Supplementary material

Lightweight attention mechanisms in breast ultrasound lesion segmentation: parameter cost, protocol choice, data leakage, and what the data can resolve

S1. Minimum detectable effect per comparison

The MDE is a function of the standard deviation of each comparison's own paired differences. That standard deviation varies by a factor of 2.8 across our nine comparisons, from 0.0731 for a mechanism added to its own reference to 0.2055 for two different architectures, giving thresholds from 0.0148 to 0.0415 at $\alpha = 0.05$. There is therefore no single detection threshold applicable to all comparisons, and the values in Table **??** should not be used to construct one.

All values are two-tailed at 80 % power, computed as MDE $= (z_{1-\alpha/2} + z_{0.80}) \cdot \mathrm{SD}/\sqrt{192}$, where 192 is the number of independent clusters the 234 test observations fall into. The corrected column uses the most stringent level of each comparison's own family under Holm–Bonferroni: $\alpha = 0.05/5 = 0.01$ for Family A and $\alpha = 0.05/4 = 0.0125$ for Family B. The two families are corrected separately and their corrected columns are not on a common scale.

Table S1: Per-comparison detection threshold, computed on 192 independent clusters.

Family	Comparison	SD(diff)	SE	MDE ($\alpha = 0.05$)	MDE (corrected)	Observed effect
B	`R50-UNet` → `R50-UNet + Aug`	0.2055	0.01483	0.0415	0.0495	+0.0634
B	`U-Net-3L` → `U-Net-4L`	0.1502	0.01084	0.0304	0.0362	+0.0400
B	`U-Net-4L` → `U-Net-4L + Aug×2`	0.1450	0.01046	0.0293	0.0349	+0.0195
B	`U-Net-4L + Aug×2` → `...+ Aug×3`	0.0968	0.00699	0.0196	0.0233	−0.0058
A	`+ ECA + DS`	0.1063	0.00767	0.0215	0.0262	+0.0137
A	`+ CBAM + DS`	0.1358	0.00980	0.0275	0.0335	+0.0082
A	`+ AG`	0.0731	0.00527	0.0148	0.0180	+0.0053
A	`+ scSE + DS`	0.0979	0.00707	0.0198	0.0242	+0.0030
A	`+ SE + DS`	0.0963	0.00695	0.0195	0.0238	+0.0020

Clustering correction

The 234 test observations are not 234 independent units, for two reasons that compound. The three splits are drawn independently rather than as folds of a partitioned scheme, so their test partitions are not disjoint: 27 images contribute more than one observation, and the 234 observations come from 205 unique images. The duplicate audit (Methods) shows that those 205 images are themselves not independent: grouping near-duplicate records collapses them to 192 clusters.

We take the cluster as the unit of information and compute every standard error on 192 units, equivalent to multiplying the nominal standard error by $\sqrt{234/192} = 1.104$. Every threshold and every confidence interval in this paper and in this supplement is computed on that basis; quantities assuming 234 independent observations are not reported. One consequence is visible in *Results*: on the widest of the four ablation-axis comparisons the interval crosses zero while the primary rank test still clears its corrected threshold. The interval is parametric and the verdict is not, and we report both rather than choosing between them. The correction is applied to the standard error and not to the effect estimates, which the pairing leaves unaffected. Treating repeated images as carrying no additional information is the conservative choice: the repeated observations come from separately trained models, so they are correlated rather than identical.

A single 80/20 split on the same data yields roughly 130 test images, about 122 clusters at the same duplicate rate, a standard error of 0.00962 and an MDE of 0.0270 — 1.25 times the threshold obtained from three independent random splits. The clustering multiplier for a single split is 1.032, not 1.104: the 234 → 205 component is undefined when there is only one partition.

Sensitivity to test and correction choice

The primary comparison was recomputed under both tests and both correction states for all nine comparisons. No p-value is selected from that recomputation; it is reported so that the reader can see the verdict does not depend on the choice. For `U-Net-4L` → `U-Net-4L + Aug×2` the two tests diverge by roughly a factor of ten (Wilcoxon Holm-4 0.0081, paired t Holm-4 0.0811), which is discussed in *Results*. The full table is in the reproduction package.

S2. Boundary metrics, full paired results

Boundary metrics were measured and are reported whether or not they produced a finding. Table **??** gives paired differences against **U-Net-4L + Aug×2** in pixels; negative favours the variant. Pairs in which either model predicted no foreground are excluded, so n differs by row.

Table S2: Paired boundary differences, Family A.

Variant	HD95	ASSD	n	Holm-5 p (HD95 / ASSD)
+ ECA + DS	−3.3189	−1.4947	232	0.1888 / 0.0503
+ CBAM + DS	−2.2560	−0.7407	230	0.4032 / 0.9375
+ AG	−0.9499	−0.6583	232	0.9370 / 0.9375
+ scSE + DS	−0.0888	+0.0864	232	0.8331 / 1.0000
+ SE + DS	+2.5550	+1.0435	231	0.9370 / 1.0000

The ECA + DS variant is again the only one with a visible signal and again at the threshold. The SE + DS variant is worse than the reference on both boundary metrics, and the scSE + DS variant is unchanged — its two differences have opposite signs and both are below 0.1 px. As with Dice, the direction is more often favourable than not and no comparison is resolvable; we report this without treating the direction consistency as a result.

Multiplicity across metrics. Boundary metrics are corrected within metric — Holm over the same five comparisons — and not across metrics. Dice, HD95 and ASSD are not independent hypotheses but three summaries of the same pair of masks, and treating them as separate tests would penalise one question asked three ways. The choice does not affect any conclusion: no boundary comparison clears its own corrected threshold, and correcting across all fifteen tests moves the smallest of them from 0.050 to 0.151 — further from the threshold rather than across it.

S3. Dice and IoU as computed here

Both are computed per image on binarized masks, the prediction thresholded at 0.5, with a smoothing constant s = 1e-6 in numerator and denominator:

$$\text{Dice} = \frac{2|P \cap G| + s}{|P| + |G| + s}$$

with IoU obtained from the same intersection and union counts. The constant keeps the value defined when both masks are empty; no image in the evaluated set has an empty reference, so it affects no reported number. The implementation is the one released with the code, and every metric in this paper is computed by it from stored prediction masks rather than during training.

S3b. IoU as a transform of Dice

At the per-image level IoU is an exact deterministic monotone transform of Dice, IoU = Dice / (2 − Dice). Across our measurements the maximum deviation from this identity is 1.3e-6, within-model Spearman correlation is 1.000000, and the sign agreement of paired differences is 100 %.

Wilcoxon p-values computed on the two metrics can nonetheless differ, because the transform rescales the magnitudes of the paired differences and therefore changes the ranks. For the ECA + DS variant the Holm-corrected value is 0.0909 on Dice and 0.0498 on IoU, and the corrected value for ASSD is 0.0503 — the three straddle the conventional threshold within 0.0005 of one another. Reporting one as significant while the others are not would be a selection among metrics rather than a result. Dice and IoU are consequently read as a single result throughout this paper, and a comparison is never reported as significant on one and not the other.

S4. Baseline parity audit

The retrained baselines use model definitions, optimizer, learning rate, loss and scheduler taken unchanged from the reference repositories; only the data loader and training-control logic were replaced. A line-by-line audit against those repositories identified and corrected three initially

mismatched items: the loss weighting, the learning-rate schedule, and weight decay for UNeXt. Table **??** gives the published hyperparameters.

Table S3: Hyperparameters of the two retrained baselines, as published.

	CMUNeXt	UNeXt
Optimizer	SGD, momentum 0.9	Adam
Initial learning rate	0.01	1e-3
Weight decay	1e-4	1e-4
Batch size	8	16
Schedule	polynomial, power 0.9	cosine, floor 1e-5

Both use the loss function from their own reference implementation ($0.5 \cdot \text{BCE} + \text{Dice}$), imported rather than reimplemented. The scheduler horizon was set to the epoch cap of 300 and early stopping used a patience of 30 — a different value from the ablation runs, which is a property of the two sets of runs and not of the comparison between them.

Normalization. Both reference implementations apply a normalization transform and additionally divide by 255 in the dataset class. We treat this published pre-processing path as the faithful arm and use it for the primary results; a single-normalization (`/255` only) arm is reported as a sensitivity analysis. The ordering between the two baselines is unchanged, but against the highest-scoring configuration of the ablation the sign does change. Neither interval excludes zero. Table **??** gives both arms.

Table S4: The two pre-processing arms. Paired difference is against the highest-scoring ablation configuration; positive favours the baseline.

Arm	UNeXt	CMUNeXt	Paired difference	95 % CI
Faithful (primary)	0.7319	0.7720	+0.0150	[−0.0121, +0.0421]
Single normalization (`/255`)	0.7466	0.7539	−0.0030	[−0.0297, +0.0236]

The pre-registered parity check on the first split (threshold 0.005) did not hold, and per the pre-registered rule the faithful arm was made primary and the full six-run set repeated. This decision is one of fidelity to the published pre-processing, not an empirical selection between the two arms.

A third baseline was removed before training. DAU-Net was listed in the study plan and dropped for three reasons recorded in advance: it is published at 128×128, so retraining at 256×256 would move its published operating point while retraining at 128×128 would make it incomparable to the other models; its reference implementation does not build under the framework version used here; and at approximately 15 M parameters it falls outside the lightweight scope. No claim of superiority over DAU-Net is made in this paper.

S4b. Training setup of the eleven ablation configurations

All eleven share the settings in Table **??**, taken from the configuration notebooks released with the code.

Table S5: Settings shared by all eleven ablation configurations.

Optimizer	Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 10^{-7}$)
Loss	binary cross-entropy + Dice term
Batch size	1
Schedule	learning rate halved after 8 epochs without improvement in the monitored quantity, floor 1e-5
Stopping	patience 20 with restoration of the best weights, epoch cap 200
Deep supervision	loss weights 1.0 on the final output and 0.4, 0.2, 0.1 on the three auxiliary heads, all removed at inference

Two settings are not shared, and we state them rather than describe the set as uniform. Table **??** gives them.

Table S6: The two settings that differ across configurations.

	Initial learning rate	Checkpoint monitored
`R50-UNet + Aug`	1e-4	validation loss
the other ten	1e-3	see below

The configuration with the pretrained ResNet-50 encoder was trained at a tenth of the learning rate used by the others. It is one of the two configurations carrying that encoder and it produces the largest resolved effect on the ablation axis, so the difference is stated here rather than left to the code.

The monitored quantity is not uniform either. The four deeply supervised configurations were checkpointed on binarized validation Dice and three others on validation loss; for the remaining four the notebooks record the monitor but we did not re-derive it per run, and we report the two we verified rather than implying a rule that covers all eleven. The asymmetry this creates is stated in *Methods*.

S5. Literature protocol audit, study by study

The record-level table — 22 studies × the protocol dimensions reported in *Results* Table 7, with each study's stratum membership, the search route that found it, its citation count and access date, and the exclusion reasons for records examined and not included — is released with the reproduction package.

Eligibility. A study is included if it reports a numerical segmentation metric on BUSI, is in English, is peer-reviewed or a preprint with a published version, and is available in full text; abstract-only availability is not sufficient. A study is excluded if it reports classification only, uses BUSI only in a qualitative figure, has been retracted and superseded (the superseding record is used), or is a duplicate version of a study already included (the journal version is used). Semi-supervised studies are included but flagged, because the labelled-data fraction is a separate degree of freedom.

Search. Studies were first identified by citation searching — backward and forward chaining from the comparison tables of recent lightweight BUSI segmentation papers — and, because citation searching under-samples work not cited in such tables, complemented with database

searching across comprehensive indexes, publisher platforms and preprint servers. All records were screened against the same criteria and each row records which route found it. Searching stopped when five consecutive new studies added no protocol pattern not already present in the table.

Extraction. The recording rule is stated in *Methods* and applied without exception here; where a repository exists, the code takes precedence over the paper text, and no protocol property is inferred from an absence.

S5b. The margin sample, and why it is separate from the protocol audit

The 22-study record described in S5 is a protocol audit: what each study states about how it was evaluated. The margin figures quoted in *Results* come from a different and larger retrieval with a different purpose, and the two samples should not be read as one. Table **??** sets them side by side.

Table S7: The two literature samples and what each supports.

	Protocol audit (S5)	Margin sample
Question	what a study states about how it was evaluated	what improvement margin a study claims
Retrieval	22 studies in two pre-defined strata	open-access breast ultrasound segmentation articles reachable in full text through Europe PMC
Unit	study	reported margin
Yield	22 studies	19 studies, 49 margins
Supports	existence claims about protocol heterogeneity	counts over the retrieved slice
Does not support	prevalence in the literature	prevalence in the literature

Retrieval. Open-access breast ultrasound segmentation articles reachable in full text through Europe PMC. Full text was parsed rather than abstracts.

Extraction. Figure and table blocks were removed from the body before parsing, so every margin counted here was stated in running text. Numerical values were aligned to the metric they belong to by position — in a sentence reporting *"0.32 %, 0.45 % and 0.28 % in DSC, VOE and recall"*, only the first value is recorded as Dice — and phrases describing a decrease were excluded. An earlier version of this extraction assigned every percentage in a Dice-containing sentence to Dice; a manual read of 20 values showed 11 of them belonged to other metrics, and the rule above replaced it. The replacement was audited against 12 hand-read values and agreed on 11.

Yield. 19 studies stated at least one Dice improvement margin in running text, giving 49 margins.

What this sample is not. Most studies proposing an architecture state their improvement margin only inside a comparison table, and those margins are invisible to a running-text extraction. The 19 studies are therefore a minority slice of the retrieved literature, selected by a reporting habit rather than by content. The proportions in *Results* are counts over that slice. They are not estimates of how the literature as a whole reports improvements, and we do not use them as such.

Sensitivity to the threshold. The proportions are read against a single comparison's corrected threshold (0.0262). The nine thresholds in this study span 0.0148 to 0.0415 (Table **??**), so a margin near the middle of the observed range can fall on either side depending on which comparison it is scaled against. This is the same point the paper makes about its own comparisons: there is no single ruler.

S6. Reproduction package and provenance

Contents. Code and the per-image measurement tables required to reproduce every number in this paper, together with the mapping from the labels used here to the internal identifiers used in the released files, are deposited at https://doi.org/10.5281/zenodo.21789807; the trained model weights are deposited at https://doi.org/10.5281/zenodo.21870651. Per-split tables, empty prediction counts, test-time augmentation results, the foundation-model reference by split and the directory-forensics output are in the reproduction package rather than in this supplement. The per-pass file counts behind the stale-accumulation finding are included there: in the first two splits two augmentation indices cover 613 and 615 source images while the third covers 500, and the 500 are a subset of the 613 and 615 — the signature of a second write pass in the first two splits and a single pass in the third.

What re-execution establishes, and what it does not. An earlier version of the package contained the evaluation paths for the retrained baselines and the foundation-model reference, but not the path that produced the per-image metrics for the ablation configurations; that step lived in the training notebooks. On a first attempt to regenerate prediction masks from the stored weights we recovered the recorded per-image Dice to within 0.0002 on some images and 0.008 on others, depending on which of two standard resampling implementations was used — a disagreement of the same order as the effects this study reports.

The path has since been extracted and added. It fixes the four choices in Table **??**, none of which the stored weights determine.

Table S8: The four pre-processing choices that stored weights leave open.

Choice	Fixed value
Interpolation	bilinear for images, nearest-neighbour for masks
Reference threshold	127.5, not 127
Division by 255	after resizing, not before
Output head, deeply supervised models	the first of the four supervision outputs

Regenerating every per-image value with it reproduces the recorded table exactly, at the precision at which that table is stored, across eleven configurations, three splits and 78 images per split, in the environment in which those values were produced. One exception is measured rather than assumed: on a single image the regenerated value differs by 3×10^{-5} — a five-hundredth of the smallest threshold in this study — in roughly one run in seven of the same configuration. In one clean re-installation, with CPU inference and a newer array library, we observed differences of up to 0.0003. We measured these on the environments available to us and do not claim a

bound for all of them. The residual is in the forward pass rather than the pre-processing: mask reading, resizing and the overlap computation are integer operations, while a prediction lying within rounding distance of the decision threshold can fall either way — across environments, and between runs on the same hardware.

Two dependencies had to be shipped with the package and neither is inferable from a model file: the convolutional block attention variant instantiates a custom layer that is not registered for serialisation, so its weights cannot be loaded without the module that defines it, and that module must be supplied explicitly rather than merely imported.

We report the sequence rather than only its outcome. The stored weights and the recorded numbers were present throughout, and for a period they did not determine the reported values to the precision at which our comparisons operate; four undocumented choices stood between the artifacts and the numbers. This is the same distinction the paper draws elsewhere, applied to itself.

Remaining provenance gaps. The model files record the Keras version under which the ablation runs executed but not the TensorFlow version, and per-epoch training histories were retained for eight of the eleven ablation configurations — not for `U-Net-4L`, `U-Net-3L` or `R50-UNet`. Early-stopping behaviour is therefore reported for eight configurations and documented only in the training code for three. Neither gap affects the comparisons reported here, which are within-study, computed by one module from one set of stored masks, and paired image by image.