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.

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

Gökhan Uçar

gokhan.ucar@bilecik.edu.tr

Bilecik Şeyh Edebali Üniversitesi

Research Article

Keywords: Breast ultrasound, Image segmentation, Deep learning, Attention mechanisms, Data leakage, Statistical power, BUSI

Posted Date: August 21st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10657323/v1

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

Additional Declarations: No competing interests reported.

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

Gökhan Uçar[1*]

[1]*Department of Computer Engineering, Faculty of Engineering, Bilecik Şeyh Edebali University, Bilecik 11100, Türkiye*

[*]Corresponding author. E-mail: gokhan.ucar@bilecik.edu.tr

ORCID: https://orcid.org/0000-0003-1840-8613

Abstract

Background. Reported segmentation results on the Breast Ultrasound Images (BUSI) dataset span a wide range, and the protocols behind them differ in ways not always stated. Whether the architectural differences those results support are large enough to be resolved by the data has not been established.

Methods. We trained eleven U-Net variants under a single evaluation protocol on the 647 lesion-bearing images, using three independent random splits (Monte Carlo cross-validation) with a held-out test partition separate from validation, and computed all metrics from stored masks with one evaluation module. Five attention mechanisms were compared against a common reference and four ablation-axis steps as a separate family, each corrected for multiplicity within its family. For every comparison we report a minimum detectable effect from that comparison's own paired differences, with standard errors on independent clusters rather than images. Parameter cost was counted from the trained model files.

Results. Parameter cost spanned three orders of magnitude across the five mechanisms (12 to 32,595 parameters in the deployed model) while measured gains spanned 0.0020 to 0.0137 Dice. The largest gain came from the cheapest mechanism, efficient channel attention with deep supervision, at 12 parameters. No gain reached the threshold its own comparison could resolve; the largest was 0.52 of a multiplicity-corrected threshold of 0.0262, with win rates between 49.1 % and 56.4 %. Network depth, by contrast, produced a resolved effect of +0.0400. Selecting the reported checkpoint on the evaluation partition rather than a held-out one moved the reported value by +0.0316 to +0.0355 in our own runs. A published lightweight baseline exceeded the best ablation configuration by +0.0150, a margin that its own comparison could not resolve either. An audit of the distributed dataset found 146 near-duplicate pairs; removing them moved absolute Dice by 0.0203 but paired differences by 0.0065.

Conclusions. On this benchmark the parameter cost of an attention module and the effect it produces are not aligned, and the effects lie below what the data can resolve. We report these magnitudes so that the cost of an architectural decision can be weighed against the evidence for its benefit.

Keywords: Breast ultrasound; Image segmentation; Deep learning; Attention mechanisms; Data leakage; Statistical power; BUSI

Background

Breast cancer is the second most frequently diagnosed cancer worldwide and the fourth leading cause of cancer death [1, 2]. Mammography is the established screening modality, but its sensitivity falls in dense tissue, and supplemental ultrasound detects additional cancers at the cost of a higher recall rate [3]. Ultrasound is non-ionising, inexpensive and widely available, which also makes it the practical primary modality where mammography is not.

Its principal drawback is operator dependence, which extends to how lesions are characterised: using the BI-RADS-US lexicon, Park et al. found only fair inter-observer agreement on mass margin ($\kappa = 0.32$) while the same readers were internally consistent ($\kappa = 0.64$) [4]. This is the clinical motivation for automated segmentation, and also a constraint on what any method can achieve — one measured directly in this modality by Webb et al., who reported Fleiss' $\kappa = 0.78$ among three experts against $\kappa = 0.79$ for a deep learning model [5].

Public, pixel-annotated breast ultrasound data is scarce. The Breast Ultrasound Images (BUSI) dataset [6] — 780 images from 600 patients — is one of the few openly available collections with segmentation masks, and has consequently become the de facto benchmark for this task. That reuse has produced a dense table of reported numbers. It has not produced a comparable one.

A recurring pattern in medical image segmentation is that architectural modifications are proposed, evaluated on one or two datasets, and reported as improvements. Isensee et al. report that *details in method configuration have more impact on performance than do architectural variations*, and, examining the top 15 entries to one segmentation challenge, that none of the commonly used architectural modifications — residual connections, dense connections, attention mechanisms or dilated convolutions — represented a necessary condition for good performance [7]; the same group later found that many subsequent claims of superiority did not hold when common validation shortcomings were controlled [8]. Varoquaux and Cheplygina summarised the situation more broadly: *many developments of model bring improvements smaller than the evaluation errors* [9]. The instruments needed to check this are also established — interval-based analyses of reported segmentation performance [10, 11], the pitfall catalogue of *Metrics Reloaded* [12], the non-independence of test cases identified by Reinke et al. [13], and the leakage taxonomy of Kapoor and Narayanan [14]. None of these principles is in question here, and none is claimed as new. What is missing is the magnitude, on a given dataset, set against the cost of the decision it is used to justify.

What we measure

We use one lightweight architecture family, evaluated on BUSI under a fixed, leakage-controlled protocol, to supply those magnitudes. Each contribution below states what is already established and what we add.

1. What five attention mechanisms cost and what they resolve. Efficient channel attention, squeeze-and-excitation, its spatial-and-channel variant, convolutional block attention and an

attention gate were each added to the same reference configuration and evaluated on the same images. Counted from the trained model files, their parameter cost spans three orders of magnitude — 12 to 32,595 parameters in the deployed model — while every measured effect falls below the threshold its own comparison could resolve. The two axes are not aligned: the largest gain came from the cheapest mechanism. Attention modules have been applied to this dataset, including in combination with deep supervision [15, 16], and three further studies in the sample we audit here use an attention-module architecture; the cost set beside the resolvable effect is what we add.

2. The magnitude of one protocol choice. Reporting on the model-selection partition rather than a held-out one inflated the reported value by +0.0316 to +0.0355 in our own runs, measured on two architectures. That such reporting inflates a result is established [14, L1.1]; the comparison with the architectural effects above is what is new here, and it is decisive.

3. What duplicate records displace, and what they do not. We detect near-duplicate records in BUSI, quantify how many reach the test partition, and separate their effect on absolute scores from their effect on paired differences. That duplicates constitute leakage is established [14, L1.4] and duplicates in BUSI have been documented [17]; the decomposition is what makes the consequence actionable.

4. A documented case in which artifact-level evidence and code-level verification disagreed about test-set leakage. Artifacts and code can disagree about what a pipeline did [18], and reproducing a number does not establish that the process was sound [19]. We report a case in which the disagreement runs in the direction that manufactures a finding, and to the best of our knowledge we are not aware of a prior report of an artifact-level audit producing a false-positive leakage finding.

To place these magnitudes in context we also audited 22 studies reporting a numerical segmentation metric on BUSI under a stated protocol. These are counts over the studies we examined, not estimates of a population proportion.

This work is deliberately narrow: lightweight, edge-deployable segmentation models on a single dataset. BUSI was chosen for a reason connected to the paper's own subject — its normal class supplies empty reference masks, which gives an otherwise invisible aggregation choice measurable leverage. The mechanisms we measure are not specific to BUSI, and we did not measure them elsewhere. The architecture is an instrument for these measurements, not a proposed contribution.

Methods

This section describes the dataset and the audit of its record independence, the evaluation protocol, the eleven ablation configurations and their parameter cost, the baselines and metrics, and the statistical analysis. It closes with the audits applied to our own pipeline and to the literature sample.

Dataset

We used the Breast Ultrasound Images (BUSI) dataset [6], a publicly available collection of B-mode images from 600 women in three classes, each with an expert-annotated binary lesion mask. All experiments use the 647 lesion-bearing images (437 benign, 210 malignant); the 133 normal images were excluded. Images and masks were resized to 256×256 pixels. The dataset is distributed at the level of the image rather than the patient and carries no patient identifier, which bears on how test observations are treated under *Statistical analysis*.

The exclusion of the normal class is a protocol choice with measurement consequences, and we state it because published results on BUSI are not uniform in this respect. For an image with an empty ground-truth mask the Dice coefficient is degenerate — approximately 1.0 for an empty prediction, 0 for any non-empty one — so a mean over a set including normal images is influenced by a model's empty-prediction rate independently of its segmentation quality. Restricting evaluation to lesion-bearing images removes this term from every number we report, following *Metrics Reloaded* [12]. All counts reported here are taken from the distributed directories rather than from the dataset publication, which describes the release at the level of patients.

Audit of record independence

Evaluation assumes test records are independent of training records. A curated release of BUSI reports repeated images in the original distribution [17], and duplicates in datasets are a documented leakage channel [14, L1.4]. Because this paper reports on protocol control, we audited our own copy rather than assuming the channel absent.

The audit hashes each file, each decoded greyscale array and each mask separately, and flags near-duplicate pairs by 64-bit difference hash at a Hamming distance of at most 5. Because perceptual hashing can flag unrelated pairs in low-texture greyscale imagery, the threshold was calibrated against a null distribution of randomly drawn pairs rather than adopted from convention. The pair list is then intersected with the test assignments of every split. Full procedure and outputs are in the released code.

Duplication versus augmentation. The two are distinct. The operation is similar; the difference is its position relative to the split. Augmentation applied after splitting places all copies on one side and does not affect measurement. Repeated records present before splitting can place copies on both sides, which shifts the test mean and makes its variance smaller than the nominal sample size implies. We did not retrain on a deduplicated release; the measurements reported here are made on the dataset as distributed, which is also the condition under which the published values we compare against were produced.

Evaluation protocol

All configurations, retrained baselines and the foundation-model reference were evaluated under one protocol, fixed before any comparison reported here was run. Figure 1 summarises the

configurations and the partitioning.

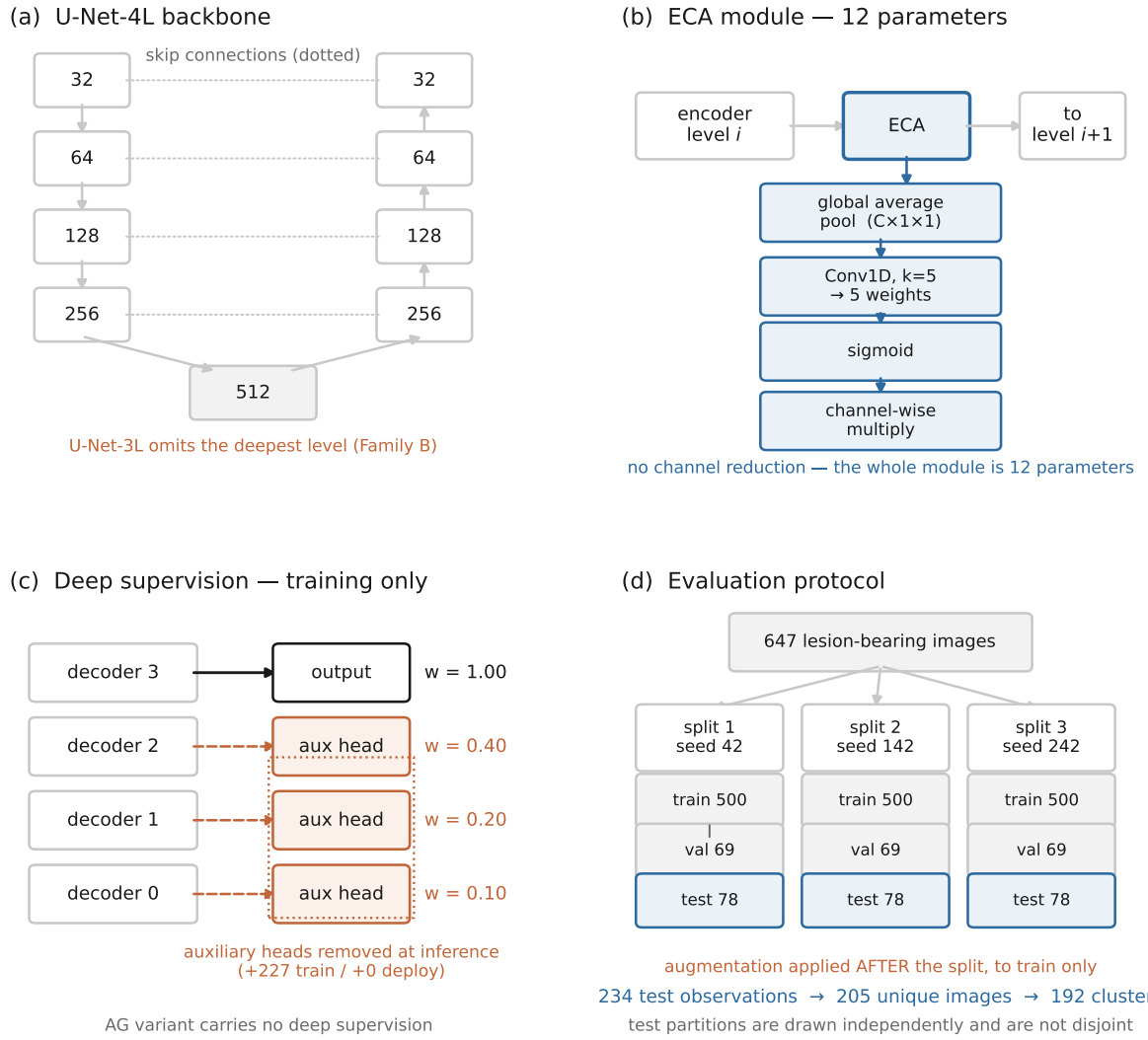


Figure 1: Configurations and partitioning. Monte Carlo cross-validation: three independent random splits. Standard errors are computed on the 192 clusters.

Splitting. Three independent random splits — Monte Carlo cross-validation — with seeds 42, 142 and 242. Each assigns 12 % of the 647 images to test and 12 % of the remainder to validation, giving 500 training / 69 validation / 78 test images. The design decouples the number of repetitions from the test fraction, which a partitioned k-fold scheme fixes to $1/k$. Test partitions are drawn independently and are therefore not disjoint: the 234 test observations comprise 205 unique images, 27 of them appearing in more than one test partition. All configurations share the same three partitions, so every comparison is paired on identical images. Splitting is performed before any augmentation, which is applied to the training partition only, in two layers: an offline pass writing two augmented copies of every training image (500 → 1,500), then online augmentation during training.

Model selection and stopping. Early stopping with patience 20 and an epoch cap of 200; checkpoint selection on the validation partition, every reported number from the held-out test partition. The monitored quantity is not the same for all configurations: binarized validation

Dice in the four deeply supervised ones (12/12 runs), validation loss in three others (9/9). This heterogeneity is not neutral — selecting on validation Dice is the more favourable criterion for the metric we report, and it is the criterion used by the attention configurations rather than by their reference. The asymmetry works against the null result reported here. Per-image metrics were retained for every model on every test image, giving n = 234 paired observations per comparison.

Ablation configurations and their parameter cost

Eleven configurations were trained under the protocol above, forming two comparison families defined in advance. Table 1 gives them with their family membership and the parameter cost of each mechanism. The mapping from the configuration names used here to the internal identifiers of the released per-image tables is given in `configuration_map.csv` in the code record, so each row can be traced to the file it was computed from.

Table 1: The eleven configurations, the comparison family each belongs to, and the parameter cost of each attention mechanism in the deployed model.

Configuration	Family / role	Deployed parameters	Added (deployed)
U-Net-3L	B — depth	—	—
U-Net-4L	B — baseline depth	—	—
R50-UNet [20]	B — encoder	—	—
U-Net-4L + Aug×2	B; reference for Family A	2,143,905	—
U-Net-4L + Aug×3	B — augmentation saturation	—	—
R50-UNet + Aug	B — augmentation on encoder	—	—
+ ECA + DS [21]	A	2,143,917	+12
+ SE + DS [22]	A	2,154,913	+11,008
+ CBAM + DS [23]	A	2,155,305	+11,400
+ scSE + DS [24]	A	2,155,393	+11,488
+ AG [25]	A	2,176,500	+32,595

Family A rows are mechanisms added to U-Net-4L + Aug×2; Family B rows are independent configurations. All derive from a U-Net backbone [26]; the deep-supervision variants [27] use three supervision levels and four output heads, with loss weights 1.0 on the final output and 0.4, 0.2 and 0.1 on the auxiliary heads. Optimizer, schedule, batch size, loss and per-configuration settings are given in the supplementary material (S4b); one configuration was trained at a tenth of the learning rate used by the others, and that difference is stated there rather than absorbed.

Parameter counts. Counts depend on which form of a deeply supervised model is described, so we state the form. The column above is the deployment form, in which supervision heads are removed and a single output remains; mechanism cost is that figure minus the Family A reference. In training form each deeply supervised variant carries a further 227 parameters that are removed at inference. Counts were taken from the trained model files, and the three splits agree exactly. Parameter cost spans three orders of magnitude across the five mechanisms, from 12 to 32,595.

The cost axis and its limit. Parameter count is the cost measure available here. Whether it predicts inference latency, or the behaviour of the model under quantization, is a separate question, and we do not address it.

Why these mechanisms. They are not untried on this dataset. Combining deep supervision with channel attention has been published and evaluated on BUSI itself [15], so one variant we evaluate — SE with deep supervision — is that published combination; three of the five have been evaluated on BUSI as attention baselines in a published comparison [16]; and the ECA-with-deep-supervision pairing has been published for liver CT [28]. Those two reports share a first author, and both appear in the literature sample audited here, so we give a count that excludes them: three of the remaining studies in that sample of 22 use an attention-module architecture, one of them the same scSE module we evaluate.

Baselines, foundation-model reference and metrics

Two lightweight architectures were retrained from scratch under the protocol above: UNeXt [29] (1,472,952 parameters) and CMUNeXt [30] (3,163,116), counted from the trained checkpoints. CMUNeXt was designed for ultrasound and operates at our input resolution. Hyperparameters were taken from the original publications and not tuned; full values, the parity audit, the single-normalization sensitivity arm and the reasons a third baseline was removed before training are in the supplementary material (S4).

Training loop. The loop is ours, not the reference repositories'. This was necessary rather than incidental: the reference loops embed their own evaluation protocols — a single split, no held-out partition separate from validation, no pre-augmentation stage — which is precisely the variable under study. Model definitions, optimizer, learning rate, loss and scheduler were taken unchanged; only the data loader and training-control logic were replaced.

SAM ViT-B [31] was evaluated zero-shot on the same three test partitions using a box prompt derived from the ground-truth mask. This is an oracle upper bound and not a competing method: the prompt supplies the lesion location that the automatic models must infer. General SAM was used rather than a medically fine-tuned variant, which makes the value an *under*estimate of the bound.

Because metric implementations differ across papers, every model in this study writes binarized prediction masks to disk and all metrics are computed from those masks by a single evaluation module. Predictions were binarized at 0.5; Dice and IoU use a smoothing constant of 1e-6 and are defined in the supplementary material (S3). Boundary metrics — HD95 and ASSD [32] — were computed with medpy 0.5.2 [33] in pixels; both behave differently from overlap measures on small or fragmented structures [34] and are undefined when either mask is empty, so image–model pairs with an empty prediction are excluded from boundary means and counted separately.

The checkpoint-selection measurement. One quantity reported here is not produced by the protocol above but by deliberately departing from it. For each retrained baseline we record two values from the same run: the best epoch scored on the partition that is then reported, and the value of the held-out test partition under our own protocol. Their difference is the optimism introduced by selecting the reported checkpoint on the reported partition. It is a difference between two partition means rather than a paired per-image difference, so it carries no paired interval, no win rate and no comparison-specific detectable effect, and it is not scaled against the

thresholds computed for the ablation comparisons. We have three such measurements, from two architectures and two pre-processing arms.

Statistical analysis

All primary comparisons are paired at the image level (n = 234), each variant against its reference on the same image in the same split. The Wilcoxon signed-rank test [35] is the primary test throughout, because the distribution of paired per-image Dice differences is heavy-tailed and skewed; the choice was fixed before the comparisons were run. The paired t-test is reported as a secondary quantity and does not override the primary verdict; where the two diverge this is reported rather than resolved.

Two families were defined in advance and Holm–Bonferroni correction [36] is applied within each: Family A, five attention mechanisms added to U-Net-4L + Aug×2 (Holm-5); Family B, four steps along the ablation axis (Holm-4). No correction is applied across families and effects from the two are not ranked against one another, because they are corrected at different levels and each step starts from a different baseline. Boundary metrics are a secondary check, corrected within metric and not across: Dice, HD95 and ASSD are three summaries of the same pair of masks, not independent hypotheses.

Each comparison is reported with its mean paired difference, a 95 % confidence interval computed on the same 192 independent clusters as the thresholds, and the win rate — the proportion of test images on which the variant exceeds the reference. Both are reported for every comparison because they can point in different directions: a variant can have a positive mean carried by the magnitude of its wins rather than their frequency. Neither statistic is reported without the other. IoU is not independent evidence: at the per-image level it is an exact monotone transform of Dice, IoU = Dice/(2 − Dice), with maximum deviation 1.3e-6 across our measurements, so the two are read as a single result; the transform is verified in the supplementary material (S3b).

Detection threshold. For each comparison we report the minimum detectable effect (MDE) at 80 % power, two-tailed, from that comparison's own paired differences as MDE = $(z_{1-\alpha/2} + z_{0.80}) \cdot \mathrm{SD}/\sqrt{n}$. It is computed per comparison rather than assumed constant, because that standard deviation varies with how similar the two configurations are — by a factor of 2.8 across our nine comparisons; the per-comparison values are in the supplementary material (S1). An MDE is the interval width used elsewhere for segmentation metrics [10, 11] under a different parametrization. The corrected column uses the most stringent level of each comparison's own family, so no single threshold applies to all comparisons.

Correction for non-independent test records. The MDE formula assumes independent observations; non-independence of test cases is a documented pitfall whose recommended response — aggregating at the level of the independent unit — is set out with a worked example by Reinke et al. [13]. Here the 234 observations reduce to 205 unique images and those to 192 clusters once near-duplicate records are grouped. We treat the cluster as the unit of information, equivalent to multiplying the nominal standard error by $\sqrt{234/192} = 1.104$. Every threshold reported in this paper is computed on this basis; thresholds assuming 234 independent observations are not

reported. The correction is applied to the standard error, not to the effect estimates. The unit is an approximation: BUSI distributes no patient identifiers, so patient-level grouping is impossible [37], and image similarity errs in both directions — it groups images a patient-level split would group, but not two views of the same lesion that fall outside the threshold.

Post-hoc sample size and hierarchical testing. Figures back-calculated from an observed effect are fragile, so we report a required-sample-size figure only for the comparison whose point estimate supports one. One Family A comparison was proposed chronologically before the others, so treating it as a pre-specified primary endpoint would have been defensible; we did not, because no formal pre-registration exists. We state the option and report the corresponding uncorrected value alongside the corrected result.

Audits of our own pipeline and of the literature

Because this paper reports on the protocol reporting of others, we audited our own pipeline at the level of source code rather than by inspecting the artifacts it left on disk, verifying for each configuration notebook that the split precedes augmentation, that the pre-augmentation function receives and iterates over the training partition only, and that training consumes its in-memory return value. As reported in *Results*, the two strategies disagree on our own pipeline, and the artifact inspection is the one that is wrong. Confirming that stored weights reproduce a recorded metric establishes that weights and number correspond, not that the process was sound [19]; published leakage-detection tooling accordingly operates on source code [38].

To characterize the protocols behind published BUSI results we assembled and audited a sample of studies under a stated protocol; the record-level table is released with the code. Two strata were defined — studies from 2024 onward, and the most-cited studies with no date restriction — and every entry was verified against the full text or the publisher record rather than a metadata index. Full eligibility rules, search routes and the stopping criterion are in the supplementary material (S5). For each study we record only what the paper or its released code states explicitly. An item a study does not state is recorded as not stated, which is an observation about the record and not about the study. The sample supports existence claims about the heterogeneity of reported protocols; it does not support estimates of prevalence, and none is made.

Software and hardware

Ablation configurations were implemented in Keras 3.13.2 — the version recorded in the metadata of the released weight files — with the TensorFlow backend; the environment in which the released reproduction package was verified is TensorFlow 2.21.0. The retrained baselines were implemented in PyTorch 2.7.1 (CUDA 11.8) in a separate environment. Boundary metrics use medpy 0.5.2. The retrained baselines and the foundation-model reference ran on an NVIDIA RTX 2070 (8 GB); the hardware used for the ablation training runs was not recorded. Code and the per-image measurement tables required to reproduce every number, with the mapping from the labels used here to the internal identifiers, are available at https://doi.org/10.5281/zenodo.21789807.

Results

Results follow the order in which the protocol produces them: descriptive values for all eleven configurations, then the paired comparisons within each family, then the record-independence and pipeline audits, the lightweight baselines, and the protocol-reporting audit of the literature sample.

What the five attention mechanisms cost and what they resolve

All eleven configurations were trained and evaluated under the single protocol described in *Methods*. Table 2 reports the descriptive result for each; values are means over the three held-out test partitions (234 observations), computed by the shared evaluation module from stored prediction masks. Figure 2 shows the per-image distributions behind those means.

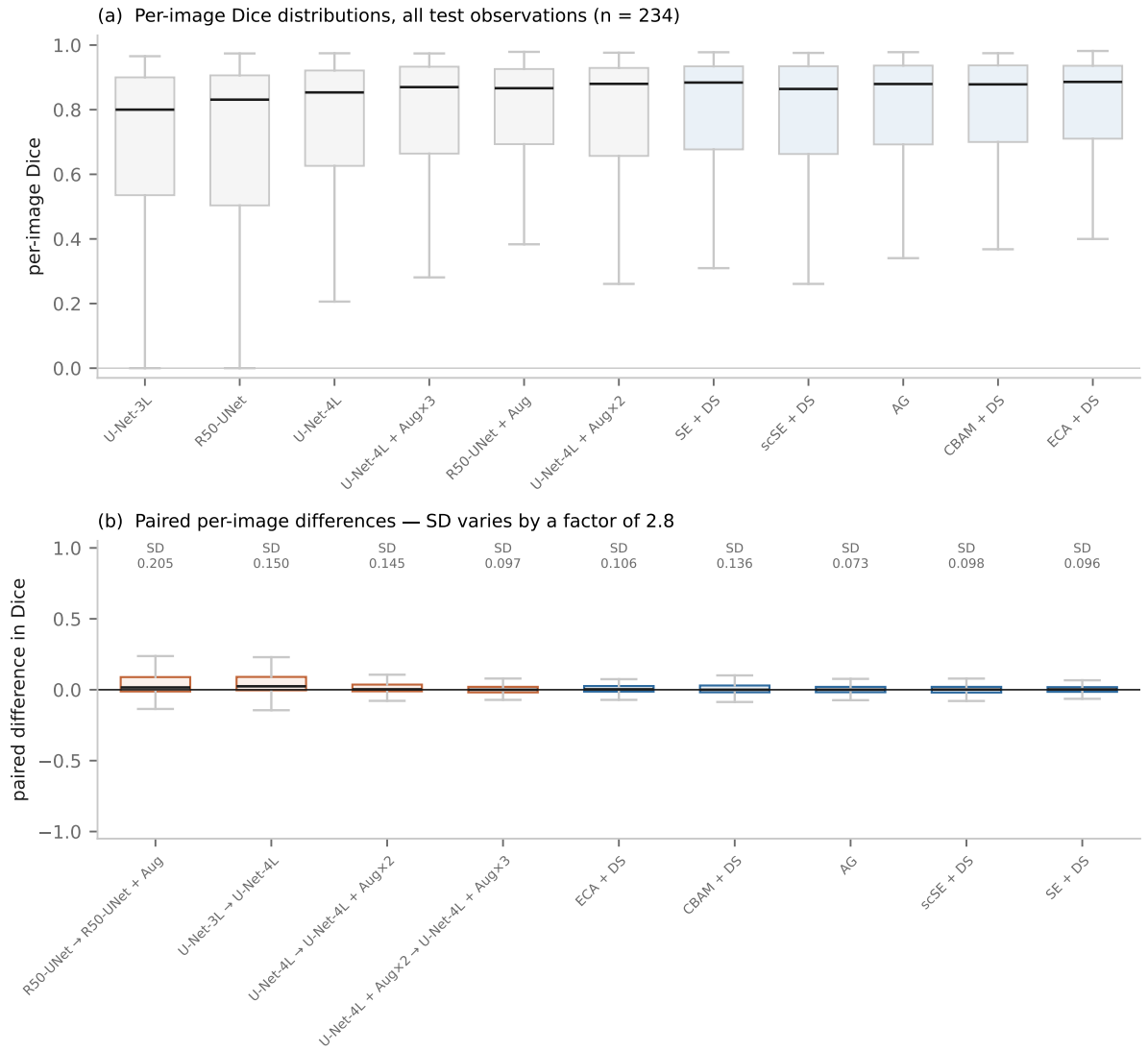


Figure 2: Per-image distributions behind the means of Table 2, and the paired differences on which the comparisons rest. Distributions span the full range in every configuration; the SD of paired differences is why detection thresholds are computed per comparison.

Table 2: Ablation configurations under one protocol. Boundary metrics are in pixels; lower is better. n(empty) is the number of test images on which the model predicted no foreground. Added parameters are counted in the **deployed form** of each model, relative to the Family A reference. ‡ Counts are for the deployed form, in which the supervision heads are removed and a single output remains. In training form each deeply supervised variant carries a further 227 parameters, removed at inference. The attention gate carries no deep supervision, so its training and deployed forms coincide; the same holds for the Family A reference.

Configuration	Added parameters (deployed)	Dice	IoU	HD95	ASSD	n(empty)
R50-UNet	—	0.6748	0.5827	37.18	16.52	6
U-Net-3L	—	0.6838	0.5791	52.55	18.16	1
U-Net-4L	—	0.7238	0.6315	36.78	14.47	1
U-Net-4L + Aug×3	—	0.7376	0.6506	35.05	13.40	2
R50-UNet + Aug	—	0.7382	0.6504	32.32	12.53	4
U-Net-4L + Aug×2	*reference*	0.7433	0.6543	34.53	12.86	2
+ SE + DS	+11,008	0.7453	0.6590	37.18	14.07	1
+ scSE + DS	+11,488	0.7463	0.6580	35.15	13.48	0
+ AG	+32,595	0.7486	0.6596	34.01	12.42	1
+ CBAM + DS	+11,400	0.7515	0.6641	33.34	13.16	2
+ ECA + DS	+12	0.7570	0.6713	31.87	11.75	0

Among the five attention mechanisms, efficient channel attention with deep supervision produced the largest measured gain over the common reference (+0.0137 Dice) at the lowest parameter cost: 12 parameters added to the deployed model, 0.0006 % of its total. The squeeze-and-excitation variant, at 11,008 parameters, produced +0.0020, and the attention gate, at 32,595 parameters, produced +0.0053. Across the five mechanisms parameter cost spans three orders of magnitude — a factor of 2,716 — while the measured gains span 0.0117 Dice. The two axes are not aligned: the cheapest mechanism produced the largest gain and the most expensive produced less than half of it. Figure 3 places every configuration on both axes.

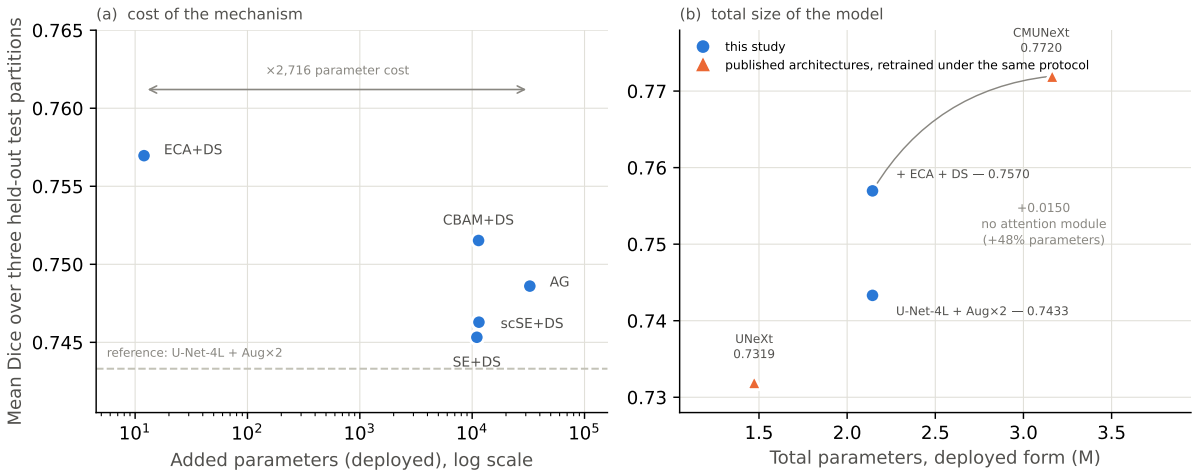


Figure 3: (a) Parameters each mechanism adds to the deployed model, against mean Dice. The x-axis is logarithmic because the costs span three orders of magnitude (12 to 32,595); the reference configuration adds none and is drawn as a horizontal line rather than a point. Labels match the configuration names used in Tables 2 and 3: deep supervision adds a further 227 parameters in training form only, removed at inference, and the attention gate carries none. (b) Total deployed size against mean Dice, with the two published baselines retrained under the same protocol. No effect in either panel clears the threshold its own comparison could resolve (Figure 4); selection asymmetry and the capacity difference are discussed in the text.

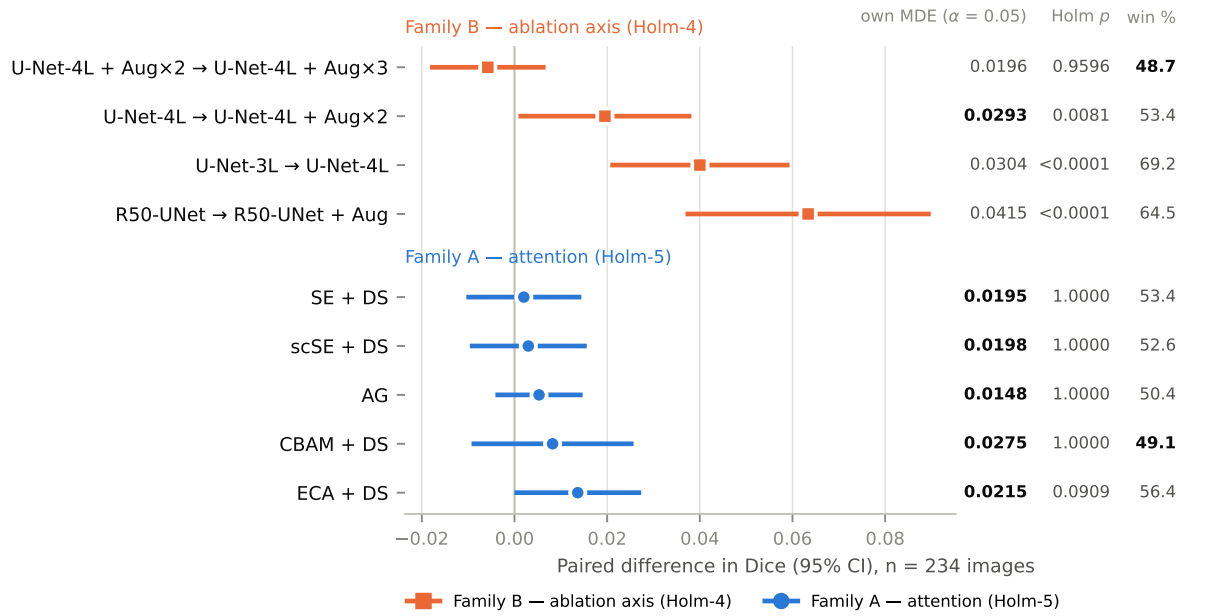


Figure 4: Each effect against the threshold its own comparison could resolve. Family A effects are relative to U-Net-4L + Aug×2; Family B rows are independent configurations. Detection thresholds are per comparison (80 % power); bold marks an effect below its own threshold.

Table 3 gives the paired comparisons: each mechanism against the same reference on the same images, with Holm–Bonferroni correction over the five. Figure 4 places each effect against the threshold its own comparison could resolve, and Figure 5 shows the reference and the ECA + DS variant on the image at the median of their paired difference in each split, and on the image with the largest paired difference across all three.

Table 3: Family A — mechanisms added to **U-Net-4L + Aug×2**. Paired per-image differences in Dice, 234 observations. Win rate is the proportion of test images on which the variant exceeds the reference. Added parameters are in the deployed form (see the footnote to Table 2).

Variant	Added parameters (deployed)	Effect	95 % CI	Win rate	Wilcoxon p	Holm-5 p
+ ECA + DS	+12	+0.0137	[−0.0015, +0.0288]	56.4 %	0.0182	0.0909
+ CBAM + DS	+11,400	+0.0082	[−0.0111, +0.0275]	49.1 %	0.2639	1.0000
+ AG	+32,595	+0.0053	[−0.0051, +0.0157]	50.4 %	0.4804	1.0000
+ scSE + DS	+11,488	+0.0030	[−0.0110, +0.0169]	52.6 %	0.5007	1.0000
+ SE + DS	+11,008	+0.0020	[−0.0117, +0.0157]	53.4 %	0.3331	1.0000

Family A. No variant exceeds the reference at the corrected threshold. The largest effect has an uncorrected Wilcoxon $p = 0.0182$ and a Holm-corrected $p = 0.0909$, and its confidence interval includes zero. All five mean effects are positive, and taken alone that direction consistency would suggest a shared effect; the win rates qualify this substantially. They span 49.1 % to 56.4 %, and the variant with the second-largest mean effect wins on fewer than half of the test images (49.1 %) — its positive mean is produced by the magnitude of its wins rather than their frequency. Mean effect and win rate are two statistics computed from the same measurements, and here they tell different stories.

Dice and IoU are read as a single result (*Methods*). The transform moves the corrected p-value for the ECA + DS variant across the conventional threshold — 0.0909 on Dice, 0.0498 on IoU —

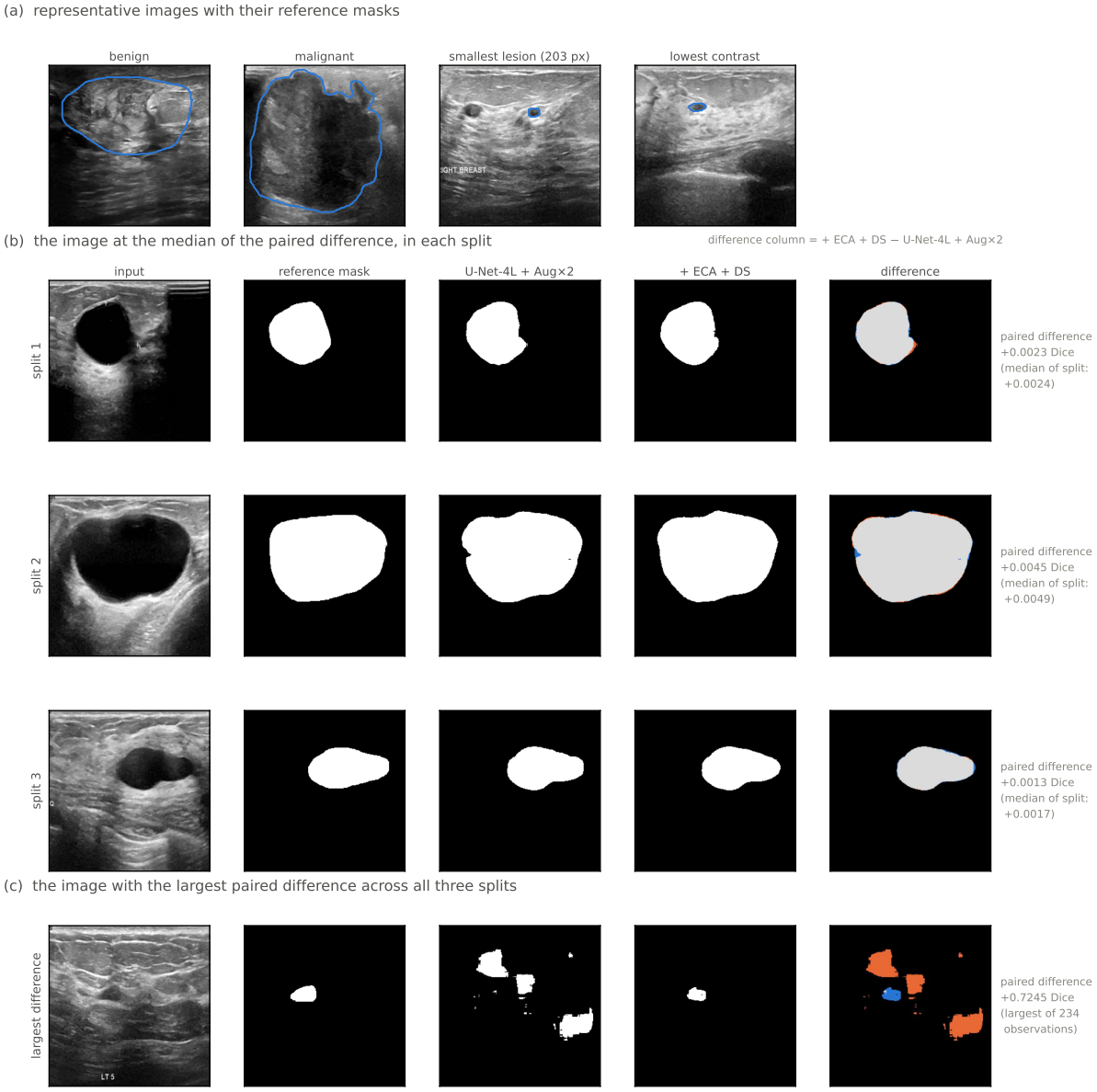


Figure 5: (a) Representative images from the dataset with their reference masks; a display of scope, not a measurement. (b) The image at the median of the paired difference between U-Net-4L + Aug×2 and the + ECA + DS variant in each split; by construction one on which the two configurations largely agree. (c) The image with the largest paired difference across all three splits; on that image the reference produces scattered false positives. Large single-image differences of this kind are consistent with the small mean effect: the paired differences for this comparison have a standard deviation of 0.106 (Figure 2b). In the difference column, grey is agreement, blue is predicted by the variant only and orange by the reference only. Both selections were made from the recorded per-image tables before the masks were regenerated; regenerated Dice matched the recorded values exactly ($|\Delta| = 0.000000$). Images reproduced under CC BY 4.0; the predicted masks are derived works of the same release.

and the corrected values for IoU (0.0498) and ASSD (0.0503) differ by 0.0005. Reporting either as a significant finding while the others are not would be a selection among metrics, not a result, and we do not do it; the defensible statement is that this effect sits at the resolution limit of the design. The ECA + DS comparison was proposed chronologically before the others, and treating it as a pre-specified primary endpoint would have yielded an uncorrected $p = 0.018$; we report the corrected result and state the alternative so the reader can apply either standard.

What this design could have detected, and where the literature sits

Whether a null result is informative depends on what the study could resolve, so we report the detection threshold alongside the result rather than after it; Figure 6 relates a difference to the number of independent units needed to detect it.

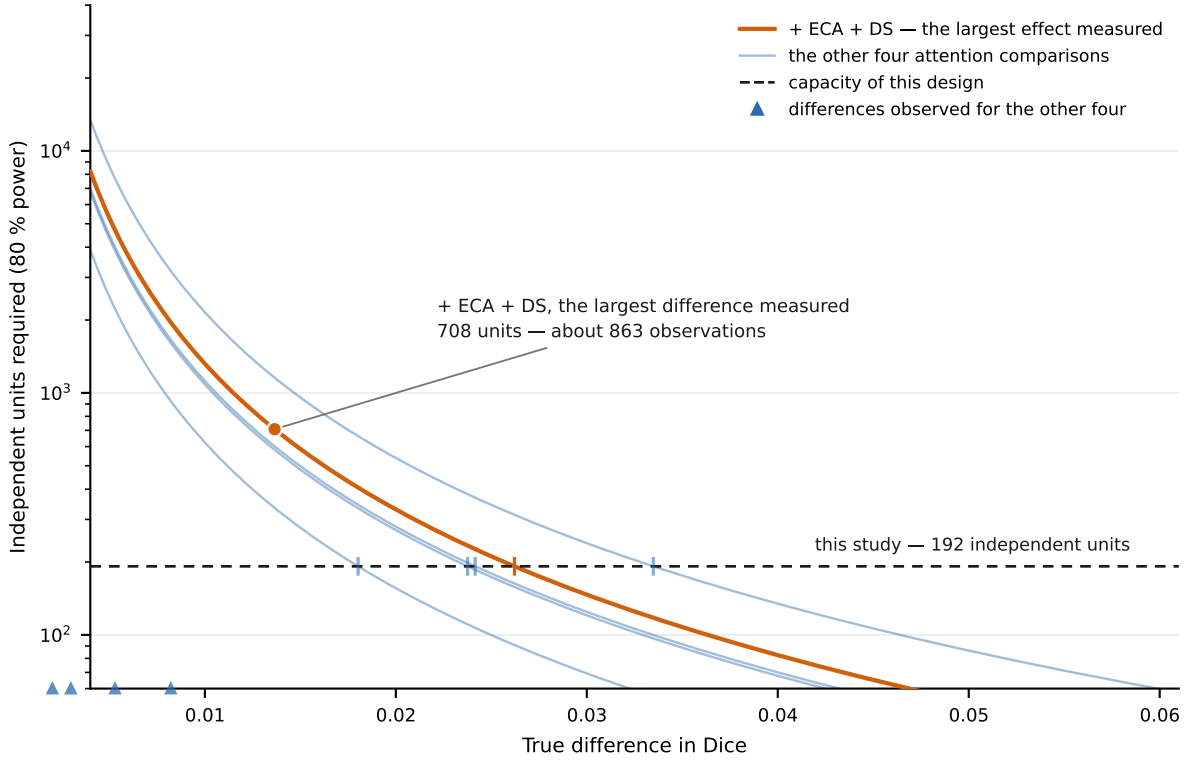


Figure 6: Units required to detect a given difference at 80 % power, two-tailed, after multiplicity correction within the family, using each comparison's own paired standard deviation. Ticks on the dashed line are each comparison's detectable difference at the present size. A required size is given for one comparison only; for the others the observed difference is marked on the axis alone.

For the ECA + DS comparison the observed standard deviation of paired per-image differences is 0.1063. Computed on the 192 independent clusters described in *Methods*, this gives a standard error of 0.00767 and an MDE of 0.0215 at $\alpha = 0.05$ and 0.0262 at the Holm-corrected level. The largest observed effect, +0.0137, is 0.52 of the corrected threshold. Treating clusters rather than images as the unit of information raises every threshold by 10.4 %; no verdict changes and no effect changes side relative to its own threshold (S1) — not because the comparisons are unusually robust, but because the correction is modest relative to the margins, the smallest of which is 32 % of its uncorrected threshold; read against the corrected thresholds from which the

verdicts are taken, the same margin is 10 %.

Detecting an effect of the observed size at 80 % power after correction would require approximately 708 independent units, about 3.7 times what this design provides. We report this figure for one comparison only; for the remaining four the confidence intervals span zero and a back-calculated sample size is not stable enough to quote. A single 80/20 split on this dataset is weaker still, at an MDE of 0.0270. The MDE varies across comparisons because it is a function of each comparison's own standard deviation, which ranges from 0.0731 to 0.2055 — a factor of 2.8 — giving thresholds from 0.0148 to 0.0415 at $\alpha = 0.05$. There is consequently no single ruler against which all comparisons can be scaled.

This places the effect we report in a specific context. Among the open-access BUSI segmentation studies we could retrieve in full text, 49 numerical Dice improvement margins were stated in running text across 19 studies. Roughly half of them (47 %) fall below the corrected threshold this comparison could resolve, and roughly a quarter (27 %) fall below the effect we report here. These are counts over the records we retrieved, not estimates of a population proportion. The sample is a minority slice and we do not read it as typical practice: most studies proposing an architecture state their margin only inside a table, and margins reported that way are not captured by this extraction. The retrieval route, the extraction rule and its audited accuracy are given in the supplementary material (S5b), which also states why this sample is separate from the 22-study protocol audit reported below.

Ablation axis: what is resolved

The four comparisons along the ablation axis were corrected separately (Holm-4), as a pre-defined family distinct from Family A. Table 4 reports each with its paired difference, interval, win rate and detection threshold.

Table 4: Family B — ablation axis. Paired per-image differences in Dice, 234 observations.

Comparison	Intervention	Effect	95 % CI	Win rate	Holm-4 p	Resolved (Wilcoxon, Holm-4)
`R50-UNet` → `R50-UNet + Aug`	augmentation on the ResNet-50 encoder	+0.0634	[+0.0341, +0.0926]	64.5 %	0.0000	yes
`U-Net-3L` → `U-Net-4L`	depth, 3 → 4 levels	+0.0400	[+0.0186, +0.0614]	69.2 %	0.0000	yes
`U-Net-4L` → `U-Net-4L + Aug×2`	augmentation ×2	+0.0195	[−0.0011, +0.0402] §	53.4 %	0.0081	yes
`U-Net-4L + Aug×2` → `+ Aug×3`	augmentation ×3 (saturation)	−0.0058	[−0.0195, +0.0080]	48.7 %	0.9596	no

§ The interval is parametric and the verdict is not: for this comparison the paired t-test does not clear its corrected threshold while the primary rank test does, so the interval and the verdict come from different tests. This is the divergence described below, seen as an interval rather than as a p-value.

Three comparisons are resolved. Adding a network level produced the largest resolved architectural effect in this study, +0.0400 at a win rate of 69.2 % (162 wins, 7 ties, 65 losses) — approximately three times the largest attention-related effect, which was not resolvable. Encoder augmentation is larger still (+0.0634, 64.5 %) but changes the training data regime rather than the architecture.

A third augmentation pass. This produced no improvement: the point estimate is negative (−0.0058), the win rate is below chance (48.7 %), and the comparison does not clear its threshold.

Moving from two to three offline augmentation passes takes the pre-augmented training set from 1,500 images to 2,000 — a third more storage and training time — and returns nothing measurable — a cost with no accompanying resolved benefit, on the same axis on which the attention mechanisms are evaluated.

Divergence between tests. For `U-Net-4L` → `U-Net-4L + Aug×2` the Wilcoxon test clears the corrected threshold (0.0081) while the paired t-test does not (Holm-4 0.0811). The cause is the shape of the difference distribution, which is heavy-tailed and skewed; under those conditions a rank-based test has more power. The MDE is computed from the parametric standard error and therefore describes the power of the t-test. The verdict is read from the primary test, and the ruler and the verdict are not on the same scale. We report the divergence rather than resolving it by selecting the more favourable test.

We do not rank Family A effects against Family B effects: the families are corrected differently, were defined to answer different questions, and the configurations they compare do not share a common baseline.

Boundary metrics. These did not produce a finding separate from Dice: the scenario the design allowed for — a mechanism leaving Dice unchanged while improving boundary agreement — did not occur. Full paired results are in the supplementary material (S2).

Records that are not independent

The audit of the distributed release returned the counts in Table 5.

Table 5: Non-independent records in the 647 lesion-bearing images.

Quantity	Value
Byte- and pixel-identical pairs	1
Near-duplicate pairs (dHash, Hamming ≤ 5)	146
Unique images involved	241 of 647
Median mask IoU of flagged pairs	0.844
Flagged pairs with identical pixel dimensions	0 of 146
Pairs whose members carry different class labels	8
Test images with a counterpart on the training side	26.9 % – 38.5 % per split

Three flagged pairs are shown in Figure 7. The threshold was calibrated rather than adopted: the median mean-absolute pixel difference of randomly drawn pairs is 46.5 against 10.1 for the flagged pairs, and **every flagged pair falls below the first percentile of the random distribution**. The flagged images are not byte-level copies — none shares its pixel dimensions with its counterpart — but their reference masks agree closely (median IoU 0.844), which is the combination that matters for evaluation: **the records are not independent, whatever produced them.** Eight pairs carry different class labels, three at a perceptual-hash distance of zero, which a difference in acquisition cannot account for. **An independent audit of this dataset, using structural similarity and manual review rather than perceptual hashing, gives as its worked example a triplet of repeated images [17]; two of our flagged cross-class pairs are exactly those records** — two methods with no implementation in common converging on the same images.

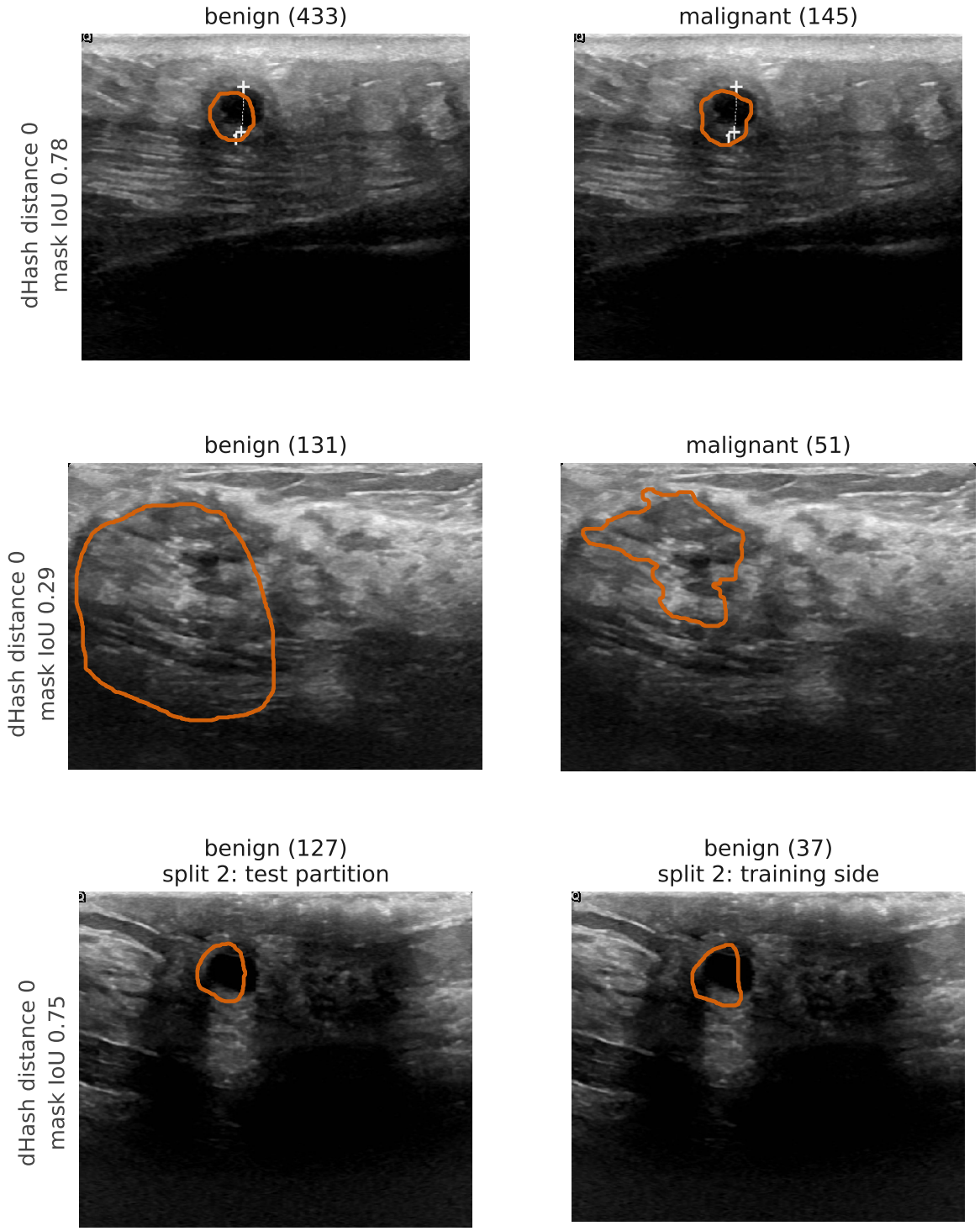


Figure 7: Three of the 147 duplicate and near-duplicate pairs found in the distributed BUSI release (1 exact, 146 near). Rows 1 and 2 carry different class labels; row 3 straddles a split boundary. Contours are the distributed reference masks. The dataset publication states that "the data included duplicated images that required to be removed" during its preparation; the pairs shown here are present in the release as distributed. Images reproduced under CC BY 4.0.

Effect on the measurements reported here. Recomputing the headline comparisons on the duplicate-free subset shifts the absolute Dice of all eleven configurations downward by 0.0203 on average, and shifts the paired differences by 0.0065 — below the smallest threshold in this study (0.0148). The absolute level is optimistic; the paired comparisons, which are what this paper reports, are largely unaffected.

We give the sensitivity analysis rather than only its summary. Recomputing all nine pre-specified comparisons on the subset that excludes every image belonging to a flagged pair gives shifts of 0.0005 to 0.0120, median 0.0056. No comparison changes sign, and the largest shift is below the smallest threshold in this study. The subset is also smaller, so its own thresholds are higher: excluding every flagged image leaves roughly two thirds of the clusters and raises each threshold by about a fifth. We therefore read the shift against the full-sample thresholds and report the subset as a sensitivity check rather than as a second set of verdicts. The two largest shifts are on the ablation axis (`U-Net-3L` → `U-Net-4L`, 0.0120; `U-Net-4L` → `+ Aug`×2, 0.0093), and both remain on the same side of their own threshold.

Pairings that are not among the nine — a configuration compared directly against `R50-UNet + Aug` across families, for instance — do change sign on this subset. Those pairings are not reported in this paper, because the two families are corrected separately and are not ranked against one another; we state the instability here so that the released per-image tables are not used to construct such a comparison without it.

A retracted reading. A first reading of these data was wrong and is reported as such. We initially read the score premium on leaked test images, which varied by a factor of 5.3 across configurations, as an architecture-dependent memorization signature. Stratifying by class and then by lesion area removes the premium entirely and reverses the ordering; it was a lesion-size composition artifact. Duplicate leakage does not favour the attention configurations.

Lightweight baselines and the effect of one protocol choice

Two published lightweight architectures were retrained from scratch under this protocol and evaluated with the same metric code. Table 6 places them against the ablation configurations.

Table 6: Baselines and ablation configurations under one protocol. All values measured in this study from stored prediction masks; published values are not shown in this table.

Model	Params (training form)	Mean Dice	IoU	HD95	ASSD
CMUNeXt	3,163,116	0.7720	0.6876	30.41	11.52
+ ECA + DS (this study)	2,144,144	0.7570	0.6713	31.87	11.75
U-Net-4L + Aug×2 (this study)	2,143,905	0.7433	0.6543	34.53	12.86
UNeXt	1,472,952	0.7319	0.6471	34.86	13.91

Parameter counts in this table are in the training form, which is the form the two baselines are published in; the deeply supervised configuration therefore appears here 227 parameters larger than in Table 2, where counts are in the deployed form. The comparison is unaffected: the capacity difference is 47 % in either form.

The baseline margin. CMUNeXt exceeds the best configuration of our ablation by +0.0150. We report this as measured, and two observations bear on how it should be read. The selection asymmetry runs against our own configuration: the ECA + DS variant was selected as the best of eleven evaluated on these same partitions and its training callback went through three iterations, whereas the baselines were each trained once with published hyperparameters and no selection advantage — on an equal-selection basis the margin is likely larger, not smaller. And CMUNeXt contains no attention mechanism: a source-level scan of the reference implementation returns no SE, CBAM or ECA module. It also carries 47 % more parameters, so this is not a comparison at matched capacity.

Resolution of that margin. It is itself below what this design can resolve: as a paired per-image difference it is +0.0150 (95 % CI [−0.0121, +0.0421], win rate 56.4 %) against a detectable effect of 0.0385 for that comparison — the same regime as the attention mechanisms, on the other side of the reference. In the single-normalization arm it is −0.0030, in the other direction, and neither arm resolves it. The faithful arm is primary because that rule was fixed in advance, not because it is the more favourable one — it is not.

The UNeXt publication uses the same dataset definition as this study and reports a Dice of 0.7937 against the 0.7466 we measure for the same architecture in our sensitivity arm. This comparison is secondary and was constructed after the results were seen. Two of the three components of the difference were measured directly: metric aggregation accounts for less than 10 % of it, and the single largest identified component is the choice of split on which the checkpoint is selected: +0.0316. In the faithful arm the same quantity is +0.0355 for UNeXt and +0.0320 for CMUNeXt.

Reading the comparison. This is the central qualitative comparison of the paper, and it must be read with its base stated. The largest architectural effect we could not resolve is +0.0137, a paired per-image difference against a corrected threshold of 0.0262 computed from the same paired differences. The protocol effect is +0.0316, and it is a different kind of quantity: the difference between the best epoch scored on the reported partition and the value that partition gives when it is held out — a difference of two partition means, not a paired per-image difference. It therefore has no paired confidence interval, no win rate and no comparison-specific detectable effect, and we do not place it on the same ruler as 0.0262. What the two magnitudes support is an ordering of kind rather than of scale: the architectural effects could not be resolved by the design that measured them, while the protocol effect is large enough to be visible in every one of the three measurements we have of it (+0.0316, +0.0355, +0.0320).

The quantity cannot be given an interval from our own eleven configurations either, and the reason is structural rather than an omission: measuring it requires scoring the test partition at every epoch, which the protocol deliberately forbids. A related quantity — the best epoch on a disjoint validation partition minus the held-out test value — is computable from our runs and gives +0.0178 over eight configurations, but it captures only the generalisation gap and not the selection optimism, so it is a smaller and different measurement and we do not report it as this one. It is not a claim about how often the practice occurs: among the three baseline repositories whose code we could inspect, two select the checkpoint on the evaluation split and one does not, and $n = 3$ cannot support a prevalence claim. The correct reading of +0.0316 is the effect when

it occurs.

Protocol reporting in the audited studies

We audited 22 studies under the rules given in *Methods*. Reported Dice values on this dataset span 0.528 to 0.93, a range of 40 points that architecture differences alone do not account for. Table 7 gives how many of the 22 state each protocol dimension, and Figure 8 gives the record study by study.

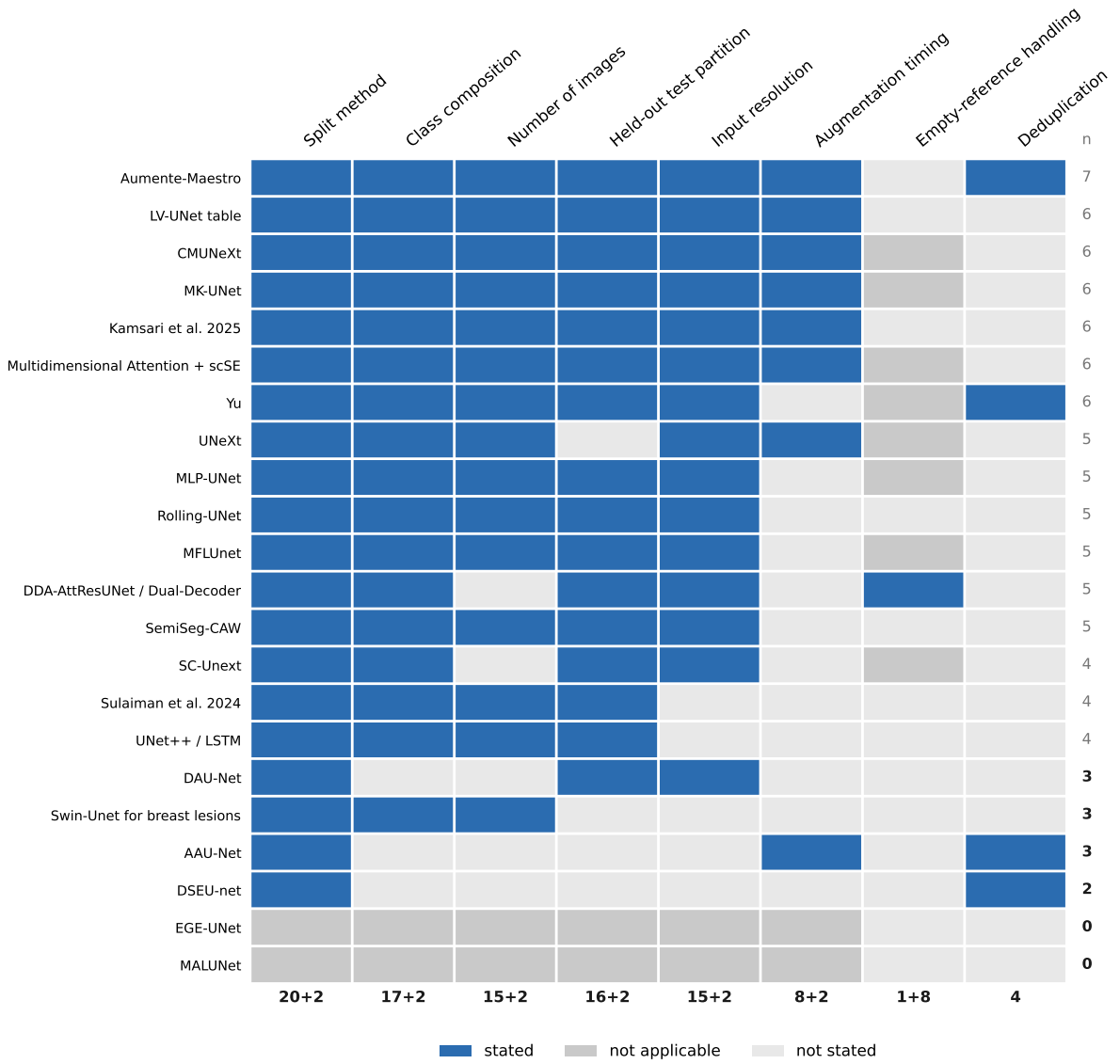


Figure 8: Each cell is one study's record on one protocol dimension. Counts below each column are studies stating it (plus, where shown, studies for which it does not apply); the column at the right is dimensions stated per study.

What is least often stated. The dimensions that determine whether two reported values are comparable are the ones least often stated. Two of the 22 report training and validation splits with no third partition and present validation-set numbers; one reports removing duplicated records — the study that identified them — and one acknowledges the risk without addressing it.

Table 7: Protocol dimensions and how many of the 22 audited studies state them. An absence is recorded as not stated, which is an observation about the record.

Protocol dimension	Studies stating it
Split method	20 of 22
Class composition (normal included or excluded)	17 of 22
Held-out test partition separate from validation	16 of 22
Number of images used	15 of 22
Input resolution	15 of 22
Empty-reference handling	9 of 22
Augmentation timing relative to the split	8 of 22
Deduplication	4 of 22

Reporting completeness is not predicted by venue: the most complete protocol reporting in the sample comes from a six-page conference paper, while internal inconsistencies appear in papers from top-tier conferences and established journals.

Audit of our own pipeline

The code-level audit returned a clean result for all eight configuration notebooks: the split precedes augmentation, the pre-augmentation function receives and iterates over the training partition only, and training consumes its in-memory return value. No configuration reads augmented images back from disk.

An inspection of the artifacts on disk returns the opposite conclusion: the pre-augmentation output directories contain augmented copies of test and validation images — 58 test and 55 validation for the first split, 60 and 55 for the second; the third is clean. An audit performed by inspecting these directories would report test-set contamination in two of three splits (Figure 9).

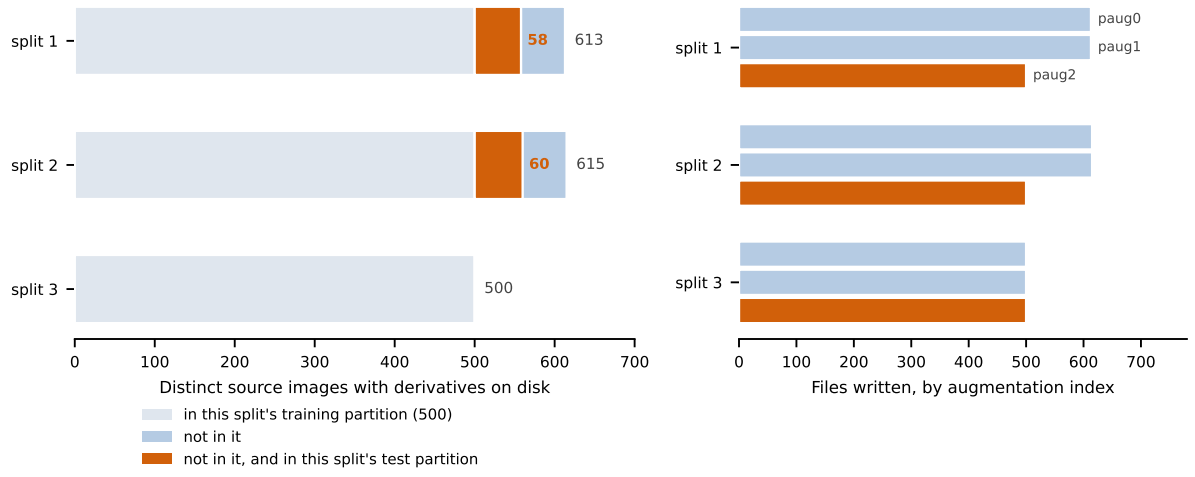


Figure 9: Left: source images with pre-augmented derivatives in each split's cache directory, against that split's training partition of 500. Right: the same files by augmentation index. In splits 1 and 2 two indices cover more images than the third, which is the signature of a second write pass under a different partition; split 3 shows one pass. The training list is built from the partition rather than from the directory, so the additional files are on disk but are not read (see text).

Cause. The directories are a stale-accumulation artifact. The pre-augmentation routine writes

to a per-seed directory with `exist_ok=True` and never clears it, and the split ratios were varied over development, so each run wrote its own train-only output without removing its predecessor's. An image that belonged to the training partition under an earlier ratio, and that falls into test or validation under the one actually used, remains on disk as an augmented file the final training run never read; the per-pass file counts are in the supplementary material (S6).

Effect on the partitions. Those actually used are reproducible. Regenerating the splits from the three recorded seeds with the released code returns test partitions identical to those in the per-image tables in all three splits, and 90 % of the unique originals in the first split's directory appear in the union of the training partitions of six historical split variants. The third split, which received fewer development runs, is clean — the pattern follows development history rather than partition identity.

The two audits therefore disagree on the same pipeline, and the artifact-based one is the one that is wrong. We report this because the failure mode is not specific to our code: any pipeline that caches augmented images to a non-versioned directory and is run more than once with different split parameters will accumulate the same appearance of leakage while remaining correct.

Discussion

Two results carry the discussion: what the cost axis shows about the five mechanisms, and how one protocol choice compares in size with the architectural effects. The section closes with the limitations of the design.

What the cost axis shows

Across five attention mechanisms evaluated under one protocol on the same images, parameter cost spanned three orders of magnitude while every measured effect fell below the threshold its own comparison could resolve. The two axes are not aligned: efficient channel attention with deep supervision added 12 parameters to the deployed model and produced the largest measured gain (+0.0137 Dice), while the attention gate added 32,595 — a factor of 2,716 — and produced less than half of it. A design decision on this axis can cost three orders of magnitude more and return nothing the data can distinguish.

This is not a statement that the mechanisms do nothing. An effect below a detection threshold is not an effect of zero, and *"could not be distinguished from zero at this sample size"* is a different claim from *"equivalent"*. What the study establishes is the resolution, and what that resolution implies for the comparisons this literature routinely reports. Among the open-access BUSI segmentation studies whose full text we could retrieve, roughly half of the Dice improvement margins stated in running text fall below the threshold this comparison could resolve, and roughly a quarter fall below the effect we report. Among the margins we could retrieve, the improvements reported on this benchmark lie below the resolution of the data they are reported on — and our own effect sits inside that group, not outside it. We state this as a property of the records

we retrieved rather than of the literature, because most studies report their margin only in a table and those margins are not in our sample. The same point has been made for the MICCAI corpus using confidence intervals [10] and for brain MRI using test-set-size analysis [11]; it is reproduced here with the parameter cost placed beside it.

Two results on the ablation axis bear on the same decision: adding a network level produced a resolved effect roughly three times the largest attention-related one, and a third augmentation pass produced none. On this dataset, at this scale, the decisions that moved the result were depth and the training-data regime, not the attention module.

We note one further asymmetry without building a ranking on it: augmentation was roughly three times more effective on the ResNet-50 encoder than on the four-level U-Net, but the two comparisons start from different baselines on a curve with diminishing returns.

One protocol choice is visible where the architectural effects are not

The clearest quantitative result of this study is a comparison between two measured magnitudes on the same data. Selecting the reported checkpoint on the evaluation partition rather than a held-out partition moved the reported Dice by +0.0316 in the sensitivity arm and +0.0355 and +0.0320 in the primary arm — three measurements of the same quantity on two architectures. Two architectures are not enough to call this a property of the protocol rather than of the architecture, and we do not.

The quantity cannot be recovered for the eleven ablation configurations, because under our design the selection partition is disjoint from the reported one, so the test partition was never scored during training: the measurement is available only where the practice being measured is present. That the practice inflates results is established [14]; what was not established is by how much on this task. This is the effect when it occurs, not an estimate of how often it occurs.

The consequence for reading this literature is narrow but firm. When a table places values from different studies side by side, a difference of the size typically separating adjacent rows can be produced by a protocol choice the studies are not required to report — and, in the sample we examined, often do not. The benchmark remains usable for comparisons under one protocol, which is what an ablation is; it does not support ranking across studies.

Removing the affected records moves the absolute Dice of all eleven configurations down by 0.0203 on average, and moves the paired differences by 0.0065, with no comparison changing sign and the largest single shift at 0.0120 — below the smallest threshold in this study. This is the distinction that determines how the benchmark can still be used: for absolute reporting it is inflated, for within-protocol ablation it is serviceable.

We report the retraction described in *Results* — the score premium we first read as an architecture-dependent memorization signature, and which stratification showed to be an artifact of lesion size — because the first reading would have supported a more striking claim than the data sustain.

The two available audit strategies disagreed on our own pipeline: inspecting the artifacts on disk indicates test-set contamination in two of three partitions, while the source code shows the

training pipeline clean in all eight configurations. That reproduction does not establish validity is established [19], and that artifact-level checks can diverge from what execution did is also established — a forensic audit of a radiology benchmark reports that *"agreement across these artifacts is usually assumed rather than tested"* [18].

Direction of the discrepancy. Our case runs the other way. There, artifact-level inspection concealed a defect; here it manufactures one on a pipeline the code shows to be correct, and to the best of our knowledge we are not aware of a prior report of an artifact-level audit producing a false-positive leakage finding. The asymmetry is practical rather than formal: a false negative misses a problem, while a false positive places a correct study in the position of arguing against its own released files.

SAM ViT-B, prompted with a box derived from the ground-truth mask, reaches 0.8324 on the same partitions. This is not a competing method and is not comparable with the automatic models: the prompt supplies the lesion location. Read as an oracle bound, it indicates that a substantial part of the distance between our models and a perfect segmentation is not attributable to localization failure, and that a general-purpose model given the location outperforms task-specific models that must find it. Because a general rather than a medically fine-tuned variant was used, the bound is likely an underestimate.

The comparisons this paper reports are intervals on differences, not on absolute metrics. Reporting *"Dice 0.75 (95 % CI 0.70–0.80)"* is the common pattern in this literature; reporting *"the difference is +0.0137 (95 % CI −0.0015, +0.0288)"* is not, and it is the second that determines whether a reported improvement can be distinguished from zero. Tables 3 and 4 give the second for all nine comparisons.

Precedent. This is not unprecedented here, and we cite where it appears. One study on breast ultrasound including BUSI reports image-level Dice and IoU with paired Wilcoxon tests, Benjamini–Hochberg correction across baselines, and bootstrap confidence intervals [39]; another reports a 95 % confidence interval for the mean paired Dice difference across its test patients [40]; a third compares a proposed method against three references with Holm correction over six contrasts [41]. We name these rather than assert that the practice is rare, which would require a prevalence estimate our sampling does not support.

None of the principles this paper relies on is new: selection on the evaluation partition inflates results, duplicate records break independence, and effects can be smaller than the errors against which they are measured. Each is published and cited here; we add measurements rather than claims of priority — and we regard that as a strength, because a claim of priority can be refuted by a single earlier paper while a measurement cannot.

What the measurements together suggest is narrower than a recommendation and firmer than an impression: on this benchmark, at this scale, the parameter cost of an architectural decision can be stated precisely while the benefit cannot, and the quantities that move a reported number most are not the ones the reports are about.

The parameter cost of an attention module is one input to a deployment decision; whether it transfers to quantized or embedded settings is not established by these measurements.

Limitations

The baseline comparison is not at matched capacity, and both that difference and the selection asymmetry run against our own configuration rather than for it. The training loop is ours rather than the reference repositories', which was necessary because those loops embed the evaluation protocols under study — a property of the intervention, not a confound in it — but it does mean the baselines were evaluated under a regime their authors did not choose.

Three further constraints bound what these measurements support. The protocol effect was measured on two architectures, which is not enough to establish it as a property of the protocol rather than of the architecture. Models were trained on the dataset as distributed, without deduplication; the effect of the repeated records was measured on the existing per-image outputs and is smaller than the smallest threshold here on paired differences, but a retrained comparison on a curated release [17] would be a stronger test and we did not run it. And the clustering correction rests on an approximation: with no patient identifiers in the release, image similarity is the only available proxy, and it errs in both directions.

Two properties of the design bound what the Family A comparisons can attribute. Deep supervision is never varied on its own: four of the five mechanisms carry it and neither the attention gate nor the reference does, so a deeply supervised variant's effect cannot be apportioned between mechanism and supervision. Family A's internal ordering is unaffected — all four carry the same arm — but the contrast with the attention gate is. This bears on the result we state most prominently. The largest gain, from efficient channel attention with deep supervision, is reported beside a deployed-form cost of 12 parameters, but that figure is the inference footprint of the attention module alone: the effect is produced by the deeply supervised training configuration, whose supervision heads add parameters that are removed at inference and whose checkpoint criterion, as the Methods notes, is the one more favourable to the metric we report. The twelve parameters state what the mechanism costs to deploy, not that twelve parameters produced the gain. And the training partitions of the three splits overlap: two independent draws of 500 images from 647 share 386 on average, so the splits are not independent replicates. The correction we apply treats dependence among test records only; dependence transmitted through shared training data is not modelled, and it works in the direction of making the resolved effects less certain rather than more.

The study was designed to detect effects of a certain size and did not; it was not designed to establish equivalence, and a null result at this sample size does not support one. The mechanisms whose magnitudes we report are dataset-independent and we measured them on one dataset. Excluding the normal class removes the empty-reference term from every number we report, but means we cannot report how these models behave on images containing no lesion. Architectures outside the lightweight scope, including nnU-Net and the studies reporting values at the top of the observed range, were not retrained; what we report is the heterogeneity of the protocols behind those values, not a re-evaluation of the values themselves.

Finally, this paper argues that stored artifacts are not by themselves evidence about the process that produced them, and that argument applies to our own package. Two provenance gaps

remain — the TensorFlow version of the ablation runs, which the released weight files do not record although they do record the Keras version, and per-epoch histories for three of the eleven configurations — and a third, the evaluation pre-processing path, has since been closed; the supplementary material (S6) gives what re-execution establishes and what it does not. Re-executing a package is evidence that a number can be recovered, not that the process recorded it faithfully. The sequence and the four undocumented choices it turned on are in the supplementary material (S6).

Conclusions

Under a single evaluation protocol on breast ultrasound lesion segmentation, the parameter cost of five attention mechanisms spanned three orders of magnitude while none of their measured effects reached the threshold its own comparison could resolve. The largest gain came from the cheapest mechanism, and the two axes were not aligned. Along the same ablation axis, network depth produced a resolved effect three times larger, and a third augmentation pass produced none. A single protocol choice — selecting the reported checkpoint on the evaluation partition — moved the reported value in each of the three measurements we have of it, where no attention mechanism produced an effect the design could resolve. The two magnitudes are an ordering of kind rather than of scale, and we do not place them on one ruler.

We report these magnitudes rather than a recommendation. The cost side of an architectural decision on this benchmark can be stated exactly; the benefit side, at this sample size, cannot. This benchmark remains usable for ablation under one protocol and does not support ranking across studies.

List of abbreviations

AG	Attention gate
ASSD	Average symmetric surface distance
BUSI	Breast Ultrasound Images (dataset)
CBAM	Convolutional block attention module
CI	Confidence interval
DS	Deep supervision
ECA	Efficient channel attention
HD95	95th percentile Hausdorff distance
IoU	Intersection over union
MDE	Minimum detectable effect
SAM	Segment Anything Model

scSE	Spatial and channel squeeze-and-excitation
SD	Standard deviation
SE	Squeeze-and-excitation

Declarations

Ethics approval and consent to participate

This study used a publicly available, de-identified dataset and involved no new data collection from human participants; ethical approval and informed consent were obtained by the authors of the original dataset publication [6]. Images are reproduced from the public release as distributed; the only text burned into them is anatomical laterality and device annotation.

Consent for publication

Not applicable.

Availability of data and materials

The Breast Ultrasound Images (BUSI) dataset analysed in this study is publicly available from the dataset publication [6] under a CC BY 4.0 licence, at the address given there: `https://scholar.cu.edu.eg/?q=afahmy/pages/dataset`. The code and the per-image measurement tables required to reproduce every number reported here, including the mapping from the configuration names used in this paper 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`.

Competing interests

The author declares that he has no competing interests.

Funding

No funding was received for this work.

Authors' contributions

G.U. is the sole author. He designed the study, implemented the experiments, performed the analysis, and wrote the manuscript. The author read and approved the final manuscript.

Acknowledgements

During the preparation of this manuscript, the author used Anthropic Claude as an assistive tool for language, presentation, and reference management, and to help review and revise drafts of the text. All study design, data analysis, statistical computation, and conclusions are the author's own; every reported value and margin was computed exclusively by the author's own scripts from the raw data and verified manually. AI was not used to generate data, results, or conclusions, and no AI system is listed as an author, in line with ICMJE recommendations.

References

[1] Ervik M, Laversanne M, Lam F, Colombet M, Ferlay J, Filho AM, et al.. Global Cancer Observatory: Cancer Today (2024 estimates, version 1.0); 2026. Lyon, France: International Agency for Research on Cancer. Available from: `https://gco.iarc.who.int/today`.

[2] Sung H, Filho AM, Laversanne M, Ferlay J, Siegel RL, Soerjomataram I, et al. Global cancer statistics 2024: GLOBOCAN estimates of incidence and mortality worldwide for 34 cancers in 186 countries. CA: A Cancer Journal for Clinicians. 2026;76.

[3] Berg WA, Blume JD, Cormack JB, Mendelson EB, Lehrer D, Böhm-Vélez M, et al. Combined Screening With Ultrasound and Mammography vs Mammography Alone in Women at Elevated Risk of Breast Cancer. JAMA. 2008 May;299(18):2151-63.

[4] Park CS, Lee JH, Yim HW, Kang BJ, Kim HS, Jung JI, et al. Observer Agreement Using the ACR Breast Imaging Reporting and Data System (BI-RADS)-Ultrasound, First Edition (2003). Korean Journal of Radiology. 2007;8(5):397-402.

[5] Webb JM, Adusei SA, Wang Y, Samreen N, Adler K, Meixner DD, et al. Comparing deep learning-based automatic segmentation of breast masses to expert interobserver variability in ultrasound imaging. Computers in Biology and Medicine. 2021;139:104966.

[6] Al-Dhabyani W, Gomaa M, Khaled H, Fahmy A. Dataset of breast ultrasound images. Data in Brief. 2020;28:104863.

[7] Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods. 2021;18(2):203-11.

[8] Isensee F, Wald T, Ulrich C, Baumgartner M, Roy S, Maier-Hein K, et al. nnU-Net Revisited: A Call for Rigorous Validation in 3D Medical Image Segmentation. In: Linguraru MG, Dou Q, Feragen A, Giannarou S, Glocker B, Lekadir K, et al., editors. Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. Lecture Notes in Computer Science. Cham: Springer Nature Switzerland; 2024. p. 488-98.

[9] Varoquaux G, Cheplygina V. Machine learning for medical imaging: methodological failures and recommendations for the future. npj Digital Medicine. 2022;5(1):48.

[10] Christodoulou E, Reinke A, Houhou R, Kalinowski P, Erkan S, Sudre CH, et al. Confidence Intervals Uncovered: Are We Ready for Real-World Medical Imaging AI? In: Medical Image Computing and Computer Assisted Intervention (MICCAI 2024). vol. 15010 of Lecture Notes in Computer Science. Springer, Cham; 2024. p. 124-32.

[11] El Jurdi R, Varoquaux G, Colliot O. Confidence Intervals for Performance Estimates in Brain MRI Segmentation. arXiv preprint arXiv:230710926. 2023.

[12] Maier-Hein L, Reinke A, Godau P, Tizabi MD, Buettner F, Christodoulou E, et al. Metrics Reloaded: recommendations for image analysis validation. Nature Methods. 2024;21(2):195-212.

[13] Reinke A, Tizabi MD, Sudre CH, Eisenmann M, Rädsch T, Baumgartner M, et al.. Common Limitations of Image Processing Metrics: A Picture Story; 2021.

[14] Kapoor S, Narayanan A. Leakage and the reproducibility crisis in machine-learning-based science. Patterns. 2023;4(9):100804.

[15] Chen G, Liu Y, Qian J, Zhang J, Yin X, Cui L, et al. DSEU-net: A novel deep supervision SEU-net for medical ultrasound image segmentation. Expert Systems with Applications. 2023;223:119939.

[16] Chen G, Li L, Dai Y, Zhang J, Yap MH. AAU-Net: An Adaptive Attention U-Net for Breast Lesions Segmentation in Ultrasound Images. IEEE Transactions on Medical Imaging. 2023;42(5):1289-300.

[17] Aumente-Maestro C, Díez J, Remeseiro B. A multi-task framework for breast cancer segmentation and classification in ultrasound imaging. Computer Methods and Programs in Biomedicine. 2025;260:108540.

[18] Kozłowski M. Forensic Reproducibility Audit of a Radiology Vision-Language Model Benchmark: From Intended Protocol to Released Artifact. arXiv:2607.25589 [cs.CL]; 2026.

[19] National Academies of Sciences, Engineering, and Medicine. Reproducibility and Replicability in Science. Washington, DC: The National Academies Press; 2019.

[20] He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Los Alamitos, CA, USA: IEEE Computer Society; 2016. p. 770-8.

[21] Wang Q, Wu B, Zhu P, Li P, Zuo W, Hu Q. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA; 2020. p. 11531-9.

[22] Hu J, Shen L, Sun G. Squeeze-and-Excitation Networks. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA; 2018. p. 7132-41.

[23] Woo S, Park J, Lee JY, Kweon IS. CBAM: Convolutional Block Attention Module. In: European Conference on Computer Vision (ECCV); 2018. .

[24] Roy AG, Navab N, Wachinger C. Concurrent Spatial and Channel Squeeze and Excitation in Fully Convolutional Networks. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI); 2018. .

[25] Oktay O, Schlemper J, Folgoc LL, Lee M, Heinrich M, Misawa K, et al. Attention U-Net: Learning Where to Look for the Pancreas. In: Medical Imaging with Deep Learning (MIDL); 2018. .

[26] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI); 2015. .

[27] Lee CY, Xie S, Gallagher P, Zhang Z, Tu Z. Deeply-Supervised Nets. In: Artificial Intelligence and Statistics (AISTATS); 2015. .

[28] Li J, Liu K, Hu Y, Zhang H, Heidari AA, Chen H, et al. Eres-UNet++: Liver CT image segmentation based on high-efficiency channel attention and Res-UNet++. Computers in Biology and Medicine. 2023;158:106501.

[29] Valanarasu JMJ, Patel VM. UNeXt: MLP-Based Rapid Medical Image Segmentation Network. In: Wang L, Dou Q, Fletcher PT, Speidel S, Li S, editors. Medical Image Computing and Computer Assisted Intervention – MICCAI 2022. vol. 13435 of Lecture Notes in Computer Science. Cham: Springer; 2022. p. 23-33.

[30] Tang F, Ding J, Quan Q, Wang L, Ning C, Zhou SK. CMUNeXt: An Efficient Medical Image Segmentation Network Based on Large Kernel and Skip Fusion. In: IEEE International Symposium on Biomedical Imaging (ISBI); 2024. .

[31] Kirillov A, Mintun E, Ravi N, Mao H, Rolland C, Gustafson L, et al. Segment Anything. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France; 2023. p. 3992-4003.

[32] Taha AA, Hanbury A. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. BMC Medical Imaging. 2015;15(1):29.

[33] Maier O, et al.. MedPy: Medical image processing in Python;. Available from: `https://github.com/loli/medpy`.

[34] Reinke A, Tizabi MD, Baumgartner M, Eisenmann M, Heckmann-Nötzel D, Kavur AE, et al. Understanding metric-related pitfalls in image analysis validation. Nature Methods. 2024;21(2):182-94.

[35] Wilcoxon F. Individual Comparisons by Ranking Methods. Biometrics Bulletin. 1945;1(6):80-3.

[36] Holm S. A Simple Sequentially Rejective Multiple Test Procedure. Scandinavian Journal of Statistics. 1979.

[37] Yu S, Huang M, Chen E, Zhu B, Liang X, Xie Y. Benchmarking YOLOs in breast ultrasound lesion segmentation. Quantitative Imaging in Medicine and Surgery. 2026;16(7).

[38] LeakageDetector: An Open Source Data Leakage Analysis Tool in Machine Learning Pipelines. arXiv preprint arXiv:250314723. 2025.

[39] Mallina R, Shareef B. XBusNet: Text-Guided Breast Ultrasound Segmentation via Multimodal Vision-Language Learning. Diagnostics. 2025;15(22):2849.

[40] Minoccheri C, Hodgman M, Ma H, Merchant R, Wittrup E, Williamson C, et al. LoRA-based methods on Unet for transfer learning in aneurysmal subarachnoid hematoma segmentation. BMC Medical Imaging. 2025;26(1):58.

[41] Kang J, Kim HG, Shin NY, Nam Y. Automatic lateral ventricle and choroid plexus segmentation method in infant brain MR images. BMC Medical Imaging. 2026;26(1):251.

Supplementary Files

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

- supplementary.pdf