A Lightweight Network Matches Larger Models in Cross-Dataset Polyp Segmentation under a Controlled Multi-Seed Protocol

Mostofa Adib Shakib[a], Lotifur Shabbir[b], Abraar A Rehman[c], Sajidur Rahman Siam[b], Nazifa Mouli[d], Sadia Hamid[e], Adnan Safin[b] and Tauhidur Rahman Sakib[b,*]

[a]Variant Labs, Newark, 19713, Delaware, USA
[b]Department of Computer Science and Engineering, BRAC University, Dhaka, 1212, Bangladesh
[c]Department of Computer Science, BRAC University, Dhaka, 1212, Bangladesh
[d]Department of Computer Science, Iowa State University, Ames, 50011, Iowa, USA
[e]Department of Information and Communication Technology, Bangladesh University of Professionals, Mirpur Cantonment, Dhaka, 1216, Bangladesh

ARTICLE INFO

Keywords:
polyp segmentation
domain generalization
discrete wavelet transform
feature perturbation
lightweight neural networks
colonoscopy
cross-dataset evaluation

ABSTRACT

Cross-dataset generalization and deployment cost are two largely independent obstacles to the clinical translation of polyp segmentation. We present WaveLite-DG, a compact encoder–decoder addressing both. It combines a MobileNetV3-Large encoder, Wavelet Frequency Modules that perturb approximation-band feature statistics during training while enhancing detail-band structure, a gated context scan at the bottleneck, and a boundary-supervised decoder. We derive in closed form the energy this perturbation injects. Trained on Kvasir-SEG and CVC-ClinicDB, with CVC-ColonDB, ETIS-LaribPolypDB and CVC-300 withheld from every fitting and selection decision, WaveLite-DG reaches a three-seed mean macro unseen Dice of 0.8300 ± 0.0004, the highest of six architectures retrained under one protocol without photometric augmentation and statistically equivalent to Polyp-PVT and VM-UNet within a pre-set margin of 0.02. With colour jitter, or with the perturbation added to those two baselines, they lead by under 0.008. In every regime tested WaveLite-DG stays within 0.008 of the best model with one third or less of the parameters (7.68 M) and one fifth or less of the computation (1.93 giga multiply–accumulate operations), and it is 0.0427 more accurate than the best of four lightweight networks. Approximation-band perturbation adds +0.0651 macro Dice without photometric augmentation but only +0.0185 with colour jitter, which recovers most of the gain, and perturbing the undecomposed feature at matched energy shows no detectable difference. The contribution therefore lies in efficiency at parity rather than in a new mechanism.

1. Introduction

Colorectal cancer is among the most frequently diagnosed malignancies worldwide [1], and colonoscopic removal of precursor adenomas substantially reduces long-term mortality [2]. Tandem-colonoscopy studies nevertheless report clinically important adenoma miss rates, concentrated among small, flat and low-contrast lesions [3]. Computer-aided segmentation can assist endoscopists by delineating lesion boundaries and directing visual attention.

Encoder–decoder networks have driven rapid progress on this task, from U-Net [4] and PraNet [5] to the transformer-based Polyp-PVT [6] and SSFormer [7] and the state-space VM-UNet [8]. Their clinical value, however, depends on two further properties: whether accuracy persists when hardware, illumination, bowel preparation and annotation convention differ from the training data, and whether the model fits on endoscopy processors and shared workstations that operate well below the resources of a research GPU.

The standard protocol introduced with PraNet [5] — training on Kvasir-SEG and CVC-ClinicDB and testing on CVC-ColonDB, ETIS-LaribPolypDB and CVC-300 — is already cross-dataset, and we do not claim otherwise. Our protocol refines it in two respects: checkpoints are selected solely on a held-out split of the source pool, so no target-set quantity enters any selection decision; and the headline score is the unweighted macro-average over target datasets, so that each acquisition domain counts once. We also measure accuracy and cost for all six models under one protocol rather than collating figures from heterogeneous sources.

We propose WaveLite-DG. Its Wavelet Frequency Module (WFM) applies a one-level Haar decomposition to encoder skip features; during training the approximation band receives a stochastic affine perturbation of its statistics, while the detail bands pass through a lightweight structural enhancement path. A gated context scan supplies bottleneck context, and a training-only boundary head adds contour supervision. Our contributions are:

1. A lightweight architecture that reaches parity with much larger models on unseen data: without photometric augmentation, the highest three-seed mean macro unseen Dice (0.8300 ± 0.0004) among six models, statistically equivalent to Polyp-PVT and VM-UNet, and within 0.008 of the best model in every regime tested, with the lowest parameter count, computation and memory of the six (7.68 M, 1.93 GMACs, 79 MB)

* Corresponding author at: Department of Computer Science and Engineering, BRAC University, Dhaka 1212, Bangladesh.
E-mail address: tauhidurrahmansakib@gmail.com (T. R. Sakib).

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7615100

against 24.44–44.27 M and 10.00–20.90 GMACs.
2. A frequency-selective feature perturbation with zero parameters and zero inference cost, whose expected injected energy we derive in closed form (Section 3.3), making exact how it differs from whole-feature perturbation.
3. A parameter-matched, three-seed ablation isolating the perturbation from capacity, with controls for colour jitter, for the bottleneck module and for whole-feature perturbation, and comparisons against MixStyle, DSU and FDA in the same architecture and against four lightweight segmentation networks.

The principal claim is efficiency at parity: the efficiency advantage is structural, whereas the accuracy margins over the strongest baselines are within run-to-run variance. We do not claim that the wavelet decomposition is *necessary*. The controls reported in Section 5.6 show that perturbing the undecomposed feature at matched energy makes no detectable difference, that ordinary colour jitter recovers most of the perturbation's benefit, and that established domain-generalization mechanisms perform comparably in the same architecture.

2. Related work

2.1. Polyp segmentation

U-Net [4] established the skip-connected encoder–decoder, and UNet++ [9] narrowed its semantic gap with nested pathways. PraNet [5] added reverse-attention refinement on a Res2Net-50 backbone [10]. Polyp-PVT [6] and SSFormer [7] build on PVTv2 [11], and selective state-space models [12, 13] underlie VM-UNet [8]. Recent work continues to refine attention and fusion design [14, 15, 16]. A parallel line targets deployment cost: Mobile-PolypNet [17] and later efficient designs [18, 19, 20, 21] show that compact models can be accurate. Our contribution is not that claim itself, but evidence that a compact model holds parity with three families of much larger architectures when all are trained under one protocol.

2.2. Feature-statistic domain generalization

Adaptive instance normalization showed that channel-wise feature statistics encode visual style [22]. MixStyle [23] mixes these statistics across samples, and DSU [24] samples them from an estimated distribution. Our perturbation is generative in the manner of DSU, needing no partner sample, but is restricted to the wavelet approximation band. Reported margins in domain generalization are often comparable to seed variation [25], which shapes how we report ours.

2.3. Frequency and wavelet methods

FDA [26] swaps low-frequency Fourier amplitude between images, an idea extended to domain generalization [27] and federated medical segmentation [28]. The discrete wavelet transform adds spatial localization [29]; MWCNN [30] embedded it in CNNs, and Wave-SAN [31] swaps low-frequency wavelet components for cross-domain few-shot learning. WaveLite-DG differs in operating on learned multi-scale features rather than images, in requiring no second image, and in routing the detail bands around the perturbation through a separate enhancement path — an option unavailable to whole-feature methods such as MixStyle and DSU. These are architectural distinctions with an analytical consequence (Section 3.3.4), not an established empirical advantage.

3. Method

3.1. Architecture overview

For a 352×352 input, an ImageNet-pretrained MobileNetV3-Large encoder [32], fully fine-tuned, yields features at strides 4, 8, 16 and 32 with 24, 40, 112 and 960 channels. The three higher-resolution features each pass through a WFM before entering the decoder; the stride-32 feature is refined by the gated context scan (Fig. 1).

3.2. Wavelet Frequency Module

Let $x \in \mathbb{R}^{B\times C\times H\times W}$ be a skip feature. A one-level two-dimensional Haar transform gives

$$\mathrm{DWT}(x) = \{\mathrm{LL},\ \mathrm{LH},\ \mathrm{HL},\ \mathrm{HH}\}, \tag{1}$$

with each band of size $(H/2) \times (W/2)$. The filters are fixed and differentiable, and the inverse transform is exact.

3.2.1. High-frequency path

The three detail bands are concatenated ($3C$ channels) and processed by a grouped convolution with one group per directional band:

$$\begin{aligned}&\{\mathrm{LH}', \mathrm{HL}', \mathrm{HH}'\}\\ &= \mathrm{split}\big(\mathrm{ReLU}(\mathrm{BN}(\mathrm{Conv}_{g=3}([\mathrm{LH};\mathrm{HL};\mathrm{HH}])))\big).\end{aligned} \tag{2}$$

3.2.2. Approximation-band perturbation

Let $\mu, \sigma \in \mathbb{R}^{B\times C}$ be the per-sample, per-channel spatial mean and population standard deviation (divisor N) of LL. During training,

$$\mathrm{LL}' = \gamma\,(\mathrm{LL} - \mu) + \mu + \beta\,\sigma, \tag{3}$$

$$\gamma = 1 + \lambda\varepsilon_\gamma, \qquad \beta = \lambda\varepsilon_\beta, \tag{4}$$

with $\varepsilon_\gamma, \varepsilon_\beta \sim \mathcal{N}(0,1)$ drawn once per sample, shared across channels and independent across scales. We use $\lambda = 0.3$ and application probability $p = 0.5$; γ is not clamped ($P(\gamma < 0) \approx 4 \times 10^{-4}$). Eq. (3) is the reduced form of the AdaIN-style normalize-then-rescale operation [22], in which σ cancels. It alters statistics associated predominantly with illumination, contrast and colour, and is disabled at inference, adding no parameters or cost.

3.2.3. Reconstruction

The sub-bands are recombined and fused residually (Fig. 2):

$$\begin{aligned}\mathrm{WFM}(x) = x\\ + \mathrm{Conv}_{1\times1}\big(\mathrm{IDWT}(\mathrm{LL}', \mathrm{LH}', \mathrm{HL}', \mathrm{HH}')\big).\end{aligned} \tag{5}$$

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7615100

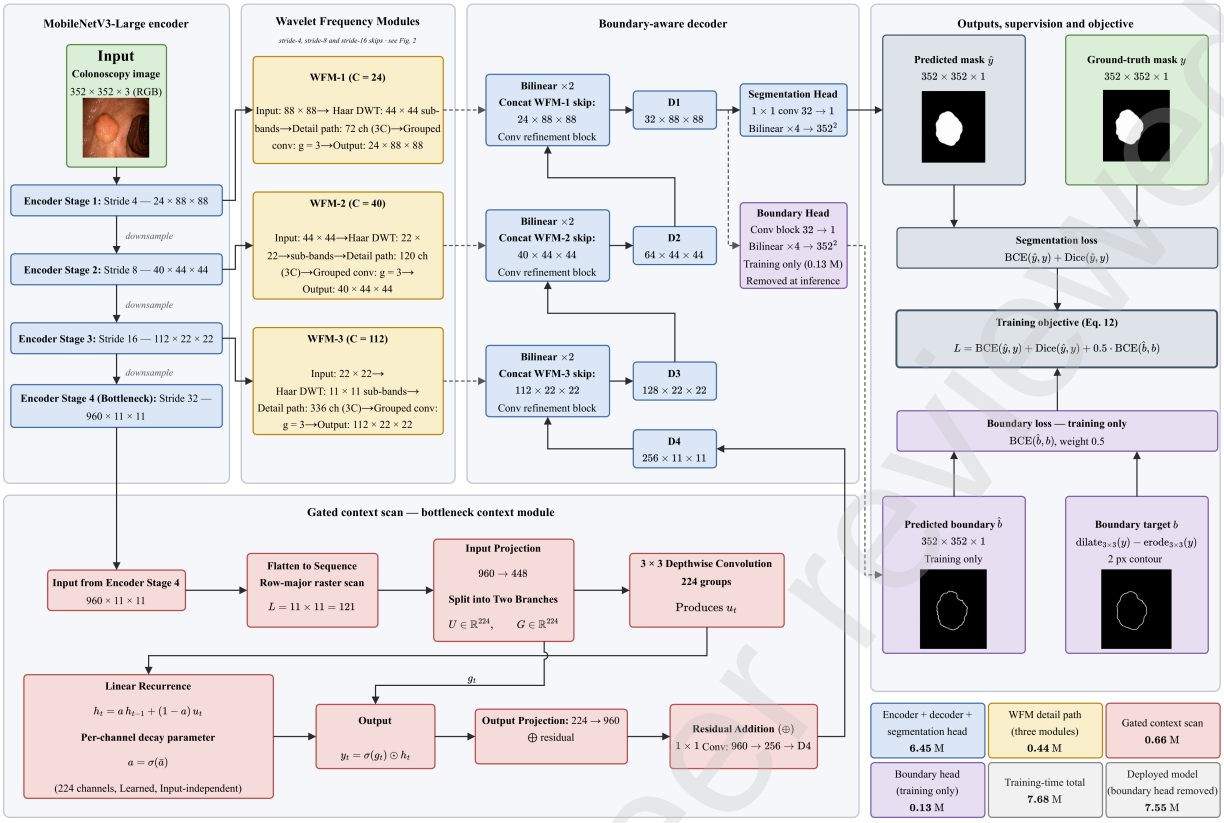


Fig. 1. Overall architecture of WaveLite-DG. Wavelet Frequency Modules process the stride-4, 8 and 16 skip connections; the gated context scan refines the stride-32 bottleneck. Four decoder stages (256, 128, 64 and 32 channels) fuse the skip pathways. The boundary head (dashed) is used during training only.

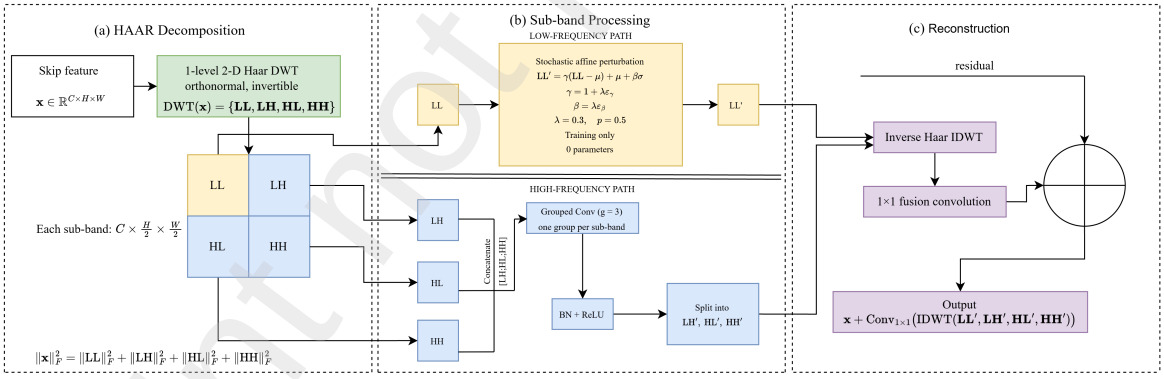


Fig. 2. Wavelet Frequency Module. **(a)** Haar decomposition into LL, LH, HL and HH. **(b)** Training-only perturbation of LL (Eq. (3)) and grouped-convolutional enhancement of the detail bands (Eq. (2)). **(c)** Inverse transform, 1 × 1 fusion and residual addition (Eq. (5)).

Fig. 3 shows the module on a real feature map: the change in the approximation band is modest relative to that in the learned detail path, as the analysis of Section 3.3 predicts.

3.3. Signal-processing analysis

3.3.1. Exact control of the injected energy

The Haar analysis operator W, built from $h = [1, 1]/\sqrt{2}$ and $g = [1, -1]/\sqrt{2}$, is orthonormal, so Parseval's relation holds:

$$\|x\|_F^2 = \|\mathrm{LL}\|_F^2 + \|\mathrm{LH}\|_F^2 + \|\mathrm{HL}\|_F^2 + \|\mathrm{HH}\|_F^2 . \quad (6)$$

Because only LL is modified, $\|\Delta x\|_F = \|\mathrm{LL}' - \mathrm{LL}\|_F$ exactly; in a redundant decomposition the delivered perturbation would instead depend on the frame bounds. From Eq. (3) and (4), $\Delta\mathrm{LL}_b = \lambda\varepsilon_\gamma^{(b)}(\mathrm{LL}_b - \mu_b) + \lambda\varepsilon_\beta^{(b)}\sigma_b$. The cross term vanishes identically, since $\sum_n(\mathrm{LL}_{b,c,n} - \mu_{b,c}) = 0$, and $\sum_n(\mathrm{LL}_{b,c,n} - \mu_{b,c})^2 = N\sigma_{b,c}^2$, so

$$\mathbb{E}\left[\,\|\Delta x_b\|_F^2\,\right] = 2\lambda^2\,\|\mathrm{LL}_b - \mu_b\|_F^2 , \quad (7)$$

as $\varepsilon_\gamma^2 + \varepsilon_\beta^2 \sim \chi_2^2$ has mean 2. Thus $\lambda = 0.3$ injects, in expectation, 0.18 times the centred approximation-band

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7615100

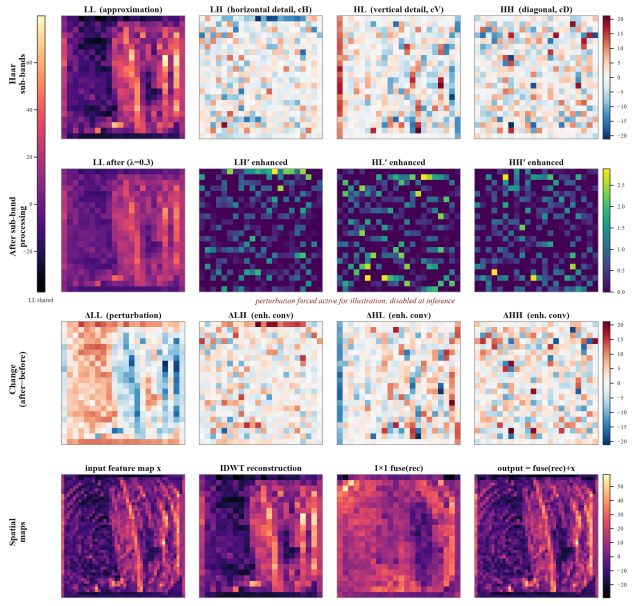


Fig. 3. WFM internals for a trained model (stride-8 skip, channel 0, a CVC-ColonDB image). Rows: Haar sub-bands; the same bands after processing ($\lambda = 0.3$, perturbation forced on); their differences; input, reconstruction, fusion output and residual sum.

energy (0.09 marginally, with $p = 0.5$). The injected energy scales with each feature's own variance, so it is calibrated across scales without per-scale tuning, and it is right-skewed: about 14% of applied perturbations exceed twice the expectation and 5% exceed three times it, exposing the network to occasional large appearance excursions.

3.3.2. A graded frequency split

The analysis filters have magnitude responses

$$|H(\omega)| = \sqrt{2}\left|\cos(\omega/2)\right|, \quad |G(\omega)| = \sqrt{2}\left|\sin(\omega/2)\right|, \quad (8)$$

crossing at $\omega = \pi/2$ with the widest transition band of any orthonormal wavelet. The approximation band attenuates rather than removes mid-frequency content, so the split isolates a smooth, locally averaged component, not an "appearance" component; anatomy and appearance are present in all four bands.

3.3.3. Why Haar, and one level

The skip features are 88×88, 44×44 and 22×22. Haar ($L = 2$) tiles these exactly without boundary extension, whereas an 8-tap Daubechies or Symlet filter would perturb about seven samples at each edge — most of a 22×22 map — and it is a fixed stride-2 depthwise convolution with $\pm 1/2$ kernels, carrying no learned parameters. A second level is not uniformly available: the stride-16 approximation band is 11×11, which a critically sampled Haar step cannot tile without explicit padding.

3.3.4. Relation to whole-feature perturbation

Applied to the undecomposed feature, the same derivation gives $2\lambda^2 \left\|x - \mu_x\right\|_F^2$. Since the Haar highpass annihilates constants, W maps the centred tensor to $\{\mathrm{LL} - \mu_{\mathrm{LL}}, \mathrm{LH}, \mathrm{HL}, \mathrm{HH}\}$, and Eq. (6) yields

$$\mathbb{E}\left[\left\|\Delta x_{\mathrm{full}}\right\|_F^2\right] - \mathbb{E}\left[\left\|\Delta x_{\mathrm{LL}}\right\|_F^2\right] = 2\lambda^2\left(\|\mathrm{LH}\|_F^2 + \|\mathrm{HL}\|_F^2 + \|\mathrm{HH}\|_F^2\right), \quad (9)$$

with equality of the two only when the detail bands vanish, that is, for a feature that is constant within every 2×2 Haar block. Whole-feature perturbation rescales edge and texture coefficients in the same proportion as the smooth component; ours leaves them untouched and routes them through Eq. (2). This establishes non-equivalence at matched λ, and by a computable amount. It does not establish that the restriction is *beneficial*; that requires the matched-strength control reported in Section 5.6.

3.4. Gated context scan

The bottleneck $z \in \mathbb{R}^{B\times 960\times 11\times 11}$ is projected to 448 channels and split into branches $U, G \in \mathbb{R}^{224}$; U passes through a 3×3 depthwise convolution. The sequence, in row-major raster order, is processed by

$$h_t = a\, h_{t-1} + (1 - a)\, u_t, \quad (10)$$

$$y_t = \sigma(g_t) \odot h_t, \qquad a = \sigma(\bar{a}), \quad (11)$$

where $\bar{a}$ is a learned per-channel vector, followed by a $224 \to 960$ residual projection. Because the decay a is input-independent, we call this a gated scan rather than a selective state-space operator [12, 13]. At sequence length 121 its linear cost confers no practical advantage over attention; it is used because it supplies whole-map context for only 0.66 M parameters. The unidirectional scan imposes a directional asymmetry that multi-directional scanning would remove at higher cost.

3.5. Decoder, objective and complexity

A 1×1 projection reduces the bottleneck to 256 channels; four stages of bilinear upsampling, skip concatenation and convolutional refinement follow, with widths 256, 128, 64 and 32. A segmentation head and a training-only boundary head map the final representation to one channel each. The loss is

$$\mathcal{L} = \mathcal{L}_{\mathrm{BCE}}(\hat{y}, y) + \mathcal{L}_{\mathrm{Dice}}(\hat{y}, y) + \lambda_b\, \mathcal{L}_{\mathrm{BCE}}(\hat{b}, b), \quad (12)$$

with $\lambda_b = 0.5$ and boundary targets from a morphological gradient,

$$b = \mathrm{dilate}_{3\times 3}(y) - \mathrm{erode}_{3\times 3}(y). \quad (13)$$

The model has 7.68 M training-time parameters (encoder, decoder and mask head 6.45 M; detail path 0.44 M; gated scan 0.66 M; boundary head 0.13 M) and 7.55 M when deployed. It requires 1.93 GMACs at 352×352. Comparison tables use the conservative training-time total.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7615100

Table 1
Datasets and partitions. Kvasir-SEG and CVC-ClinicDB are seen (source) domains, split 80/20 into training and validation; the validation split is used only for checkpoint selection. The other three are unseen and used only for testing.

Dataset	Images	Resolution	Train	Val.	Test
Kvasir-SEG [33]	1000	Variable	791	209	–
CVC-ClinicDB [34]	612	384 × 288	498	114	–
CVC-ColonDB [35]	379	574 × 500	–	–	379
ETIS-LaribPolypDB [36]	196	1225 × 966	–	–	196
CVC-300 [37]	60	574 × 500	–	–	60
Total	2247		1289	323	635

4. Experimental setup

4.1. Datasets and protocol

Kvasir-SEG [33] and CVC-ClinicDB [34] form the source domain; CVC-ColonDB [35], ETIS-LaribPolypDB [36] and CVC-300 [37] are used only for final testing (Table 1). The 1612 source images are split 80/20 at image level with a fixed seed (1289 training: 791 Kvasir, 498 ClinicDB; 323 validation), and the checkpoint with the highest validation Dice is retained. Seen-domain scores (†) are therefore selection results, not independent test estimates; all comparative conclusions rest on the 635 unseen images. Because ClinicDB frames come from a limited number of videos, correlated frames may straddle the split and inflate validation Dice. The CVC-ColonDB copy used (Kaggle distribution) contains 379 of the 380 standard test frames, and all ColonDB results are computed on these 379. Exact- and near-duplicate checks found no shared frames between the source pool and the target sets, nor between any two target sets (Supplementary Table S1). CVC-300 and CVC-ColonDB share resolution and acquisition source, so the macro average may weight that acquisition setting twice. Split manifests, SHA-256 hashes and the duplicate audit are available from the corresponding author on request.

4.2. Implementation and baselines

Table 2 lists the configuration, applied to all six models. Augmentation is limited to flips: no photometric augmentation is used, so that the approximation-band perturbation is the only appearance-space regularizer and its ablated contribution is not confounded. This will have depressed every model's unseen scores, so the comparison is one under a common regime, not of each model's best achievable accuracy. To measure that effect, the six models and A4 were also retrained with colour jitter, and the perturbation was added, without jitter, to the skip features of Polyp-PVT and VM-UNet (Section 5.6).

The baselines span three families (Table 3). U-Net is the *segmentation_models_pytorch* ResNet-34 variant [41], denoted U-Net (R34); the others use official implementations. All are retrained through our pipeline rather than transcribed. Each retains its published loss, so the comparison controls data, augmentation, optimizer, schedule, duration, checkpoint rule, resolution, test-time augmentation and threshold, but not the objective. As lightweight competitors we also trained, over three seeds and with the same configuration, U-Net with a MobileNetV3-Large encoder [4, 32], DeepLabV3+ with MobileNetV2 [42, 43], FPN with EfficientNet-B0 [44, 45] and UNeXt [46].

Table 2
Training configuration, shared by all six architectures.

Setting	Value
Input / interpolation	352 × 352; bilinear image, nearest mask
Augmentation	H/V flips ($p = 0.5$); no photometric
Optimizer	AdamW [38], lr 10^{-4}, wd 10^{-4}
Scheduler	Cosine [39], $T_{max} = 100$, $\eta_{min} = 10^{-6}$
Epochs / batch / precision	100 / 8 / mixed [40]
Checkpoint criterion	Highest validation Dice
Test-time augmentation	Three views, probability averaging
Threshold / seed	0.5 / 42 (repeats: 43, 44)

Table 3
Baselines. All use ImageNet initialization and the configuration of Table 2. Code: *smp.Unet*; *DengPingFan/PraNet*; *DengPingFan/Polyp-PVT*; *Qiming-Huang/ssformer*; *JCruan519/VM-UNet*. Bnd = boundary loss.

Model	Backbone	Loss
U-Net (R34) [41]	ResNet-34	BCE + Dice
PraNet	Res2Net-50 [10]	Structure loss
Polyp-PVT	PVTv2-B2 [11]	Structure loss
SSFormer-S	PVTv2-B2 [11]	BCE + Dice
VM-UNet	VSSM [13]	BCE + Dice
WaveLite-DG	MobileNetV3-L [32]	BCE + Dice + 0.5 Bnd

4.3. Profiling, metrics and statistics

Profiling uses batch size 1, 352 × 352 and FP32 on an NVIDIA RTX 3050 Laptop GPU (4 GB; CUDA 12.1, PyTorch 2.5.1), with 10 warm-up and 40 synchronized timed iterations. Throughput is a single forward pass. Peak memory is *torch.cuda.max_memory_allocated()* after warm-up (weights and activations). Parameters and MACs are measured with *thop* for all six models; we report GMACs, since many papers label MACs as "FLOPs".

Dice and IoU are primary; sensitivity, specificity and precision are complementary. The headline score is the unweighted mean of the three unseen dataset means; the pooled per-image view is also reported. For each baseline, per-image Dice differences on the pooled unseen set ($n = 635$) are summarized by a 95% percentile bootstrap (10 000 resamples) [47], a two-sided Wilcoxon signed-rank test [48] with Holm correction [49], and the standardized effect size d_z (mean/SD), since at $n = 635$ significance is reachable at negligible magnitudes. These quantities are conditional on the seed-42 checkpoint of each model and assume image-level independence.

Retraining variation is assessed at the seed level, on the macro unseen Dice of three runs per model (seeds 42, 43 and 44; sample SD, divisor $n - 1$). Equivalence with Polyp-PVT and VM-UNet, on the benchmark targets and on PolypGen, is tested by two one-sided Welch t-tests (TOST) [50, 51] at $\alpha = 0.05$ with margin $\delta = 0.02$ Dice. The margin was

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7615100

Table 4
Dice (seed-42 checkpoints) and computational cost (352 × 352, batch 1, FP32). Seen-domain values (†) are validation results used for checkpoint selection. Accuracy uses three-view test-time augmentation; FPS is single-view. Best value per column in bold.

Method	Params (M)	GMACs	FPS	Mem. (MB)	Kvasir†	ClinicDB†	ColonDB	ETIS	CVC-300	Macro unseen
U-Net (R34) [4, 41]	24.44	14.80	**94.4**	239	0.9315	0.9347	0.6902	0.6471	0.8803	0.7392
PraNet [5]	32.55	13.20	29.3	284	0.9295	0.9403	0.7206	0.5989	**0.8992**	0.7396
Polyp-PVT [6]	25.11	10.00	23.6	386	**0.9419**	**0.9470**	0.7783	0.8030	0.8822	0.8212
SSFormer-S [7]	29.96	20.90	12.5	324	0.9351	0.9408	0.7417	0.6998	0.8868	0.7761
VM-UNet [8]	44.27	18.20	22.4	378	0.9408	0.9463	0.7872	**0.8137**	0.8566	0.8192
WaveLite-DG (ours)	**7.68**	**1.93**	31.9	**79**	0.9385	0.9453	**0.8093**	0.7965	0.8854	**0.8304**

Table 5
Macro unseen Dice per training seed. SD is the sample standard deviation over three seeds. Best mean in bold.

Model	Seed 42	Seed 43	Seed 44	Mean ± SD
U-Net (R34)	0.7392	0.7287	0.7524	0.7401 ± 0.0119
PraNet	0.7396	0.7513	0.7268	0.7392 ± 0.0123
SSFormer-S	0.7761	0.7924	0.7657	0.7781 ± 0.0135
Polyp-PVT	0.8212	0.8316	0.8279	0.8269 ± 0.0053
VM-UNet	0.8192	0.8269	0.8246	0.8236 ± 0.0040
WaveLite-DG	0.8304	0.8296	0.8300	**0.8300 ± 0.0004**

fixed before the additional baseline seeds were run, as roughly twice the seed-to-seed range then observed for these baselines (0.0077–0.0104). Differences from U-Net (R34), PraNet and SSFormer-S are tested by two-sided Welch t-tests with Holm correction over the three comparisons [49].

5. Results

5.1. Training behaviour

Validation Dice is essentially flat from about epoch 40, with the retained checkpoint at epoch 78 (0.9411), so selection is insensitive to the stopping epoch; the stochastic perturbation causes no visible instability (Fig. 4).

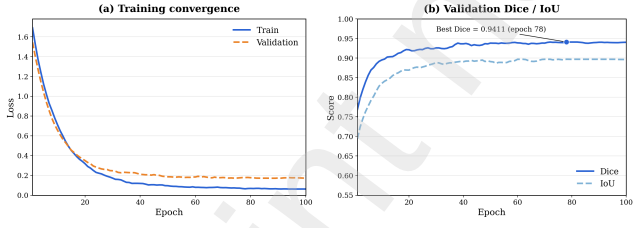


Fig. 4. Training behaviour of WaveLite-DG. **(a)** Training and validation loss. **(b)** Validation Dice and IoU, with the retained checkpoint marked (epoch 78, Dice 0.9411).

5.2. Cross-dataset accuracy

All six models perform similarly on seen data (0.9295–0.9470); separation emerges only on unseen data, where macro Dice spans 0.7392–0.8304 (Table 4, Fig. 5). On seed 42, WaveLite-DG ranks first, ahead of Polyp-PVT (+0.0092) and VM-UNet (+0.0112), with larger margins over SSFormer-S (+0.0543), PraNet (+0.0908) and U-Net (R34) (+0.0912). No model dominates every domain: WaveLite-DG leads on CVC-ColonDB, VM-UNet on ETIS and PraNet on CVC-300.

Table 6
Seed-level comparison, WaveLite-DG minus baseline (three seeds each, Welch). Equivalence: two one-sided tests, margin ±0.02, $\alpha = 0.05$. Superiority: two-sided p, Holm-adjusted over the three weaker baselines.

Baseline	Δ	90% CI	TOST p	p	Holm p
Polyp-PVT	+0.0031	[−0.0058, +0.0120]	0.016	0.42	–
VM-UNet	+0.0064	[−0.0003, +0.0131]	0.014	0.11	–
SSFormer-S	+0.0519	[+0.0293, +0.0746]	–	0.022	0.022
PraNet	+0.0908	[+0.0701, +0.1114]	–	0.006	0.017
U-Net (R34)	+0.0899	[+0.0699, +0.1099]	–	0.006	0.017

WaveLite-DG's aggregate advantage comes chiefly from avoiding the severe ETIS degradation of U-Net, PraNet and SSFormer-S. Its seed-42 margins over Polyp-PVT and VM-UNet are not by themselves evidence of an advantage: with seed 43, Polyp-PVT reaches 0.8316 (ColonDB 0.8126, ETIS 0.7814, CVC-300 0.9007), above every WaveLite-DG run, and VM-UNet 0.8269 (0.7589, 0.8425, 0.8794). Over three seeds per model (Table 5), WaveLite-DG has the highest mean macro unseen Dice, 0.8300 ± 0.0004, against 0.8269 ± 0.0053 for Polyp-PVT and 0.8236 ± 0.0040 for VM-UNet. Neither lead is significant (two-sided Welch $p = 0.42$ and 0.11), whereas both 90% confidence intervals lie inside the ±0.02 margin (TOST $p = 0.016$ and 0.014; Table 6), so WaveLite-DG is statistically equivalent to both; neither interval lies inside ±0.01. Against SSFormer-S, PraNet and U-Net (R34) the three-seed differences, +0.0519 to +0.0908, remain significant after Holm correction ($p \leq 0.022$). We therefore describe the comparison with the two strongest baselines as parity, consistent with seed variation reported elsewhere [25].

Table 7 shows that the result is not an artefact of one metric: on seed 42, WaveLite-DG has the highest macro unseen IoU (0.7517) and precision (0.8225), sensitivity marginally below Polyp-PVT, and specificity above 0.99, so accuracy is not bought by over-segmentation.

5.3. Computational efficiency

WaveLite-DG has 68.6–82.7% fewer parameters than every full-size baseline, needs 5.2–10.8× fewer MACs, and uses 79 MB of peak memory, about one third of the next-lowest model (Table 4). Throughput does not follow MACs: U-Net (R34) is fastest (94.4 FPS) despite 7.7× more computation, while WaveLite-DG is second (31.9 FPS). Depthwise separable convolutions have low arithmetic intensity and

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7615100

Table 7
Macro unseen complementary metrics, and pooled unseen Dice (n = 635; ColonDB 379, ETIS 196, CVC-300 60), seed-42 checkpoints. Best in bold.

Model	IoU	Sens.	Spec.	Prec.	Pooled Dice
U-Net (R34)	0.6757	0.7656	0.9900	0.7658	0.6949
PraNet	0.6793	0.7556	**0.9960**	0.7547	0.6999
Polyp-PVT	0.7441	**0.8822**	0.9896	0.8126	0.7957
SSFormer-S	0.6997	0.8368	0.9884	0.7639	0.7425
VM-UNet	0.7457	0.8726	0.9911	0.8151	0.8019
WaveLite-DG	**0.7517**	0.8780	0.9914	**0.8225**	**0.8125**

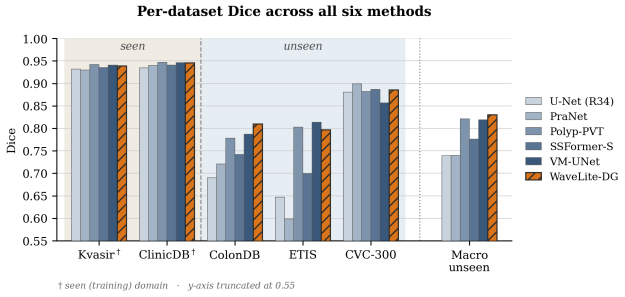


Fig. 5. Per-dataset Dice for all six methods, seed-42 checkpoints; seen domains marked †, macro unseen average at right. The y-axis is truncated at 0.55, which exaggerates visual separation.

Table 8
Lightweight networks under the same protocol: macro unseen Dice, mean ± SD over three seeds; parameters in millions, cost at 352×352, batch 1, single-view GPU throughput. Best in bold.

Model	Params	GMACs	FPS	Macro unseen
U-Net/MobileNetV3-L	5.41	2.36	27.4	0.7642 ± 0.0068
DeepLabV3+/MobileNetV2	4.87	2.78	24.9	0.7516 ± 0.0094
FPN/EfficientNet-B0	6.12	2.21	26.8	0.7873 ± 0.0071
UNeXt	**1.47**	**0.59**	**48.6**	0.7185 ± 0.0127
WaveLite-DG	7.68	1.93	31.9	**0.8300 ± 0.0004**

are bandwidth-bound [52]; effective utilization is about 0.12 TFLOP/s for WaveLite-DG against 2.8 TFLOP/s for U-Net (R34). Because accuracy uses three-view test-time augmentation while FPS is single-view, Fig. 6 pairs different operating points; the between-model comparison is valid because both conventions apply to all six models, but the absolute frontier is optimistic. Measured end to end, the three-view rate is 10.6 FPS for WaveLite-DG; the 11.0 FPS measured for the same architecture without the perturbation is measurement noise, since the perturbation is disabled at inference and the two deployed networks are identical.

These advantages hold within the six-model comparison only. Among lightweight networks (Table 8), UNeXt is smaller, cheaper and faster than WaveLite-DG (1.47 M, 0.59 GMACs, 48.6 FPS), but WaveLite-DG is the most accurate, 0.0427 above the best of them, FPN with EfficientNet-B0 (0.7873 ± 0.0071).

5.4. Paired per-image analysis

Every mean difference is positive and every Holm-adjusted p is below 0.05 (Table 9, Fig. 7), but significance and magnitude diverge. Against U-Net (R34), PraNet and SSFormer-S effects are large (d_z = 0.90–1.20). Against Polyp-PVT and VM-UNet they are small (d_z = 0.25, 0.16; medians 0.0069, 0.0058), so on a typical image the three models are nearly interchangeable, and the small p-values reflect sample size. The effect sizes are stable across retraining: on seeds 43 and 44 they stay small against Polyp-PVT (0.21, 0.27) and VM-UNet (0.19, 0.23) and large against the three weaker baselines (0.84–1.27). This is consistent with parity. The intervals and p-values are conditional on the seed-42 checkpoints; equivalence across retraining is tested at the seed level (Section 5.2, Table 6).

Pooled Dice falls below macro Dice for every model (Table 7), because the easiest dataset, CVC-300, carries one third of the macro weight but under one tenth of the pooled weight. CVC-ColonDB supplies 59.7% of pooled images, and WaveLite-DG's largest margin is there; hence pooled differences exceed macro differences for four baselines. VM-UNet is the exception, as its strength on ETIS gains weight in the pooled view.

5.5. Ablation and interaction

The boundary head adds +0.0107 macro Dice, the detail path +0.0105, the perturbation without the gated scan +0.0109 at zero parameters, and the gated scan without perturbation +0.0068 (Table 10). The decisive comparison is A4 versus A5: identical architecture and 7.68 M parameters, differing only in whether the perturbation is active. Enabling it raises macro Dice by 0.0646 (+0.0774 ColonDB, +0.1067 ETIS, +0.0098 CVC-300), a gain that cannot come from capacity. Repeating both configurations over three seeds reproduces it: 0.8300 ± 0.0004 against 0.7649 ± 0.0020, a difference of +0.0651 that is more than thirty times the seed spread. This gain is measured without photometric augmentation; with colour jitter it falls to +0.0185 (0.8369 against 0.8184; Section 5.6), so ordinary colour jitter recovers most of it. Overall, A5 improves on A0 by 0.0926 at a cost of 7.8 FPS. The base model A0 (0.7358 ± 0.0040) is itself weaker than an off-the-shelf U-Net on the same MobileNetV3-Large encoder (0.7642 ± 0.0068, Table 8), so the components added in A1–A5 build on a weaker starting point than an off-the-shelf design.

A2–A5 form a complete 2×2 factorial in perturbation and gated scan, with the detail path and boundary head active. Over three seeds the additive prediction $D_{A3} + D_{A4} - D_{A2} =$ 0.7738 falls 0.0562 short of the observed 0.8300, a residual an order of magnitude larger than either main effect and more than ten times the largest seed spread in the factorial. In three-seed means it is large on the most shifted datasets (ColonDB +0.0742, ETIS +0.1003) and slightly negative on CVC-300 (−0.0058). One consistent hypothesis is that diversified approximation-band statistics force the scan to encode appearance-invariant spatial relationships. Two limits remain: the factorial is conditional on the detail path and

Preprint not peer reviewed
This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7615100

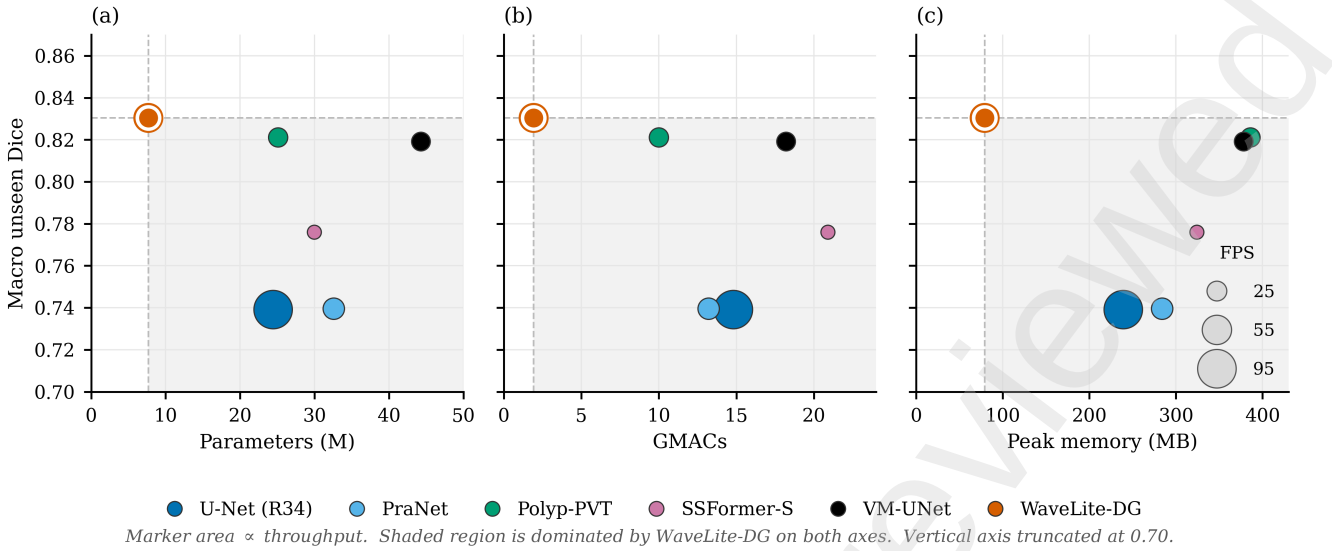


Fig. 6. Accuracy–efficiency frontier: macro unseen Dice (seed-42 checkpoints) against parameters, GMACs and peak memory. Marker area encodes single-view throughput; the proposed model is drawn as a ringed circle. Upper-left is preferable.

Table 9
Paired comparison on pooled unseen per-image Dice ($n = 635$), WaveLite-DG minus baseline, seed-42 checkpoints except the last two columns. 95% CI is a percentile bootstrap; Holm p is Holm-adjusted. d_z is the standardized paired effect size, given for each seed.

vs. baseline	Mean Δ	95% CI	Median Δ	SD	d_z	Wilcoxon p	Holm p	d_z, s43	d_z, s44
U-Net (R34)	0.1177	[0.1103, 0.1256]	0.1071	0.098	1.20	2.11×10^{-95}	1.06×10^{-94}	1.14	1.27
PraNet	0.1126	[0.1041, 0.1209]	0.0933	0.108	1.04	2.49×10^{-86}	9.95×10^{-86}	1.11	0.98
SSFormer-S	0.0701	[0.0640, 0.0762]	0.0522	0.078	0.90	4.78×10^{-75}	1.43×10^{-74}	0.84	0.96
Polyp-PVT	0.0168	[0.0116, 0.0221]	0.0069	0.068	0.25	1.57×10^{-7}	3.13×10^{-7}	0.21	0.27
VM-UNet	0.0106	[0.0054, 0.0160]	0.0058	0.068	0.16	2.78×10^{-5}	2.78×10^{-5}	0.19	0.23

boundary head being active, and no direct evidence supports the mechanism — the MMD analysis below and the attribution maps in Section 5.8 both complicate it.

The size of the interaction is specific to the gated scan. Replacing it with other bottleneck modules of about 0.66 M parameters (Supplementary Table S2), the perturbation adds +0.0088 with no bottleneck module, +0.0166 with a plain convolution block, +0.0213 with self-attention and +0.0651 with the gated scan. Like the factorial, this is shown only without photometric augmentation; under colour jitter the perturbation gain with the gated scan is +0.0185.

Varying the perturbation on seed 42 (Supplementary Table S3), $\lambda = 0.3$ and $p = 0.5$ give both the highest source validation Dice (0.9411) and the highest unseen Dice (0.8304), although validation differences between settings are only about 0.001. Unseen Dice stays within 0.007 for λ from 0.2 to 0.4 (0.8239–0.8304).

Table 10
Component ablation on unseen data. HF = high-frequency path; LL = approximation-band perturbation; Bnd = boundary head; Gate = gated context scan. Per-dataset values are seed 42; the last column is the mean ± SD of the macro score over seeds 42, 43 and 44 where three runs were made. A4–A7 have identical parameter counts. A6 and A7 replace the approximation-band perturbation with the same operation applied to the undecomposed feature, at matched λ and at matched injected energy respectively.

ID	Configuration	HF	LL	Bnd	Gate	Params (M)	ColonDB	ETIS	CVC-300	Macro	Macro, 3 seeds
A0	Base model	–	–	–	–	6.45	0.7012	0.6598	0.8524	0.7378	0.7358 ± 0.0040
A1	+ boundary head	–	–	✓	–	6.58	0.7146	0.6721	0.8587	0.7485	0.7469 ± 0.0030
A2	+ HF path	✓	–	✓	–	7.02	0.7258	0.6849	0.8662	0.7590	0.7605 ± 0.0015
A3	+ LL perturbation	✓	✓	✓	–	7.02	0.7364	0.6937	0.8796	0.7699	0.7693 ± 0.0011
A4	Full arch., no LL pert.	✓	–	✓	✓	7.68	0.7319	0.6898	0.8756	0.7658	0.7649 ± 0.0020
A6	Whole-feature pert., matched λ	✓	–	✓	✓	7.68	0.7917	0.7469	0.8890	0.8092	0.8195 ± 0.0093
A7	Whole-feature pert., matched energy	✓	–	✓	✓	7.68	0.8008	0.7731	0.9019	0.8253	0.8286 ± 0.0029
A5	Full WaveLite-DG	✓	✓	✓	✓	7.68	**0.8093**	**0.7965**	**0.8854**	**0.8304**	**0.8300 ± 0.0004**

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7615100

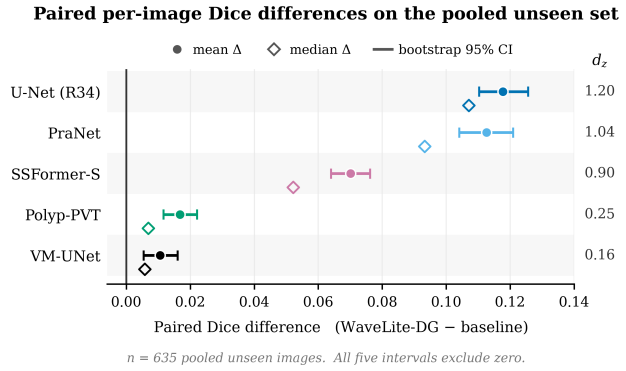


Fig. 7. Paired per-image Dice differences on the pooled unseen set: mean (circle), median (diamond) and 95% bootstrap interval, with d_z at right.

5.6. Is the wavelet restriction necessary?

Section 3.3.4 showed that restricting the perturbation to the approximation band is not equivalent to perturbing the undecomposed feature, but not that it is preferable. Table 10 now reports that control. Perturbing the whole feature at matched λ (A6) reaches 0.8195 ± 0.0093 and at matched injected energy (A7) 0.8286 ± 0.0029, against 0.8300 ± 0.0004 for the wavelet-restricted operation. The difference at matched energy, +0.0014, is smaller than the SD of A7 (0.0029): there is no detectable difference between the wavelet-restricted and the whole-feature perturbation. The benefit comes from perturbing feature statistics, not from confining that perturbation to the approximation band.

Established feature-statistic methods inserted at the same three positions support this reading (Table 11): DSU, FDA and MixStyle land within 0.010 of WaveLite-DG, all differences being small relative to the 0.0093–0.0123 seed spread of the weaker entries.

Colour jitter changes the picture (Table 12). It raises every model: the three weaker baselines by +0.0345 to +0.0436, Polyp-PVT and VM-UNet by +0.0169 and +0.0175, A4 by +0.0535, but WaveLite-DG by only +0.0069. The perturbation gain therefore falls from +0.0651 to +0.0185: most of what the perturbation contributes without photometric augmentation is recovered by ordinary colour jitter. With jitter, Polyp-PVT and VM-UNet lead WaveLite-DG by 0.0069 and 0.0042. Adding the perturbation to the skip features of Polyp-PVT and VM-UNet, without jitter, raises them to 0.8378 and 0.8346, 0.0078 and 0.0046 above WaveLite-DG, so the perturbation transfers to other architectures and is not what sets WaveLite-DG apart. Equivalence under jitter is not tested. In every regime tested, WaveLite-DG remains within 0.008 of the best model with one third or less of the parameters and one fifth or less of the computation.

Two further results bear on the mechanism. A maximum-mean-discrepancy analysis of bottleneck features does *not* support the account offered in Section 5.5: the mean pairwise MMD between the three unseen domains is larger with the perturbation (0.1478) than without it (0.1273). On contour-sensitive measures WaveLite-DG records the lowest HD95 (26.3 px against 32.6–88.5) and a Boundary IoU of 0.612, inside the baselines' 0.554–0.623 range.

Table 11
Domain-generalization mechanisms in the same architecture and protocol, macro unseen Dice, mean ± SD over three seeds. Polyp-PVT and VM-UNet, separate architectures, are shown for reference.

Mechanism	Where applied	Macro unseen Dice
WaveLite-DG (A5)	LL band, 3 scales	**0.8300 ± 0.0004**
A7	whole feature, matched energy	0.8286 ± 0.0029
DSU [24]	same 3 skip features	0.8275 ± 0.0016
Polyp-PVT [6]	separate architecture	0.8269 ± 0.0053
FDA [26]	input amplitude	0.8254 ± 0.0048
VM-UNet [8]	separate architecture	0.8236 ± 0.0040
MixStyle [23]	same 3 skip features	0.8202 ± 0.0123
A6	whole feature, matched λ	0.8195 ± 0.0093

Table 12
Augmentation and transfer controls: macro unseen Dice, three-seed mean. + jitter: retrained with colour jitter. + perturbation: approximation-band perturbation added to the baseline's skip features, without jitter. Best per column in bold.

Model	No jitter	+ jitter	+ perturbation
U-Net (R34)	0.7401	0.7837	–
PraNet	0.7392	0.7764	–
SSFormer-S	0.7781	0.8126	–
Polyp-PVT	0.8269	**0.8438**	**0.8378**
VM-UNet	0.8236	0.8411	0.8346
A4 (no perturbation)	0.7649	0.8184	–
WaveLite-DG (A5)	**0.8300**	0.8369	–

5.7. Multi-centre evaluation

As a check beyond the three standard target sets, the checkpoints of all three seeds were evaluated without any retraining on PolypGen [53] (1537 single frames from six clinical centres; Supplementary Table S4). WaveLite-DG reaches 0.7844 ± 0.0054 macro Dice, Polyp-PVT 0.7833 ± 0.0078 and VM-UNet 0.8010 ± 0.0069. WaveLite-DG is comparable to Polyp-PVT (difference +0.0011, 90% CI [−0.0110, +0.0131], two-sided Welch $p = 0.856$; TOST $p = 0.015$, equivalent within ±0.02), whereas VM-UNet is ahead (difference −0.0166, 90% CI [−0.0276, −0.0056], $p = 0.034$; TOST $p = 0.271$, not equivalent). On multi-centre data the parity therefore holds against Polyp-PVT but not against VM-UNet. We additionally trained a much smaller baseline, Mobile-PolypNet [17] (0.22 M parameters, 2.44 GMACs, 62.3 FPS), under its published longer schedule, since it has no pretrained backbone: it reaches 0.6969±0.0100 macro unseen Dice over three seeds, 0.133 below WaveLite-DG. A model an order of magnitude smaller therefore remains well short of the accuracy reported here.

5.8. Error modes and qualitative results

Fig. 8 expresses precision and sensitivity as over-segmentation (1−precision) and missed tissue (1−sensitivity). Both rise sharply under shift: from 0.043–0.067 on seen data, over-segmentation reaches 0.164 (ColonDB) and 0.202 (ETIS), and missed tissue 0.170 and 0.164. Because the two

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7615100

rates have different denominators, only CVC-300 (0.166 against 0.031) supports an ordering. The boundary-distance profile, free of this problem, shows errors peaking at the contour and decaying within about five pixels on seen data, whereas unseen data retains a shoulder of 0.05–0.08 extending some 25 pixels into adjacent background: under shift the mask spreads into surrounding tissue.

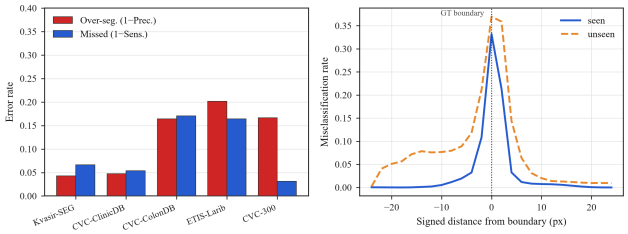


Fig. 8. Error modes of WaveLite-DG. **Left:** over-segmentation (1−precision) and missed tissue (1−sensitivity); the bars have different denominators and are not directly comparable. **Right:** misclassification rate against signed distance from the ground-truth contour (negative = exterior).

Fig. 9 shows predictions on all five datasets, including failures at Dice 0.496 (specular reflection, CVC-ColonDB) and 0.438 (a fragmented flat lesion, ETIS). As a negative result, Grad-CAM [54] at the stride-32 bottleneck does not localize the lesion on CVC-ColonDB and ETIS even where masks are accurate. Either bottleneck Grad-CAM is not diagnostic for a decoder fed by three skip pathways, or the mechanism proposed above is wrong; the observation does not support that mechanism.

5.9. *Comparison with published values*

Retrained baselines fall below their published unseen Dice in six of nine cases, for example Polyp-PVT on ColonDB (0.7783 against 0.808) and SSFormer-S on ETIS (0.6998 against 0.767) [5, 6, 7]. Our smaller training set (1289 against 1450 images), source-only checkpoint selection, the shared recipe and the absence of photometric augmentation all plausibly contribute; standardized recipes are known to shift scores by amounts comparable to architectural differences [55]. Our ColonDB margin over Polyp-PVT (+0.031) is similar to its reproduction shortfall, so we do not claim that WaveLite-DG would beat each baseline under its own tuned protocol — the more so as the shared recipe was developed with WaveLite-DG.

6. Discussion

A compact architecture reached parity with models three to six times larger on unseen data under a controlled protocol. Without photometric augmentation its three-seed mean macro unseen Dice is the highest of the six models and statistically equivalent to Polyp-PVT and VM-UNet within a 0.02 margin; with colour jitter, or with the perturbation added to those baselines, they lead by under 0.008. In every regime tested it stays within 0.008 of the best model with one third or less of the parameters and one fifth or less of the computation, and it is the only model in the six-model comparison to occupy the high-accuracy, low-cost corner of Fig. 6. Among lightweight networks UNeXt is smaller, cheaper and faster, but WaveLite-DG is 0.0427 more accurate than the best of them. The accuracy margin over the two strongest baselines is small or negative; the efficiency margin is large and structural. Approximation-band perturbation contributed +0.0651 at zero parameters and zero inference cost without photometric augmentation, but only +0.0185 with colour jitter, which recovers most of the gain; its super-additive interaction with the gated scan is likewise shown only without photometric augmentation. At matched injected energy the wavelet restriction makes no detectable difference (Section 5.6): the mechanism is best described as feature-statistic perturbation, and the approximation-band form is not shown to be better than whole-feature perturbation. On multi-centre PolypGen data WaveLite-DG is comparable to Polyp-PVT, but VM-UNet is ahead.

The efficiency figures have concrete deployment consequences. Memory, often the binding constraint, is 79 MB, leaving room to run the model alongside a detection or quality-assurance model on one device; at 31.9 FPS single-view it permits a live overlay, whereas three-view augmentation (10.6 FPS) suits still-frame review; and the 7.55 M deployed model is small enough for routine fleet updates. The MAC–throughput dissociation shows that parameters and MACs alone overstate practical speed, so measured throughput belongs alongside them.

These observations concern segmentation quality only. They say nothing about detection, clinician interaction or clinical outcomes, which require prospective video evaluation.

7. Limitations and future work

Every model, ablation cell and mechanism comparison is repeated over three seeds, but equivalence with Polyp-PVT and VM-UNet rests on three seeds per model and a 0.02 margin, would not hold at a margin of 0.01, and is untested under colour jitter. The per-image bootstrap intervals of Table 9, though not the effect sizes, remain conditional on the seed-42 checkpoints and describe test-image rather than retraining variation; image-level resampling also ignores correlated video frames. The margins over Polyp-PVT, VM-UNet, A7, DSU and FDA are all smaller than the seed spread of the weaker entries, so the accuracy claim remains one of parity, and on PolypGen VM-UNet is ahead. The interaction between perturbation and gated scan is shown only without photometric augmentation, the regime in which the perturbation's own gain is largest. The mechanism is not explained: the MMD analysis of Section 5.6 contradicts the account we find most natural, and no positive evidence replaces it. The sensitivity analysis of λ and p uses one seed. The SSFormer-S reproduction gap on ETIS remains unexplained. Throughput is measured on one laptop GPU only, object-level detection metrics are absent, and the evaluation, PolypGen included, is retrospective. Future work will test equivalence under colour jitter, seek direct evidence for the

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7615100

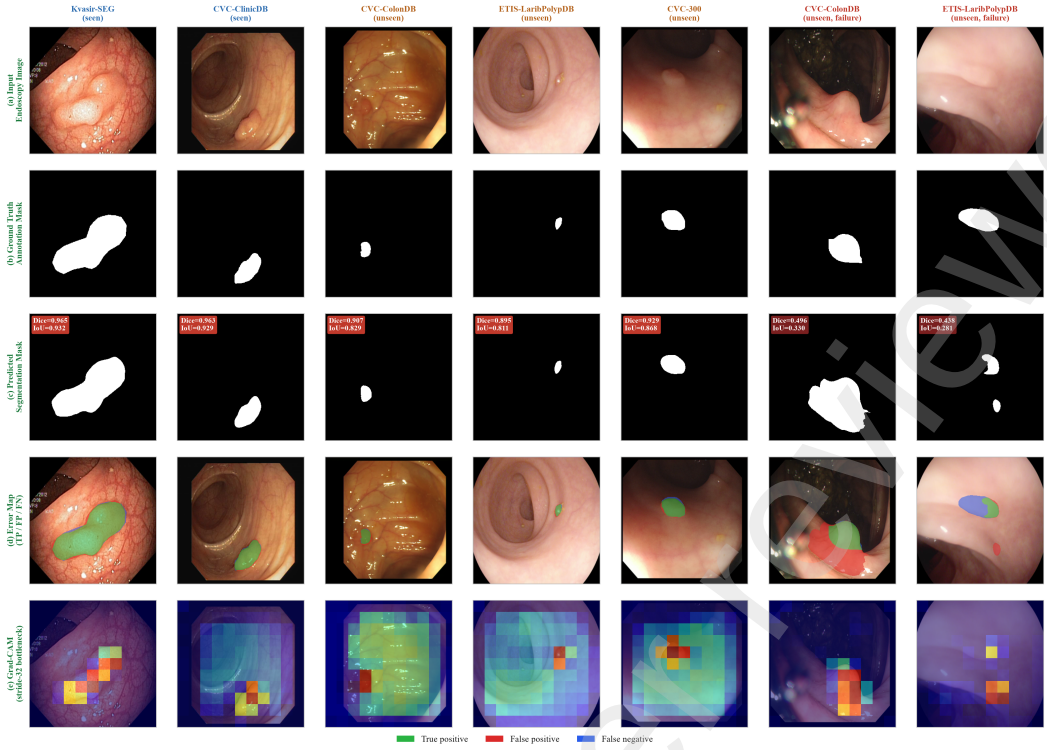


Fig. 9. Representative outputs with two failure cases. Rows: **(a)** image; **(b)** ground truth; **(c)** prediction with Dice and IoU; **(d)** error map (green TP, red FP, blue FN); **(e)** bottleneck Grad-CAM at native 11×11 resolution, min–max normalized per column.

mechanism, profile on CPU and embedded hardware, and evaluate on prospective video.

8. Conclusion

WaveLite-DG, a 7.68 M-parameter wavelet-guided network, attained the highest three-seed mean macro unseen Dice (0.8300 ± 0.0004) among six architectures retrained under a common protocol without photometric augmentation, statistically equivalent to Polyp-PVT and VM-UNet within a 0.02 margin, with the lowest computation (1.93 GMACs) and memory (79 MB) of the six. With colour jitter, or with the perturbation added to those two baselines, they lead by under 0.008; in every regime tested WaveLite-DG stays within 0.008 of the best model with one third or less of the parameters and one fifth or less of the computation. Among lightweight networks it is the most accurate, by 0.0427, though not the smallest or fastest. Its accuracy is best described as parity; its efficiency is the principal claim. Approximation-band perturbation added +0.0651 macro Dice without photometric augmentation but +0.0185 with colour jitter, which recovers most of the gain, and its +0.0562 super-additive interaction with the gated scan is likewise shown only without photometric augmentation. At matched injected energy, whole-feature perturbation shows no detectable difference, so the benefit comes from perturbing feature statistics rather than from the wavelet restriction. Parity with DSU, FDA and MixStyle in the same architecture, and with Polyp-PVT but not VM-UNet on multi-centre PolypGen data, places the contribution in efficiency rather than in a new mechanism.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work the authors used Claude (Anthropic) in order to format, language-edit, condense and revise the manuscript text, and to compute the seed-level statistical tests (Welch *t*-tests, two one-sided tests and confidence intervals) from the authors' per-seed results. The experiments and all reported model results were produced by the authors. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

CRediT authorship contribution statement

Mostofa Adib Shakib: Conceptualization, Methodology, Software, Formal analysis, Investigation, Validation, Writing – original draft, Writing – review & editing. **Lotifur Shabbir:** Methodology, Software, Formal analysis, Investigation, Data curation, Writing – review & editing. **Abraar A Rehman:** Investigation, Data curation, Validation, Writing – review & editing. **Sajidur Rahman Siam:** Investigation, Data curation, Validation, Writing – review & editing. **Nazifa Mouli:** Formal analysis, Investigation, Visualization, Writing – review & editing. **Sadia Hamid:** Methodology, Software, Formal analysis, Investigation, Validation, Visualization, Writing – review & editing. **Adnan Safin:** Investigation, Validation, Visualization, Writing – review & editing. **Tauhidur Rahman Sakib:** Conceptualization, Methodology,

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7615100

Software, Formal analysis, Investigation, Validation, Data curation, Visualization, Writing – original draft, Writing – review & editing, Supervision, Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Ethics statement

This study used publicly available, de-identified datasets and involved no recruitment of or intervention on human participants. No institutional review board approval was required. The datasets do not report patient sex or gender at the image level, so no sex- or gender-based analysis was possible, which limits the generalizability of the results across these groups.

Data availability

All six datasets are publicly available from their original sources [33, 34, 35, 36, 37, 53]. Source code, train/validation/test manifests, SHA-256 hashes and the duplicate-audit results are not deposited in a public repository because they are not maintained as a public software release; they are available from the corresponding author on reasonable request (tauhidurrahmansakib@gmail.com).

Acknowledgements

The authors gratefully acknowledge the physicians who provided valuable clinical guidance and expert feedback during the development of this study. Their insights into colonoscopy practice, polyp characteristics, clinical relevance and interpretation of the segmentation results helped strengthen the medical context of the work. The authors sincerely appreciate their time, support and constructive suggestions. These contributors were not involved in the computational experiments, data analysis or preparation of the final results and did not meet the criteria for authorship.

References

[1] F. Bray, M. Laversanne, H. Sung, J. Ferlay, R.L. Siegel, I. Soerjomataram, A. Jemal, Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, CA Cancer J. Clin. 74 (3) (2024) 229–263, https://doi.org/10.3322/caac.21834.

[2] A.G. Zauber, et al., Colonoscopic polypectomy and long-term prevention of colorectal-cancer deaths, N. Engl. J. Med. 366 (8) (2012) 687–696, https://doi.org/10.1056/NEJMoa1100370.

[3] S. Zhao, et al., Magnitude, risk factors, and factors associated with adenoma miss rate of tandem colonoscopy: a systematic review and meta-analysis, Gastroenterology 156 (6) (2019) 1661–1674, https://doi.org/10.1053/j.gastro.2019.01.260.

[4] O. Ronneberger, P. Fischer, T. Brox, U-Net: convolutional networks for biomedical image segmentation, in: Proc. MICCAI, 2015, pp. 234–241, https://doi.org/10.1007/978-3-319-24574-4_28.

[5] D.-P. Fan, G.-P. Ji, T. Zhou, G. Chen, H. Fu, J. Shen, L. Shao, PraNet: parallel reverse attention network for polyp segmentation, in: Proc. MICCAI, 2020, pp. 263–273, https://doi.org/10.1007/978-3-030-59725-2_26.

[6] B. Dong, W. Wang, D.-P. Fan, J. Li, H. Fu, L. Shao, Polyp-PVT: polyp segmentation with pyramid vision transformers, CAAI Artif. Intell. Res. 2 (2023) 9150015, https://doi.org/10.26599/AIR.2023.9150015.

[7] J. Wang, Q. Huang, F. Tang, J. Meng, J. Su, S. Song, Stepwise feature fusion: local guides global, in: Proc. MICCAI, 2022, pp. 110–120, https://doi.org/10.1007/978-3-031-16437-8_11.

[8] J. Ruan, S. Xiang, VM-UNet: Vision Mamba UNet for medical image segmentation, arXiv preprint arXiv:2402.02491 (2024), https://doi.org/10.48550/arXiv.2402.02491.

[9] Z. Zhou, M.M.R. Siddiquee, N. Tajbakhsh, J. Liang, UNet++: a nested U-Net architecture for medical image segmentation, in: Proc. DLMIA, 2018, pp. 3–11, https://doi.org/10.1007/978-3-030-00889-5_1.

[10] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, P. Torr, Res2Net: a new multi-scale backbone architecture, IEEE Trans. Pattern Anal. Mach. Intell. 43 (2) (2021) 652–662, https://doi.org/10.1109/TPAMI.2019.2938758.

[11] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, L. Shao, PVT v2: improved baselines with pyramid vision transformer, Comput. Vis. Media 8 (3) (2022) 415–424, https://doi.org/10.1007/s41095-022-0274-8.

[12] A. Gu, T. Dao, Mamba: linear-time sequence modeling with selective state spaces, in: Proc. Conf. Lang. Model. (COLM), 2024, https://doi.org/10.48550/arXiv.2312.00752.

[13] Y. Liu, Y. Tian, Y. Zhao, H. Yu, L. Xie, Y. Wang, Q. Ye, J. Jiao, Y. Liu, VMamba: visual state space model, in: Adv. Neural Inf. Process. Syst. 37 (NeurIPS), 2024, https://doi.org/10.52202/079017-3273.

[14] K. ELKarazle, V. Raman, C. Chua, P. Then, SCABNet: a novel polyp segmentation network with spatial-gradient attention and channel prioritization, Int. J. Imaging Syst. Technol. 35 (2) (2025) e70039, https://doi.org/10.1002/ima.70039.

[15] D. Li, C. Liu, Z. Xiao, CL-Polyp: a contrastive learning-enhanced network for accurate polyp segmentation, arXiv preprint arXiv:2507.07154 (2025), https://doi.org/10.48550/arXiv.2507.07154.

[16] A. Qayoom, J. Xie, H. Ali, Polyp segmentation in medical imaging: challenges, approaches and future directions, Artif. Intell. Rev. 58 (6) (2025) 169, https://doi.org/10.1007/s10462-025-11173-2.

[17] R. Karmakar, S. Nooshabadi, Mobile-PolypNet: lightweight colon polyp segmentation network for low-resource settings, J. Imaging 8 (6) (2022) 169, https://doi.org/10.3390/jimaging8060169.

[18] T. Erol, D. Sarikaya, An efficient polyp segmentation network, arXiv preprint arXiv:2203.04118 (2022), https://doi.org/10.48550/arXiv.2203.04118.

[19] H. Xu, J. Wang, Z. Li, S. Teng, X. Cheng, Lite-PolypNet: a lightweight and efficient network for polyp segmentation in colonoscopy images, Int. J. Imaging Syst. Technol. 35 (5) (2025) e70215, https://doi.org/10.1002/ima.70215.

[20] H. Wu, Z. Zhao, EPSegNet: lightweight semantic recalibration and assembly for efficient polyp segmentation, IEEE Trans. Neural Netw. Learn. Syst. 36 (8) (2025) 13805–13817, https://doi.org/10.1109/TNNLS.2025.3527557.

[21] K. ELKarazle, V. Raman, C. Chua, P. Then, M. Prabhavathy, EfficientPolySeg: a fast and accurate network for complex polyp segmentation, Biomed. Signal Process. Control 112 (2026) 108449, https://doi.org/10.1016/j.bspc.2025.108449.

[22] X. Huang, S. Belongie, Arbitrary style transfer in real-time with adaptive instance normalization, in: Proc. IEEE ICCV, 2017, pp. 1510–1519, https://doi.org/10.1109/ICCV.2017.167.

[23] K. Zhou, Y. Yang, Y. Qiao, T. Xiang, Domain generalization with MixStyle, in: Proc. ICLR, 2021, https://openreview.net/forum?id=6xHJ37MVxxp.

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7615100

[24] X. Li, Y. Dai, Y. Ge, J. Liu, Y. Shan, L.-Y. Duan, Uncertainty modeling for out-of-distribution generalization, in: Proc. ICLR, 2022, https://openreview.net/forum?id=6HN7LHyzGgC.
[25] I. Gulrajani, D. Lopez-Paz, In search of lost domain generalization, in: Proc. ICLR, 2021, https://openreview.net/forum?id=lQdXeXDoWtI.
[26] Y. Yang, S. Soatto, FDA: Fourier domain adaptation for semantic segmentation, in: Proc. IEEE/CVF CVPR, 2020, pp. 4085–4095, https://doi.org/10.1109/CVPR42600.2020.00414.
[27] Q. Xu, R. Zhang, Y. Zhang, Y. Wang, Q. Tian, A Fourier-based framework for domain generalization, in: Proc. IEEE/CVF CVPR, 2021, pp. 14383–14392, https://doi.org/10.1109/CVPR46437.2021.01415.
[28] Q. Liu, C. Chen, J. Qin, Q. Dou, P.-A. Heng, FedDG: federated domain generalization on medical image segmentation via episodic learning in continuous frequency space, in: Proc. IEEE/CVF CVPR, 2021, pp. 1013–1023, https://doi.org/10.1109/CVPR46437.2021.00107.
[29] S.G. Mallat, A theory for multiresolution signal decomposition: the wavelet representation, IEEE Trans. Pattern Anal. Mach. Intell. 11 (7) (1989) 674–693, https://doi.org/10.1109/34.192463.
[30] P. Liu, H. Zhang, K. Zhang, L. Lin, W. Zuo, Multi-level wavelet-CNN for image restoration, in: Proc. IEEE/CVF CVPRW, 2018, pp. 886–895, https://doi.org/10.1109/CVPRW.2018.00121.
[31] Y. Fu, Y. Xie, Y. Fu, J. Chen, Y.-G. Jiang, Wave-SAN: wavelet based style augmentation network for cross-domain few-shot learning, arXiv preprint arXiv:2203.07656 (2022), https://doi.org/10.48550/arXiv.2203.07656.
[32] A. Howard, et al., Searching for MobileNetV3, in: Proc. IEEE/CVF ICCV, 2019, pp. 1314–1324, https://doi.org/10.1109/ICCV.2019.00140.
[33] D. Jha, P.H. Smedsrud, M.A. Riegler, P. Halvorsen, T. de Lange, D. Johansen, H.D. Johansen, Kvasir-SEG: a segmented polyp dataset, in: Proc. MMM, 2020, pp. 451–462, https://doi.org/10.1007/978-3-030-37734-2_37.
[34] J. Bernal, F.J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, F. Vilariño, WM-DOVA maps for accurate polyp highlighting in colonoscopy: validation vs. saliency maps from physicians, Comput. Med. Imaging Graph. 43 (2015) 99–111, https://doi.org/10.1016/j.compmedimag.2015.02.007.
[35] N. Tajbakhsh, S.R. Gurudu, J. Liang, Automated polyp detection in colonoscopy videos using shape and context information, IEEE Trans. Med. Imaging 35 (2) (2016) 630–644, https://doi.org/10.1109/TMI.2015.2487997.
[36] J. Silva, A. Histace, O. Romain, X. Dray, B. Granado, Toward embedded detection of polyps in WCE images for early diagnosis of colorectal cancer, Int. J. Comput. Assist. Radiol. Surg. 9 (2) (2014) 283–293, https://doi.org/10.1007/s11548-013-0926-3.
[37] D. Vázquez, J. Bernal, F.J. Sánchez, G. Fernández-Esparrach, A.M. López, A. Romero, M. Drozdzal, A. Courville, A benchmark for endoluminal scene segmentation of colonoscopy images, J. Healthc. Eng. 2017 (2017) 4037190, https://doi.org/10.1155/2017/4037190.
[38] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, in: Proc. ICLR, 2019, https://openreview.net/forum?id=Bkg6RiCqY7.
[39] I. Loshchilov, F. Hutter, SGDR: stochastic gradient descent with warm restarts, in: Proc. ICLR, 2017, https://openreview.net/forum?id=Skq89Scxx.
[40] P. Micikevicius, et al., Mixed precision training, in: Proc. ICLR, 2018, https://openreview.net/forum?id=r1gs9JgRZ.
[41] P. Iakubovskii, Segmentation Models PyTorch [software], GitHub, 2019. https://github.com/qubvel/segmentation_models.pytorch (accessed 11 October 2026).
[42] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, in: Proc. ECCV, 2018, pp. 801–818, https://doi.org/10.1007/978-3-030-01234-2_49.
[43] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.-C. Chen, MobileNetV2: inverted residuals and linear bottlenecks, in: Proc. IEEE/CVF CVPR, 2018, pp. 4510–4520, https://doi.org/10.1109/CVPR.2018.00474.
[44] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in: Proc. IEEE CVPR, 2017, pp. 2117–2125, https://doi.org/10.1109/CVPR.2017.106.
[45] M. Tan, Q. V. Le, EfficientNet: rethinking model scaling for convolutional neural networks, in: Proc. ICML, PMLR 97, 2019, pp. 6105–6114.
[46] J. M. J. Valanarasu, V. M. Patel, UNeXt: MLP-based rapid medical image segmentation network, in: Proc. MICCAI, 2022, pp. 23–33, https://doi.org/10.1007/978-3-031-16443-9_3.
[47] B. Efron, R.J. Tibshirani, An Introduction to the Bootstrap, Chapman & Hall, New York, 1993, https://doi.org/10.1007/978-1-4899-4541-9.
[48] F. Wilcoxon, Individual comparisons by ranking methods, Biometrics Bull. 1 (6) (1945) 80–83, https://doi.org/10.2307/3001968.
[49] S. Holm, A simple sequentially rejective multiple test procedure, Scand. J. Stat. 6 (2) (1979) 65–70, https://www.jstor.org/stable/4615733.
[50] D. Lakens, Equivalence tests: a practical primer for t tests, correlations, and meta-analyses, Soc. Psychol. Personal. Sci. 8 (4) (2017) 355–362, https://doi.org/10.1177/1948550617697177.
[51] B.L. Welch, The generalization of 'Student's' problem when several different population variances are involved, Biometrika 34 (1947) 28–35, https://doi.org/10.1093/biomet/34.1-2.28.
[52] N. Ma, X. Zhang, H.-T. Zheng, J. Sun, ShuffleNet V2: practical guidelines for efficient CNN architecture design, in: Proc. ECCV, 2018, pp. 122–138, https://doi.org/10.1007/978-3-030-01264-9_8.
[53] S. Ali, D. Jha, N. Ghatwary, S. Realdon, R. Cannizzaro, O.E. Salem, D. Lamarque, C. Daul, M.A. Riegler, K.V. Anonsen, et al., A multi-centre polyp detection and segmentation dataset for generalisability assessment, Sci. Data 10 (2023) 75, https://doi.org/10.1038/s41597-023-01981-y.
[54] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: visual explanations from deep networks via gradient-based localization, in: Proc. IEEE ICCV, 2017, pp. 618–626, https://doi.org/10.1109/ICCV.2017.74.
[55] R. Wightman, H. Touvron, H. Jégou, ResNet strikes back: an improved training procedure in timm, arXiv preprint arXiv:2110.00476 (2021), https://doi.org/10.48550/arXiv.2110.00476.

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7615100