-CR-Tiny are equivalent at 512 and at 768 pixels on FGADR under both procedures: the seed-level two one-sided tests give $p_{\mathrm{TOST}} = 0.008$ and 0.020, and the image-level 90% intervals, $[-0.0105, +0.0039]$ and $[-0.0178, +0.0026]$, lie inside the margin. The external picture is weaker and we do not generalise across it: the image-level interval supports equivalence at 512 pixels only, and the seed-level test reaches it at no external resolution, while at 224 and 384 pixels Swin-Tiny is ahead externally by 0.0355 and 0.0249 and the two systems are neither equivalent nor interchangeable. Internally at 224 pixels the difference of 0.0116 is statistically detectable yet its interval, $[+0.0061, +0.0175]$, lies inside the margin: a difference can be both real and too small to act on. Because that conclusion depends on the margin, the tests were repeated at $\delta = 0.01$ and $\delta = 0.03$. Swin-Tiny and SwinV2-CR-Tiny remain equivalent at 512 and 768 pixels internally at $\delta = 0.03$ as well as $\delta = 0.02$, but at $\delta = 0.01$ no resolution supports equivalence under both procedures. Across all 40 pairs per cohort the counts meeting the image-level criterion are 1 and 0 at $\delta = 0.01$, 8 and 1 at $\delta = 0.02$, and 10 and 5 at $\delta = 0.03$, for FGADR and IDRiD respectively. The interchangeability claim is therefore contingent on treating a 0.02 macro-AUPR difference as practically negligible; a reader who would act on 0.01 should not read these two systems as interchangeable. Full pairwise and equivalence results for all ten pairs, at all three margins, are in Additional file 1.

3.6 The backbone comparison reverses when the decoder is standardised

The relative standing of ViT-Small/16 and SwinV2-CR-Tiny changed when the native decoders were replaced by the shared-final design (Fig. 6). With native decoders, SwinV2-CR-Tiny held the higher macro AUPR at both endpoint resolutions and in

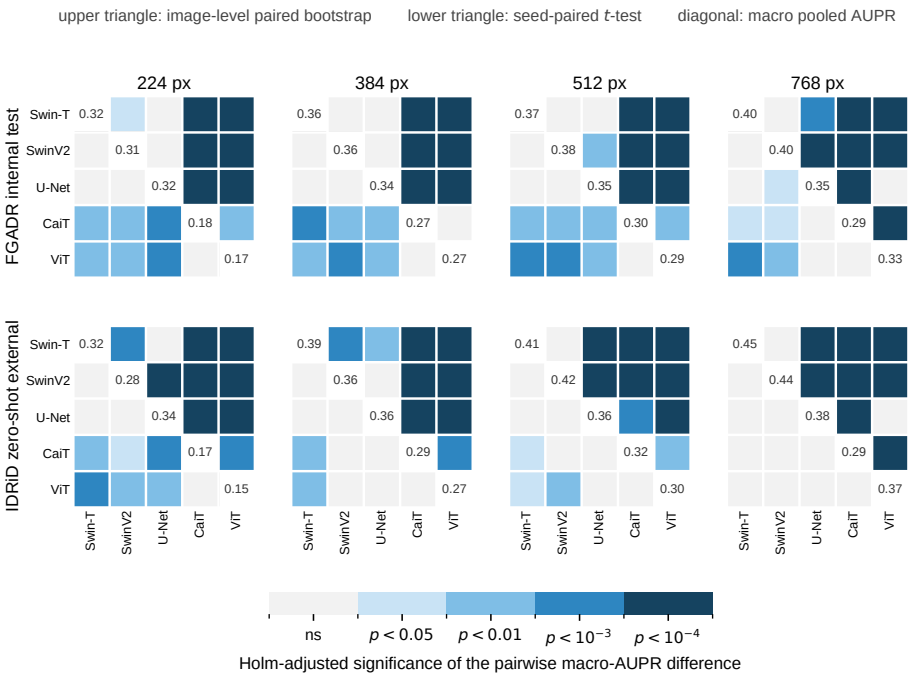


Fig. 5 Pairwise comparison of all five systems at every resolution. Cells show the Holm-adjusted significance of the difference in macro pooled AUPR between the row and column system, corrected within a 40-test family for each cohort and each test. The upper triangle of each panel is the paired image-cluster bootstrap over 2,000 replicates, which quantifies evaluation-cohort variability under fixed weights; the lower triangle is the seed-paired t-test over five matched seeds, which quantifies training-process variability. The diagonal gives the system's macro pooled AUPR. Top row, FGADR internal test; bottom row, IDRiD zero-shot external evaluation.

both datasets. With the shared-final decoder, ViT-Small/16 held the higher mean in all four comparisons: the FGADR shared-final means were 0.1432 against 0.1076 at 224 pixels and 0.3218 against 0.2940 at 768 pixels, and on IDRiD 0.1300 against 0.0686 and 0.3423 against 0.2597. The difference-in-differences estimand of Eq. (3) was positive in every condition, its nested 95% interval excluded zero in all four cases, and the four Holm-adjusted bootstrap sign p-values were 0.004 (Table 5). The observed backbone contrast therefore depended on decoder design. Because encoder capacity, feature stride, patch size and pretraining were not matched, this should not be read as a decoder-only causal effect.

3.7 Resolution changes calibration, and not in the same direction for every system

Pooled AUPR is invariant to monotone rescaling of the predicted scores, so none of the results above say whether the predicted probabilities are meaningful. Decomposing the Brier score separates the two properties, and they move together for the Transformers and apart for the convolutional reference (Fig. 7). For all four Transformer systems,

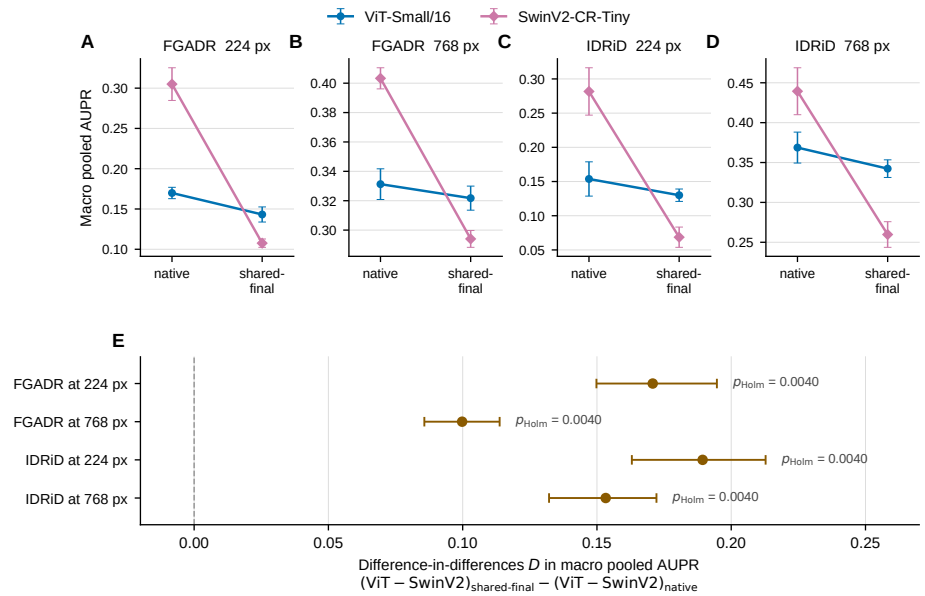


Fig. 6 The backbone comparison reverses when the decoder is standardised. a–d Macro pooled AUPR for ViT-Small/16 and SwinV2-CR-Tiny under each decoder condition, in both cohorts and at both endpoint resolutions. Points are five-seed means with t-based 95% confidence intervals. SwinV2-CR-Tiny is above ViT-Small/16 under native decoders and below it under the shared-final decoder in all four conditions. **e** Difference-in-differences estimand D from Eq. (3) with nested 95% bootstrap intervals from 2,000 replicates; positive values indicate a shift toward ViT-Small/16 under decoder standardisation.

Table 5 Post hoc macro-AUPR decoder interaction

Dataset	Resolution	D	Nested 95% CI	p_{Holm}
FGADR	224 px	0.1708	[0.1498, 0.1947]	0.004
FGADR	768 px	0.0998	[0.0857, 0.1138]	0.004
IDRiD	224 px	0.1893	[0.1630, 0.2128]	0.004
IDRiD	768 px	0.1533	[0.1322, 0.1722]	0.004

D is defined in Eq. (3). Positive values indicate a shift toward ViT-Small/16 under the shared-final decoder. Confidence intervals use 2,000 nested seed-and-image bootstrap replicates, and p-values are two-sided bootstrap sign values Holm-adjusted within the four-test primary family.

raising the input from 224 to 768 pixels reduced the macro-averaged calibration term and raised the discrimination term in both cohorts. Swin-Tiny improved from 1.02×10^{-3} to 0.87×10^{-3} on calibration and from 1.56×10^{-3} to 2.09×10^{-3} on discrimination internally, with expected calibration error falling from 4.96×10^{-3} to 3.95×10^{-3}. The external pattern is the same and stronger: expected calibration error fell from

4.92×10^{-3} to 3.21×10^{-3} for Swin-Tiny and from 6.44×10^{-3} to 4.02×10^{-3} for ViT-Small/16, and mean predicted probability converged on prevalence, so the systems became less overconfident as well as better separated.

Seed-level uncertainty qualifies how strongly that can be stated. With five seeds the 95% interval on the paired 224-to-768 change in macro expected calibration error excludes zero for all four Transformers on IDRiD (changes of -2.42, -2.46, -1.71 and -1.87×10^{-3} for ViT-Small/16, CaiT-XS24, Swin-Tiny and SwinV2-CR-Tiny) but for none of them on FGADR, where the internal changes are smaller and the intervals span zero. What supports the internal claim is consistency rather than any single interval: expected calibration error fell in all eight Transformer conditions defined by four systems and two cohorts, which an exact sign test places at $p = 0.008$. We therefore report the direction as consistent and externally supported, and do not claim a statistically resolved internal effect for any individual system.

The fixed-budget U-Net moved the other way, though less securely than its point estimates suggest. Its discrimination term barely changed (1.71 to 1.85×10^{-3} internally) while its calibration term worsened from 1.09 to 1.62×10^{-3} and its expected calibration error rose from 11.05×10^{-3} to 16.54×10^{-3}, more than 2.5 times the worst value recorded for any Transformer at any resolution, and the external cohort shows the same direction (9.63 to 14.79×10^{-3}). These are the widest intervals in the analysis: the seed-level 95% interval on the internal change is $[-10.3, +21.3] \times 10^{-3}$ and the external one $[-11.8, +22.1] \times 10^{-3}$, so neither excludes zero. The U-Net's calibration drift is a consistent point-estimate pattern with very large training-process variability, not a statistically resolved effect. Because 11 of 20 U-Net runs also met the budget-limit flag, it may reflect the fixed training budget rather than the architecture, and we do not present it as a property of convolutional decoders in general.

The macro average conceals heterogeneity that runs in the opposite direction for the class with the largest ranking gain, so the aggregate should not be read as a class-wise improvement. Microaneurysm reliability worsens between 224 and 768 pixels for all four Transformers, from 0.40 to 0.68×10^{-3} for Swin-Tiny and from 0.03 to 0.95×10^{-3} for ViT-Small/16. The near-zero values at 224 pixels are not evidence of good calibration: at that resolution the models predict close to zero almost everywhere for microaneurysms, and a model that never commits is trivially reliable. The improvement in the macro term therefore comes from the three commoner classes, while the class with the largest ranking gain moves in the opposite direction. Two boundaries apply: the terms are computed over all pixels without a field-of-view restriction, and they are pooled over pixels within an image, which understates uncertainty. Complete lesion-level values are in Additional file 1.

3.8 What the resolution gain costs

Because every run used one Tesla T4, training time is directly comparable across conditions, and the cost of moving from 224 to 768 pixels was strongly architecture-dependent (Fig. 8). Median training wall-clock per epoch rose 11.4-fold for ViT-Small/16 (19.0 to 216.2 s), 13.2-fold for Swin-Tiny (14.4 to 190.3 s) and 11.4-fold for the U-Net (17.2 to 196.0 s), but 31.8-fold for SwinV2-CR-Tiny (17.4 to 552.8 s) and 38.4-fold for CaiT-XS24 (30.2 to 1,158.8 s). Per-epoch cost is not total cost, because

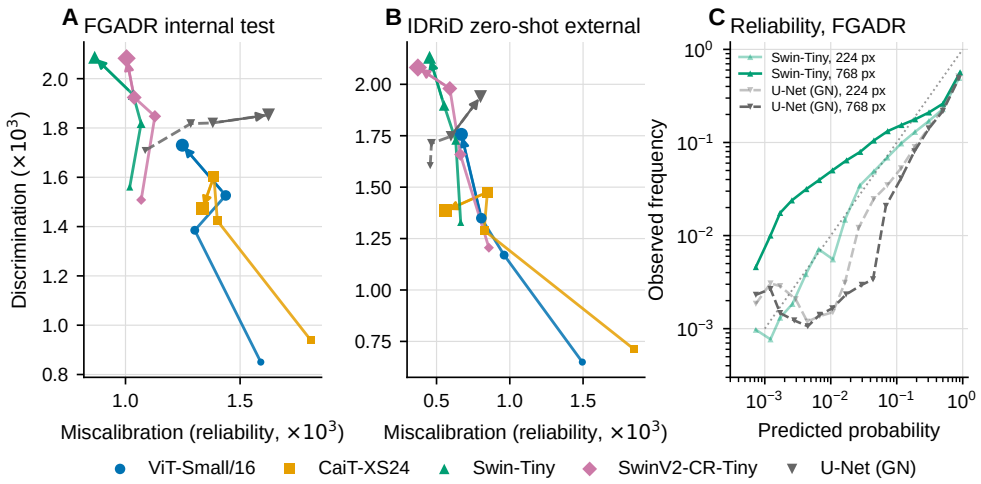


Fig. 7 Calibration and discrimination under changing input resolution. a, **b** Each trajectory follows one system across 224, 384, 512 and 768 pixels in the plane defined by the two informative terms of the Brier decomposition in Eq. (2); marker size increases with resolution and the arrow marks the final step. Better systems lie towards the upper left. All four Transformers move up and to the left; the fixed-budget U-Net moves right, becoming less well calibrated as resolution rises. **c** Reliability curves pooled over lesions and seeds on FGADR, on logarithmic axes, with the diagonal marking perfect calibration. Values are averaged over five seeds and, in **a** and **b**, over the four lesion classes.

early stopping ended runs at different epochs (Table 6). Measured to the selected checkpoint, the whole 120-run study consumed 163.1 GPU-hours. The ordering by total cost differs from the ordering by per-epoch cost, and the difference is informative: the fixed-budget U-Net ran 36 to 40 epochs at every resolution where the Transformer systems stopped at 16 to 22, so it spent more total compute than Swin-Tiny at 768 pixels despite a lower per-epoch cost and a third of the parameters. That is independent corroboration of the frozen budget-limit flag, arrived at without reference to the convergence rule. Inference latency, throughput and energy were not instrumented on the study hardware and are not reported.

Accuracy and cost did not move together. Swin-Tiny at 768 pixels cost 190 s per epoch for macro pooled AUPR of 0.3960 internally and 0.4510 externally, whereas CaiT-XS24 at its own best resolution of 512 pixels cost 331 s per epoch, nearly twice as much, for 0.3004 and 0.3181. The architecture with the steepest cost curve was also the one that lost accuracy beyond 512 pixels. Across the grid, the hierarchical systems held the accuracy–cost frontier above roughly 30 s per epoch in both cohorts.

Higher resolution also bought a noisier operating point. On FGADR, macro Dice for Swin-Tiny rose from 0.2794 at 224 pixels to 0.4080 at 768 pixels and object-level F1 from 0.2460 to 0.4696, with comparable SwinV2-CR-Tiny values; over the same range false positives per image rose from 2.45 to 7.55 for Swin-Tiny and from 2.78 to 7.33 for SwinV2-CR-Tiny, and externally from 1.43 to 3.34 for Swin-Tiny. The endpoint FROC analysis showed that the microaneurysm gains in pooled AUPR were accompanied by better object detection: on FGADR, ViT-Small/16 MA CPM increased from 0.004 to 0.083 and sensitivity at eight false positives per image from 0.010 to 0.280. This was

Table 6 Measured cost and capacity by system and resolution

System	Res. (px)	Params (M)	Epochs run	s per epoch	GPU-h total	Peak mem. (GB)
ViT-Small/16	224	23.04	16	19.0	0.63	1.54
ViT-Small/16	768	23.04	22	216.2	2.00	8.32
CaiT-XS24	224	27.71	16	30.2	0.33	2.17
CaiT-XS24	768	27.71	22	1,158.8	7.74	10.13
Swin-Tiny	224	28.88	16	14.4	0.24	1.01
Swin-Tiny	768	28.88	22	190.3	1.42	4.03
SwinV2-CR-Tiny	224	28.93	16	17.4	0.69	1.19
SwinV2-CR-Tiny	768	28.93	20	552.8	3.88	9.48
U-Net (GN)	224	7.85	40	17.2	0.48	1.55
U-Net (GN)	768	7.85	40	196.0	2.38	8.87

Medians over five seeds on one NVIDIA Tesla T4 under the frozen memory plan. Endpoint resolutions shown; all four are in Additional file 1. Epochs run reflects early stopping against a 40-epoch budget, so the U-Net used its whole budget at every resolution while no Transformer condition did.

not universal across lesions. SwinV2-CR-Tiny haemorrhage CPM fell from 0.123 to 0.086 on FGADR because at 768 pixels that model produced essentially no matched haemorrhage detections below four false positives per image, yet its sensitivity at eight false positives per image rose from 0.294 to 0.514: the decline reflects a shift in where the operating range sits, not a uniform loss of detection ability. Complete threshold-dependent values, U-Net results and FROC curves are in Additional file 1.

4 Discussion

4.1 Principal findings

Input resolution was not a neutral preprocessing choice in this study. It changed the performance of the complete segmentation system, and the direction and size of that change depended on both the lesion and the architecture. Between 224 and 768 pixels macro pooled AUPR increased for all four native Transformer systems on FGADR and in the zero-shot IDRiD evaluation, but three systems reached their highest macro means at 768 pixels whereas CaiT-XS24 peaked at 512 pixels in both cohorts. Separating the three adjacent steps sharpens the conclusion: most of the multiplicity-adjusted evidence sits at the low end of the range, and beyond 512 pixels only three lesion-by-family contrasts across both domains survived correction, all in the two smallest lesion classes. A practitioner reading only the endpoint difference would conclude that more pixels help throughout; the step-wise analysis says something narrower and more useful, that the first step of the range is where the reliable gain lives and the last step is worth paying for only when fine lesions are the target.

The annotation-footprint analysis offers a plausible geometric explanation. Downsampling did not simply remove supervision from fine annotations; it compressed their footprint and made neighbouring components more likely to merge, and the two small classes differ in which mechanism dominates. Microaneurysms are sparse and small, so

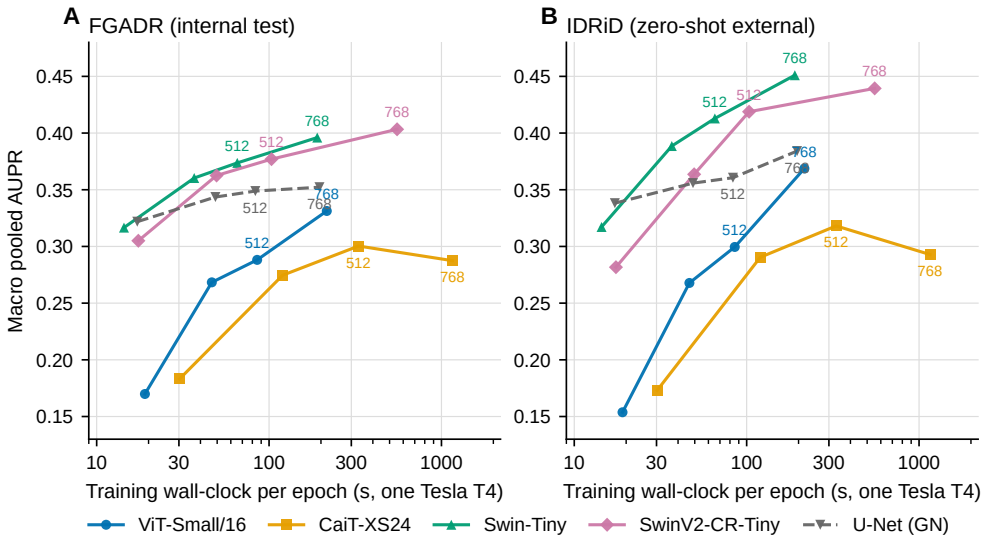


Fig. 8 Accuracy against measured training cost. a FGADR internal test. **b** IDRiD zero-shot external evaluation. Each line traces one system across 224, 384, 512 and 768 pixels; the two higher resolutions are labelled. The horizontal axis is the median wall-clock time of one training epoch across five seeds, measured on a single NVIDIA Tesla T4 under the frozen memory plan, and is shown on a logarithmic scale. Points are five-seed mean macro pooled AUPR. Times include the effect of the reduced micro-batches and gradient checkpointing used at the highest resolutions, so they reflect the cost actually incurred rather than a theoretical operation count.

compression is the binding constraint; hard exudates are small and densely clustered, so both compression and fusion apply. The scope of this account is narrow: connected annotation components are operational units rather than adjudicated clinical lesions, and the analysis supports geometric plausibility rather than a causal mechanism. The similar low-resolution profiles of MA and EX also carry a corrective, in that they do not support describing microaneurysms as uniquely the smallest lesion class in FGADR [4].

4.2 What an architecture ranking is worth

Comparative studies routinely order architectures by a mean and stop there. The pairwise analysis shows how little of that ordering survives testing, and that the answer depends on which source of uncertainty is being asked about. Evaluation-cohort variability separates most pairs; training-process variability, estimated from five seeds, separates far fewer, and externally it separates almost none. Both are real, but a practitioner who will train once and deploy once cares about the second, and a study claiming that architecture A beats architecture B in general is making a claim about the second that it has usually not tested. Requiring both tests to agree leaves five of ten pairs at 768 pixels internally and none externally.

The ordering within the leading pair is not defensible on either test, and we say so with an equivalence test rather than by pointing at a non-significant p-value. Internally, Swin-Tiny and SwinV2-CR-Tiny differ by less than 0.02 macro AUPR at 512 pixels

and above with 90% confidence, which for this task is a difference with no decision attached. Reporting that as an equivalence rather than as a narrow win changes what a reader should do: choose on cost, and Swin-Tiny costs about a third as much at 768 pixels. Two further observations follow. The equivalence appears only above 384 pixels, so architectural differences that matter on a coarse grid can vanish on a fine one, and a comparison run at a single low resolution may be measuring the resolution rather than the architecture. And with 81 external images and five seeds a gap of 0.158 macro AUPR was not separable after multiplicity correction, so external comparisons based on a single seed should be read as descriptive.

4.3 Calibration, decoder and cost are separate axes

The calibration result was not anticipated when the study was designed. Higher resolution did not merely reorder pixels more accurately for the Transformers; it also brought the predicted probabilities closer to observed frequencies, consistently across four architectures and two cohorts, while the convolutional reference gained little discrimination and lost calibration. This matters because a lesion segmentation model in a screening pathway is rarely used at a single fixed threshold: its scores may be aggregated into a burden estimate, ranked for review triage, or combined with an image-level grade, and none of those uses is measured by AUPR. Reporting a calibration term alongside a ranking metric costs nothing once per-image scores are retained. Two qualifications apply: microaneurysm reliability worsened for every Transformer, so the macro statement should not be read class by class, and because 11 of the 20 U-Net runs met the budget-limit flag its calibration drift is confounded with training budget.

The decoder analysis makes a second boundary difficult to ignore. What it establishes is an *encoder–decoder interaction*, not an isolated decoder effect: the two encoders differ in capacity, feature stride, patch structure, pretraining geometry and final feature representation, so replacing the decoder changes the pairing rather than one factor within it. The interaction is nonetheless solid – the reversal was consistent across two cohorts and both endpoint resolutions, with intervals excluding zero after multiplicity correction – and it carries the practical claim: comparing encoders only through architecture-specific segmentation heads risks attributing part of a system-level difference to the backbone, so a native ranking must not be treated as decoder-independent. Separating the decoder's own contribution would require the shared decoder across all four backbones, or matched feature projections and parameter budgets. Those conditions lie outside this design and are set out under future work.

The aggregation result sharpens the case for keeping the U-Net a labelled reference rather than a competitor. Under pooled AUPR it sits third at 768 pixels; under positive-image mAP it sits first. Both numbers come from the same 20 runs, and the gap between them is the false-positive burden it carries on lesion-negative images, which is what its calibration failure predicts. A study reporting only one aggregation would have drawn a confident and opposite conclusion about convolutional versus Transformer decoders, on a reference arm that was in any case budget-limited. The

lesson generalises past this comparison: where a screening cohort contains many lesion-negative images, the aggregation scheme is a modelling decision with the same standing as the architecture choice, and it should be pre-specified rather than selected after the fact.

Cost behaves as a third independent axis. The number of input pixels grows 11.8-fold between 224 and 768, and two of the five systems scaled roughly that way, but SwinV2-CR-Tiny and CaiT-XS24 cost 32 and 38 times more per epoch, partly reflecting the reduced micro-batches and gradient checkpointing the memory plan imposed at the highest resolution. The systems worth running at high resolution are therefore not the ones a pixel-count argument would pick: Swin-Tiny at 768 pixels was both the cheapest high-resolution option and among the most accurate in both cohorts, while CaiT-XS24 was the most expensive and the only system whose accuracy fell over the last step.

4.4 Practical implications

Combining these strands gives a concrete recommendation for this task. Where fine-lesion sensitivity is the objective, use a hierarchical encoder at 768 pixels and accept roughly a thirteen-fold increase in training cost per epoch over 224 pixels. Where a single operating point must serve a mixed lesion profile under a constrained budget, 512 pixels captures most of the achievable accuracy at about a third of that cost, and internally the two hierarchical systems are interchangeable there. Below 384 pixels the choice of architecture starts to matter more than the choice of resolution, which is the opposite of the usual assumption. Whichever setting is chosen, the false-positive burden and the calibration term belong next to the ranking metric, because they do not move together: false positives per image tripled over the same range in which Dice rose by roughly half and object-level F1 nearly doubled.

The footprint analysis has a use beyond explaining these results, because its inputs are the annotations and the resize pathway rather than a trained model. That is the reasoning behind releasing it as *maskscale*: a practitioner facing a new segmentation dataset can measure, in minutes and on a laptop, how much of each class is compressed below a few target pixels at each candidate resolution, how much fuses with a neighbour and how much disappears. What it will not do is predict which classes benefit: the apparent skill is an artefact of scaling, because relative gain is largely determined by which class starts lowest, and on an unconfounded outcome scale the ranking performs at or below chance. We report this because the temptation to promote such a statistic into a design rule is real, and because the same reasoning applies to any pre-training heuristic validated on relative improvements over weak baselines. The residual partial association supports using the tool as a descriptive screen only: a class with little compression at the lowest candidate resolution has little geometric headroom, whatever else may be true of it.

Absolute values reported here should not be compared with published FGADR or IDRiD scores, because the evaluation is pooled across every cohort image rather than averaged over lesion-positive images, one network predicts all four classes jointly rather than one specialised model per lesion, and the partition differs from those used elsewhere.

4.5 Limitations

The optimisation-level sample contained five seeds, so seed-level significance tests have limited resolution despite the complementary image and nested bootstrap analyses. FGADR was image-disjoint under exact-duplicate and identical-dHash controls, but patient identifiers were unavailable and patient-level separation could not be established. The evidence base comprised one development dataset and one zero-shot external dataset. These differ in source population and acquisition setting, so the transfer reported here crosses a real domain shift; it remains a single external cohort of 81 images carrying the same four target categories, so broader transfer across devices, populations and prospective screening workflows is still untested. Resolution was evaluated as a property of each complete pretrained system, so the study does not separate spatial resolution from positional-embedding interpolation, window configuration or transfer away from native pretraining geometry.

The shared-final, footprint and calibration analyses were post hoc, even though all used frozen evidence and multiplicity-controlled or bootstrap inference. FROC was restricted to two backbones and the endpoint resolutions, and the fixed-budget U-Net cannot support an absolute Transformer versus U-Net conclusion because 11 of its 20 runs met the frozen convergence flag; its standing also changes with the aggregation scheme, which is a further reason to read it as a reference rather than a competitor. The equivalence margin of 0.02 macro AUPR was chosen before testing on substantive grounds, but it remains a judgement, and the interchangeability conclusion holds at 0.02 and 0.03 and not at 0.01. The calibration terms are binned approximations computed without a field-of-view mask and treating pixels as exchangeable within an image, so retinal background contributes to every term; the inflation is shared across systems and so cannot by itself explain a between-system difference, but the absolute magnitudes should be read with that in mind. Seed-level intervals on the change in expected calibration error exclude zero externally but not internally, so the internal direction rests on consistency across eight conditions rather than on any single interval. The U-Net arm was trained under the same fixed budget as the Transformer systems rather than to convergence, which bounds what its comparison can establish. The cost analysis reports measured wall-clock on one device class under one memory plan; inference latency, throughput and energy were not instrumented and are not reported.

5 Conclusions

Treating input resolution as a routine image-resize setting hides a design decision that interacts with everything around it. Across 120 trained models, higher resolution improved fine-lesion segmentation, probability calibration and zero-shot transfer, but the benefit was concentrated in the first step of the range, restricted to the smallest lesion classes at the top of it, and absent under pooled AUPR for one of the four architectures beyond 512 pixels, a decline that does not survive the change to positive-image aggregation internally. Pairwise testing showed how little of a five-way ranking survives: requiring agreement between evaluation-cohort and training-process resampling separates half the pairs internally and none externally, while equivalence testing established that the two leading systems are interchangeable at 512 pixels and above

internally at margins of 0.02 and 0.03, though not at 0.01, so for a practitioner working to the coarser tolerance the choice between them is a cost decision rather than an accuracy one. Standardising the decoder reversed the ranking of two backbones in every condition tested. Because the two encoders are unmatched in several respects this is an encoder–decoder interaction rather than an isolated decoder effect, which means an encoder comparison read through architecture-specific heads may not be an encoder comparison at all.

Resolution, encoder and decoder therefore form a single choice rather than three independent ones, and that choice should be justified against the lesions that matter and the budget available. Because the geometric part of this account depends only on the annotations and the resize pathway, it can be measured before any model is trained; we tested whether that measurement could also predict which classes benefit and found that it cannot, and we release it as *maskscale*, a descriptive diagnostic with no predictive claim attached. Work that follows should test whether these patterns hold on additional cohorts and imaging devices, whether matched-capacity encoder pairs reproduce the decoder interaction, whether calibration continues to improve with resolution once the training budget is equalised across architectures, and whether resolution can be varied per lesion within one system rather than fixed once for the whole task. Retinal lesion segmentation studies should in the meantime report resolution, encoder, decoder, calibration and compute together, since any one of them can change what the others appear to show.

List of abbreviations

AUPR: area under the precision–recall curve; CLAHE: contrast-limited adaptive histogram equalisation; CPM: competition performance metric; DR: diabetic retinopathy; ECE: expected calibration error; EX: hard exudates; FGADR: Fine-Grained Annotated Diabetic Retinopathy Seg-set; FROC: free-response receiver-operating characteristic; HE: haemorrhages; IDRiD: Indian Diabetic Retinopathy Image Dataset; IoU: intersection over union; MA: microaneurysms; SE: soft exudates; TOST: two one-sided tests; ViT: Vision Transformer.

Supplementary information. Additional file 1 (Microsoft Word document) contains the supplementary methods, tables and figures referenced in the text: full layer-level architecture specifications, training hyperparameters and the computational-environment and reproducibility record; lesion-wise pooled AUPR with standard deviations for all 120 runs in both cohorts (Table S1); threshold-dependent and object-level measures (Table S2); complete pairwise and equivalence results for all ten system pairs (Table S3); the full calibration decomposition by lesion (Table S4); and Figs. S1–S4 (adjacent-resolution step effects, per-image gain, the annotation-compression criterion, and endpoint FROC curves).

Declarations

Ethics approval and consent to participate. This study used two publicly available datasets, FGADR [4] and IDRiD [5]. As the data were anonymised and obtained

from public repositories, specific ethical approval and informed consent were not required for this secondary analysis. Under the research policy of Universiti Teknologi PETRONAS, secondary analyses of publicly available, fully anonymised datasets do not require institutional ethics review, and no approval or waiver reference is issued for such work. The research was conducted in accordance with the principles of the Declaration of Helsinki and with the research use agreements of both source datasets.

Consent for publication. Not applicable.

Availability of data and materials. Both source datasets are third-party resources and are not redistributed with this article.

The FGADR Seg-set analysed during the current study is the property of the Inception Institute of Artificial Intelligence (IIAI), Abu Dhabi, United Arab Emirates, and is available for non-commercial research use only under the IIAI FGADR Dataset Research Use Agreement. Access is granted on application: the agreement must be signed and returned to the dataset maintainer (yizhou.szcn@gmail.com), and the dataset is described at https://csyizhou.github.io/FGADR/ and in the accompanying publication [4]. Because that agreement prohibits reproducing, publishing or distributing any portion of the images or of derived data, the images are not available from the authors and cannot be deposited in a public repository.

The IDRiD dataset analysed during the current study is openly available from IEEE DataPort under the persistent identifier https://doi.org/10.21227/H25W98, with further description at https://idrid.grand-challenge.org/ and in the corresponding publications [5, 6].

The derived evidence supporting this article is released in two tiers. The public tier contains lesion-wise metrics for all 120 runs in both cohorts, per-run parameters, timings and memory plan, FROC sensitivities, the seed-paired adjacent-resolution contrasts, the pairwise and equivalence results, the calibration decomposition, the annotation-footprint and retention summaries, the partial-correlation table, and the analysis and figure scripts. It regenerates ten of the twelve figures and contains the values behind every table. The public tier is available at https://doi.org/10.5281/zenodo.22290796. A second tier holds the evidence keyed to individual FGADR images: per-image average precision and object counts, per-image annotation geometry, the per-image score histograms and the image-resampling bootstrap replicates. Figures 7 and S2, and re-deriving the pairwise, equivalence and calibration tables from their inputs, additionally require that tier. Because the FGADR Research Use Agreement restricts distribution of data derived from the dataset, the image-keyed tier is not deposited publicly; it is available from the corresponding author on reasonable request to requesters holding a signed FGADR Research Use Agreement. Neither tier contains or permits reconstruction of any source image. The *maskscale* diagnostic is released as open-source software under the MIT licence and is available at https://doi.org/10.5281/zenodo.22290752.

Competing interests. The authors declare no competing interests.

Funding. This work was supported and funded by the Deanship of Scientific Research at Imam Mohammad Ibn Saud Islamic University (IMSIU) (Grant Number IMSIU-DDRSP2602). This work was also supported by the Yayasan Universiti

Teknologi PETRONAS Fundamental Research Grant (YUTP-PRG; grant 015PBC-066).

Authors' contributions. AOA conceived the study, designed the experimental matrix, implemented the training and evaluation pipeline, performed the statistical analyses, and wrote the first draft. HA supervised the project, acquired resources and revised the manuscript. MN supervised the project, contributed to the methodology, validated the analyses and revised the manuscript. WEA acquired funding and revised the manuscript. All authors read and approved the final manuscript.

Acknowledgements. The authors thank the teams behind the FGADR Seg-set and the Indian Diabetic Retinopathy Image Dataset for making expert lesion annotations available for research use.

References

[1] Teo ZL, Tham YC, Yu M, Chee ML, Rim TH, Cheung N, et al. Global Prevalence of Diabetic Retinopathy and Projection of Burden through 2045: Systematic Review and Meta-analysis. Ophthalmology. 2021;128(11):1580–1591. https://doi.org/10.1016/j.ophtha.2021.04.027.

[2] Orlando JI, Prokofyeva E, del Fresno M, Blaschko MB. An Ensemble Deep Learning Based Approach for Red Lesion Detection in Fundus Images. Computer Methods and Programs in Biomedicine. 2018;153:115–127. https://doi.org/10.1016/j.cmpb.2017.10.017.

[3] Tan JH, Fujita H, Sivaprasad S, Bhandary SV, Rao AK, Chua KC, et al. Automated Segmentation of Exudates, Haemorrhages, Microaneurysms Using Single Convolutional Neural Network. Information Sciences. 2017;420:66–76. https://doi.org/10.1016/j.ins.2017.08.050.

[4] Zhou Y, Wang B, Huang L, Cui S, Shao L. A Benchmark for Studying Diabetic Retinopathy: Segmentation, Grading, and Transferability. IEEE Transactions on Medical Imaging. 2021;40(3):818–828. https://doi.org/10.1109/TMI.2020.3037771.

[5] Porwal P, Pachade S, Kamble R, Kokare M, Deshmukh G, Sahasrabuddhe V, et al. Indian Diabetic Retinopathy Image Dataset (IDRiD): A Database for Diabetic Retinopathy Screening Research. Data. 2018;3(3):25. https://doi.org/10.3390/data3030025.

[6] Porwal P, Pachade S, Kokare M, et al. IDRiD: Diabetic Retinopathy - Segmentation and Grading Challenge. Medical Image Analysis. 2020;59:101561. https://doi.org/10.1016/j.media.2019.101561.

[7] Dutta I, Jeong T. Multi-Resolution Feature Stem for Diabetic Retinopathy Lesion Segmentation. arXiv preprint arXiv:260708679. 2026;https://doi.org/10.48550/arXiv.2607.08679.

[8] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. In: International Conference on Learning Representations; 2021. Available from: https://openreview.net/forum?id=YicbFdNTTy.

[9] Touvron H, Cord M, Sablayrolles A, Synnaeve G, Jégou H. Going Deeper With Image Transformers. In: 2021 IEEE/CVF International Conference on Computer Vision; 2021. p. 32–42.

[10] Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In: 2021 IEEE/CVF International Conference on Computer Vision; 2021. p. 10012–10022.

[11] Liu Z, Hu H, Lin Y, Yao Z, Xie Z, Wei Y, et al. Swin Transformer V2: Scaling Up Capacity and Resolution. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2022. p. 12009–12019.

[12] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015. vol. 9351 of Lecture Notes in Computer Science. Cham: Springer; 2015. p. 234–241.

[13] Ullah Z, Usman M, Latif S, Khan A, Gwak J. SSMD-UNet: Semi-supervised Multi-task Decoders Network for Diabetic Retinopathy Segmentation. Scientific Reports. 2023;13:9087. https://doi.org/10.1038/s41598-023-36311-0.

[14] Zuiderveld KJ. Contrast Limited Adaptive Histogram Equalization. In: Heckbert PS, editor. Graphics Gems IV. San Diego: Academic Press; 1994. p. 474–485.

[15] Wu Y, He K. Group Normalization. In: Computer Vision - ECCV 2018. vol. 11217 of Lecture Notes in Computer Science. Cham: Springer; 2018. p. 3–19.

[16] Wightman R.: PyTorch Image Models. GitHub. Accessed 18 August 2026. Software, GitHub. Available from: https://github.com/huggingface/pytorch-image-models.

[17] Loshchilov I, Hutter F. Decoupled Weight Decay Regularization. In: International Conference on Learning Representations; 2019. Available from: https://openreview.net/forum?id=Bkg6RiCqY7.

[18] Loshchilov I, Hutter F. SGDR: Stochastic Gradient Descent With Warm Restarts. In: International Conference on Learning Representations; 2017. Available from: https://openreview.net/forum?id=Skq89Scxx.

[19] Abraham N, Khan NM. A Novel Focal Tversky Loss Function With Improved Attention U-Net for Lesion Segmentation. In: 2019 IEEE 16th International Symposium on Biomedical Imaging; 2019. p. 683–687.

[20] Kuhn HW. The Hungarian Method for the Assignment Problem. Naval Research Logistics Quarterly. 1955;2(1-2):83–97. https://doi.org/10.1002/nav.3800020109.

[21] Holm S. A Simple Sequentially Rejective Multiple Test Procedure. Scandinavian Journal of Statistics. 1979;6(2):65–70.

[22] Efron B, Tibshirani RJ. An Introduction to the Bootstrap. New York: Chapman and Hall/CRC; 1993.

Supplementary Files

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

- Additionalfile1.docx