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.

AKGNet: A Spatially Adaptive Kernel Gating Network for Parameter- Efficient Histopathological Image Classification

Chandrapal Singh
DeepCognix AI Labs
Sumit Gupta
sumit@deepcognix.com

DeepCognix AI Labs

Research Article

Keywords: Histopathological image classification, adaptive kernel gating, dynamic convolution, lightweight deep learning, parameter efficiency, LC25000, BreakHis, patient-disjoint evaluation

Posted Date: November 27th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11554436/v1

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

Additional Declarations: No competing interests reported.

AKGNet: A Spatially Adaptive Kernel Gating Network for Parameter-Efficient Histopathological Image Classification

Chandrapal Singh [1], Sumit Gupta [2,*]

[1]Research Division, DeepCognix AI Labs, Bengaluru 560058, Karnataka, India.

Email: chandrapal1990@gmail.com

[2]Director, Research Division, DeepCognix AI Labs, Bengaluru 560058, Karnataka, India.

ORCID: 0009-0007-4766-3468. Email: sumit@deepcognix.com

[*]Corresponding Author: Sumit Gupta (sumit@deepcognix.com)

Abstract

Compact convolutional networks are attractive for digital-pathology deployment, but fixed-kernel designs apply the same receptive field at every location of tissue images whose diagnostic content varies in scale. This paper presents AKGNet, a 0.28 M-parameter network whose Smoothed Adaptive Kernel Gating (S-AKG) blocks weight 1×1 to 7×7 convolutional experts with spatially varying, temperature-scaled and smoothed routing maps, combined with a group-normalized stem, a background-suppressing spatial attention module and dual global pooling. Under a unified protocol with stain-aware augmentation and single-use test sets, AKGNet attains $97.02 \pm 0.25\%$ accuracy on LC25000 over three seeds, above a size-matched conventional CNN with 40% more parameters (94.84%) and below ImageNet-pretrained ResNet-50, EfficientNet-B4 and ViT-Base, which reach 99.9–100%. On BreakHis under a patient-disjoint split, AKGNet reaches 82.0% accuracy and 78.3% balanced accuracy, but its ROC-AUC (0.743) is below that of the size-matched CNN (0.798) and well below those of EfficientNet-B4 (0.892) and ResNet-50 (0.913); with about twelve test patients in a single split, these BreakHis comparisons are indicative only. The efficiency of AKGNet lies in parameters and memory rather than arithmetic: all experts are evaluated densely, so the model requires 1.84 GFLOPs, more than the size-matched CNN, while running 3–10× faster than the large baselines on a GPU with a 1.07 MB footprint and the lowest peak inference memory of the models tested. Single-seed ablations indicate that learned routing adds 1.36 points over uniform expert averaging on LC25000, whereas gate smoothing has no detectable effect. Because LC25000 was generated by augmenting 1,250 source images without source identifiers, a perceptual-hash analysis of its duplicate structure is also reported; it detects many near-duplicate pairs across random splits but, being insensitive to arbitrary-angle rotation, it does not bound the resulting accuracy inflation.

Keywords

Histopathological image classification; adaptive kernel gating; dynamic convolution; lightweight deep learning; parameter efficiency; LC25000; BreakHis; patient-disjoint evaluation

1. Introduction

Automated cancer diagnosis has been transformed by the advent of digital histopathology, which allows tissue morphology to be analyzed computationally at high spatial resolution. Lung, colorectal and breast carcinomas remain among the leading causes of cancer mortality worldwide [1], and histopathological examination of hematoxylin and eosin (H&E) stained tissue is the primary diagnostic method for all three; public benchmarks such as the LC25000 lung-and-colon collection [2] and the BreakHis breast collection [3] have provided the means for systematic evaluation of deep learning methods for histopathological image

classification. The classification of histopathology images, however, differs from conventional image classification in a fundamental respect: the visual evidence necessary for diagnosis is inherently heterogeneous in spatial scale. Diagnostically relevant information spans from nuclear morphology and cellular atypia at fine spatial scales, through local cellular and glandular organization, to tissue architecture and invasive growth patterns extending over much larger regions. A diagnostically important pattern may therefore be localized to a small group of hyperchromatic nuclei in one image but depend on broader architectural context in another. This multi-scale characteristic is well established: Schmitz et al. [4] showed the significance of merging information ranging from nuclear-scale aberrations to the global tissue architecture, Hashimoto et al. [5] revealed that multi-scale information helps classify cancer subtypes from unannotated slides, and Li et al. [6] found multi-resolution processing beneficial for histopathology image classification and localization. These findings suggest that no universal spatial scale is optimal across histopathological patterns, and that architectures able to resolve fine-grained cellular detail must also provide larger-scale tissue context.

The challenge is therefore not simply to extract features but to determine how much spatial context should inform the representation of a particular region. Conventional convolutional neural networks (CNNs) offer only a limited solution because their spatial operators are fixed: the same kernel configuration is applied at every spatial location, so a 3×3 kernel performs the same local aggregation whether the underlying region contains isolated nuclei, densely packed cells, glandular structures or larger tissue organization. The effective receptive field can be expanded by enlarging kernels, stacking layers or applying dilated convolution, but this comes at the cost of higher computational demands or an imposed fixed spatial structure; more importantly, none of these strategies explicitly specifies which spatial scale is most informative in a given region.

To overcome this limitation, multi-scale architectures have been developed that process multiple receptive fields or magnification levels in parallel and fuse the results [4]–[6]. These methods add complementary cellular- and tissue-level information, but every input is processed at every scale with fixed, input-independent fusion weights, irrespective of whether a scale is locally informative. This is particularly consequential for histopathological images, whose spatial content differs markedly from one region to another, with some regions requiring fine-grained nuclear information and others benefiting from a wider cellular or tissue perspective.

An alternative approach makes the convolutional operation itself input-dependent. Conditionally parameterized convolution (CondConv) showed that different input examples can be served by different sets of learned kernels [7]; dynamic convolution aggregates multiple kernels through input-conditioned attention [8]; and deformable convolution modifies the sampling locations of the kernel depending on the input [9]. These methods establish an important principle: spatial computation need not remain constant across inputs. They do not, however, directly address the scale-selection problem posed by histopathological morphology. Beyond the question of which features to extract, the relevant question is which receptive-field scale should contribute most strongly at each spatial location: fine-scale operations suit nuclear morphology and cellular atypia, intermediate scales suit local cellular organization and glandular structures, and larger receptive fields suit tissue architecture and invasive patterns.

In dynamic convolution the kernel mixture is computed once per input, and deformable convolution mainly alters sampling locations without explicitly selecting among receptive-field sizes; neither formulation therefore provides spatially varying scale selection within the same image. An alternative is to encode several receptive-field sizes as a compact collection of convolutional experts and to adjust the contribution of each expert according to local feature characteristics, so that the weight given to each scale can vary from one location to another instead of all scales being fused with fixed weights.

The need for efficiency further emphasizes this point. Deep residual networks [10], compound-scaled convolutional networks [11] and vision transformers [12] have greatly improved recognition accuracy, but they can be undesirable for resource-limited medical applications and large histopathological collections, where even moderate growth in per-image cost multiplies into significant training and inference expense when thousands of tissue fields must be processed. In such settings model size and memory footprint matter alongside arithmetic cost, while aggressive compression limits the capacity available to represent heterogeneous morphology. The present work targets parameter and memory efficiency; as Section 7 shows, the proposed gating does not reduce arithmetic relative to a plain compact CNN, and this distinction is maintained throughout the paper.

Lightweight medical architectures show that considerable parameter reduction is achievable for image classification [13], [14], but reducing parameters alone does not ensure an efficient representation of the multi-scale structure of tissue. To use the limited capacity of a compact model well, the receptive-field scales most relevant to the local morphology should be emphasized. Histopathological images offer a natural opportunity for such adaptive weighting, since a single field may contain relatively uniform background, individual nuclei, dense cellular regions, gland boundaries and larger architectural patterns side by side.

Another factor is the domain shift between natural-image pretraining and H&E histopathology. ImageNet-pretrained models are commonly applied to medical image classification, yet histopathological images differ significantly from natural photographs in their colour statistics, repetitive cellular textures, tissue organization and spatial distribution of diagnostic structures [15]. An efficient computational pathology model should therefore exploit the structural properties of tissue morphology rather than inherit architectural assumptions designed for natural image recognition. A further consideration concerns evaluation practice [16]. Patient-level partitioning is standard good practice whenever patient identifiers are available, and the original BreakHis protocol used patient-wise partitions [3], although many later BreakHis studies report image-level splits. LC25000 poses a more specific problem: the public release was generated by augmenting 1,250 source images and provides no source identifiers [2], so random image-level partitions can place augmented siblings of a test image in the training data [17]. Section 5 therefore reports a patient-disjoint BreakHis evaluation and a duplicate analysis of LC25000; both serve to qualify the architectural results rather than as contributions in their own right.

Collectively, these observations reveal a specific research gap. Existing histopathology studies emphasize the need for multi-scale information [4]–[6]; conditional and dynamic convolution show that feature extraction can be tailored to the input [7], [8]; deformable convolution adds spatial flexibility [9]; lightweight architectures address model size [13], [14]; and domain-specific studies caution against directly transferring natural-image representations to histopathology [15]. None of these directions, however, explicitly considers spatially adaptive receptive-field weighting inside a compact histopathology classifier, that is, a lightweight mechanism that decides, at each spatial location, how strongly each receptive-field scale should contribute, instead of fusing parallel scales with fixed weights. This paper addresses that gap with AKGNet, a compact Adaptive Kernel Gating Network for histopathological image classification. Instead of a fixed kernel configuration, AKGNet represents four spatial scales as convolutional experts with 1×1, 3×3, 5×5 and 7×7 kernels inside a 32-channel backbone, and a lightweight spatial router generates spatially varying soft gating maps that determine the relative contribution of each expert at every location of the feature map. The routing distribution is temperature-scaled ($\tau = 2.0$) and spatially smoothed by 3×3 average pooling, two implementation choices intended to favour gradual transitions between neighbouring regions. A group-normalized 7×7 stem avoids batch-coupled statistics under stain variability, a spatial attention module computed from channel-wise average and maximum projections reweights locations before scale allocation, and dual global pooling preserves both distributed tissue context and focal high responses. The design separates where to attend from which scale to use: spatial relevance is established by the attention module

before the gating blocks allocate receptive fields. The complete model comprises 279,555 parameters with the five-class head (1.07 MB). Because all experts are evaluated at every location, the design saves parameters and memory but not arithmetic, a distinction that Section 7 quantifies.

The principal contributions of this work are as follows.

(i) Compact adaptive architecture: a 32-channel, 0.28 M-parameter network for histopathological classification with spatially adaptive receptive-field weighting.

(ii) Spatially adaptive receptive-field selection: an adaptive kernel-gating block whose router produces spatially varying weights over 1×1 to 7×7 convolutional experts, so different regions of one image receive different scale mixtures.

(iii) Temperature-controlled and smoothed routing: temperature-scaled softmax gating with average-pool smoothing of the routing maps, included as implementation choices intended to favour spatially coherent expert weights; their measured effect on accuracy is reported rather than assumed.

(iv) Attention-guided adaptive processing: a 98-parameter spatial attention module allows the network to focus on informative tissue regions before receptive-field scales are allocated.

(v) Parameter- and memory-efficient multi-scale representation: a compact set of spatial experts is weighted per location rather than fused with fixed weights. Because the experts are evaluated densely, the benefit lies in parameters and memory rather than FLOPs, and measured latency, memory and FLOPs are reported against large pretrained baselines and a size-matched CNN.

(vi) Evaluation practice: a patient-disjoint BreakHis evaluation, a perceptual-hash duplicate analysis of LC25000 with its limitations stated, single-use test sets and checkpoint selection restricted to validation macro-F1.

(vii) Empirical analysis: accuracy with paired statistics and their limitations, component ablations, calibration, stain and corruption robustness including failure cases, Grad-CAM interpretation, and measured latency, memory and FLOPs against ImageNet-pretrained baselines and a size-matched conventional CNN.

2. Related Work

2.1 Multi-scale analysis in computational pathology

The multi-scale character of tissue morphology has motivated a family of architectures that combine information across resolutions. Schmitz et al. [4] designed multi-scale fully convolutional networks whose encoders ingest nuclear-scale and architecture-scale context jointly and showed consistent gains over single-scale processing for histopathological segmentation. Hashimoto et al. [5] combined multiple magnifications within a domain-adversarial multiple-instance framework for cancer-subtype classification of whole slides, and Li et al. [6] processed high- and low-resolution views in parallel for classification and localization under weak supervision. These designs validate multi-scale evidence but share a structural property: every input is processed by every scale pathway, and fusion occurs with fixed, input-independent structure. AKGNet does not remove this cost, since its experts are also evaluated at every location; its difference is that the fusion weights are learned and location-dependent within a small parameter budget.

2.2 Deep learning on LC25000 and BreakHis

LC25000 [2] has become a standard benchmark for lung and colon histopathology. Masud et al. [18] combined spatial- and frequency-domain features in a compact four-layer CNN and reported 96.33% five-

class accuracy, an early demonstration that small networks are viable on this dataset. Attallah et al. [19] processed the images with three lightweight CNNs, compressed the resulting features with principal component analysis and the fast Walsh–Hadamard transform, and classified with conventional learners, reaching 99.6% at low computational cost. Subsequent work has largely followed the scale-up route: Said et al. [20] designed a dedicated deep architecture, Ji et al. [21] applied Swin Transformer V2 under five-fold cross-validation, and a 2025 study fine-tuned EfficientNet-B3 to 99.39% [22]. These accuracies concentrate near or above 99% and were obtained with random image-level partitions. Because LC25000 was generated from 1,250 source images by rotation and flipping, such partitions can place augmented copies of the same tile in training and test data. Batchkala et al. [17] grouped the 25,000 images back to their source tiles using histopathology foundation-model embeddings with manual correction and released the grouping (LC25000-clean) as a leakage-free benchmark; their work is the appropriate reference for source-level evaluation on this dataset. On BreakHis [3], attention-augmented backbones dominate recent reports: the multi-branch spectral channel attention network of Cao et al. [23] enriches a ResNet-18 backbone with frequency-domain channel attention, and high-order multi-branch receptive-field modelling [24] reports up to 99.2% "patient-level" accuracy. In much of the BreakHis literature this term denotes per-patient aggregation of predictions computed on image-level splits; it does not imply that patients are disjoint between training and test partitions. The original BreakHis protocol [3], by contrast, used five random patient-wise 70/30 partitions, so patient separation is the reference practice for this dataset, and a recent patient-aware benchmarking study [25] shows that strict patient separation substantially reduces the accuracy of nine standard architectures.

2.3 Dynamic convolution and conditional computation

Input-conditioned convolution has developed along several lines. Selective kernel networks [26] fuse parallel branches of different kernel size through channel attention, adapting the receptive field per channel but uniformly over space. CondConv [7] and dynamic convolution [8] mix multiple kernel banks with example-dependent coefficients, but with one mixture per image. Deformable convolution [9] adapts sampling geometry rather than kernel scale. More recent work addresses the parameter and latency overheads of kernel mixing: KernelWarehouse [27] re-parameterizes kernel banks for efficiency, and FMDConv [28] optimizes the speed-accuracy trade-off of multi-attention dynamic convolution; sparse expert mixtures in the broader mixture-of-experts literature are surveyed in [29]. The S-AKG block differs from these methods in that routing is computed per spatial location, so a single field can emphasize nuclear-scale kernels in one region and context-scale kernels in another, and in that the routing field is temperature-scaled and smoothed. An important practical difference follows: CondConv and dynamic convolution aggregate kernels before a single convolution, so their inference cost stays close to that of one kernel, whereas S-AKG mixes the outputs of all experts at each location and therefore pays the cost of every expert (Section 7). A direct comparison with a per-image dynamic convolution under the same parameter budget was not part of the present experiments; without it, the gains reported below cannot be attributed specifically to spatial, as opposed to per-image, gating (Section 8). To our knowledge, spatially smoothed per-location kernel routing has not previously been evaluated for histopathological classification.

2.4 Architectural components and evaluation methodology

The remaining components build on established primitives. Group Normalization [30] provides batch-independent statistics, which is appropriate when staining variability makes cross-sample batch statistics unstable; the attention module follows the channel-pooling formulation popularized by CBAM [31]; Grad-CAM [32] supports qualitative analysis; and optimisation uses decoupled weight decay [33]. Stain-aware colour augmentation follows the evidence of Tellez et al. [15] that H&E channel perturbation improves generalisation in computational pathology. The evaluation draws on the taxonomy of leakage in machine-learning-based science [16], uses exact McNemar tests with Holm correction [34], and uses the NCT-CRC colorectal collection [35] for a supplementary cross-dataset experiment. Compact medical classifiers such as

attention-augmented lightweight CNNs [14] and robust medical vision transformers [13] define the efficiency regime within which the proposed model is placed.

3. Proposed Methodology

Histopathological image classification is challenged by substantial cellular heterogeneity, variation in tumor scale, stain-dependent appearance and the presence of non-diagnostic background regions; these factors make fixed-receptive-field feature extraction insufficient for reliably modelling both fine-grained cellular structure and larger tissue-level patterns simultaneously. AKGNet addresses this by adapting its effective receptive field to the spatial characteristics of the input tissue. The network is the composition of four stages, a contrast-preserving stem S, a background-suppressing spatial attention module A, three Smoothed Adaptive Kernel Gating blocks g_1, g_2, g_3 and a classification head h,

$$F(X) = h(g_3(g_2(g_1(A(S(X)))))) \tag{1}$$

where $X \in \mathbb{R}^{N\times3\times224\times224}$ is a mini-batch of N RGB images and N, C, H, W denote batch, channel, height and width dimensions throughout. The stem extracts low-level colour and texture features while avoiding batch-dependent normalization statistics; the attention module assigns spatially varying importance, attenuating responses from non-diagnostic regions before further processing; the S-AKG cascade combines parallel convolutional kernels of different receptive fields through spatially adaptive gating whose distributions are temperature-controlled and spatially smoothed; and the head condenses the final feature map into a compact global descriptor. Through this hierarchical design the network is intended to model local cellular morphology and broader tissue organization jointly. Fig. 1 shows the overall architecture using the notation of Eqs. (1)–(19). In Fig. 1, B denotes the batch size, written N in the text.

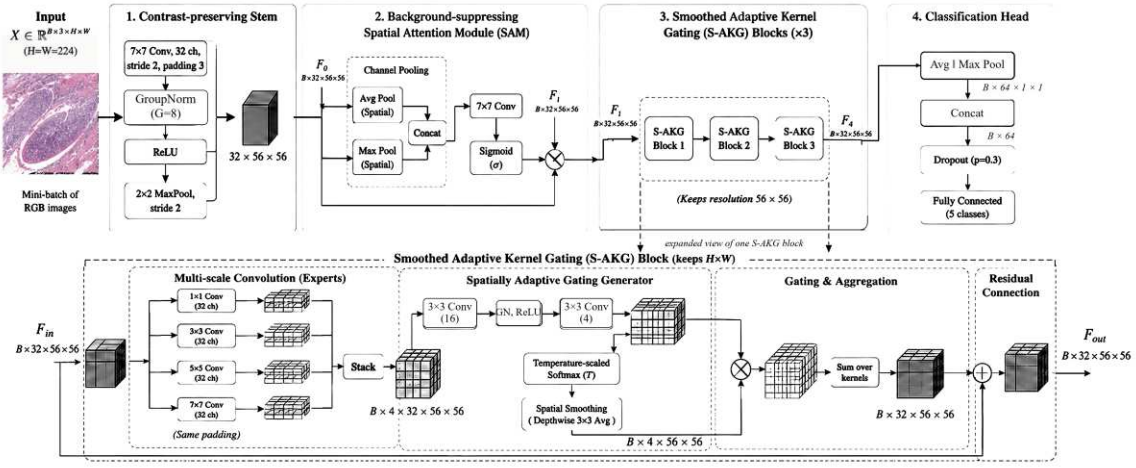


Fig. 1. Overall architecture of AKGNet

3.1 Contrast-preserving stem

The stem transforms the input into a compact low-level representation while retaining diagnostically relevant colour, texture and cellular-structure information. In H&E-stained tissue, subtle variations in nuclear intensity, cytoplasmic staining and local morphology carry cues that separate normal tissue from adenocarcinoma and squamous cell carcinoma, and excessive normalization at the first stage risks weakening exactly these stain and contrast characteristics. A 7×7 convolution with $C = 32$ output channels, stride 2 and zero-padding 3 is applied first,

$$Y = W_7 *_2 X + b,\ Y \in \mathbb{R}^{N\times 32\times 112\times 112} \tag{2}$$

where W_7 is the filter bank and $*_2$ denotes stride-2 convolution; the large first-layer kernel provides an initial receptive field spanning several nuclei at the 224×224 input resolution. The choice of normalization is deliberate. Batch Normalization estimates statistics across the mini-batch, which becomes unreliable at small batch sizes or under strong inter-slide staining variability because samples with different stain characteristics then share statistics; Instance Normalization removes per-channel, per-image statistics entirely and can discard instance-specific contrast that is informative in H&E imagery. Group Normalization [30] offers a compromise, normalizing fixed groups of channels independently for each sample. With $G = 8$ groups over $C = 32$ channels ($C/G = 4$ channels per group), for sample n and group g with channel set $\mathcal{C}_g$,

$$\mu_{n,g} = \frac{1}{4HW} \sum_{c\in\mathcal{C}_g} \sum_{h=1}^{H} \sum_{w=1}^{W} Y_{n,c,h,w} \tag{3}$$

$$\sigma^2_{n,g} = \frac{1}{4HW} \sum_{c\in\mathcal{C}_g} \sum_{h=1}^{H} \sum_{w=1}^{W} (Y_{n,c,h,w} - \mu_{n,g})^2 \tag{4}$$

$$\tilde{Y}_{n,c,h,w} = \gamma_c \frac{Y_{n,c,h,w} - \mu_{n,g(c)}}{\sqrt{\sigma^2_{n,g(c)} + \varepsilon}} + \beta_c \tag{5}$$

where $g(c)$ is the group of channel c, ε ensures numerical stability, and γ_c, β_c are learnable per-channel scale and shift parameters that allow the network to restore or amplify response magnitudes useful for classification after normalization. Rectification and 2×2 max-pooling with stride 2 complete the stem,

$$F_0 = \mathrm{MaxPool}_{2\times 2}(max(0, \tilde{Y})),\ F_0 \in \mathbb{R}^{N\times 32\times 56\times 56} \tag{6}$$

The rectification retains positive responses associated with tissue boundaries, nuclei and glandular structures, and max-pooling propagates the locally strongest of these responses while halving the resolution once more. The resulting 32×56×56 representation is passed to the attention module for spatial reweighting. Fig. 2 shows the stem.

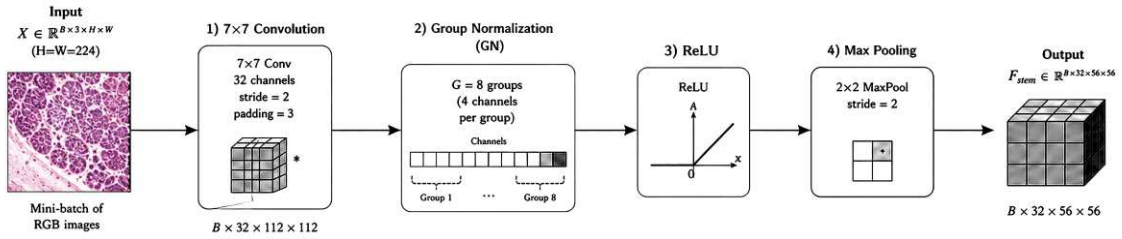


Fig. 2. Contrast-preserving stem

3.2 Background-suppressing spatial attention module

Histopathological fields commonly include regions of limited diagnostic content, such as slide background, blank glass, adipose tissue, staining artefacts and imaging noise, which can occupy a considerable portion of the input and introduce spurious activations during feature extraction. The attention module, placed immediately after the stem, computes two single-channel spatial descriptors from F_0 by aggregation over the channel dimension, following the channel-pooling formulation of [31],

$$M_{\mathrm{avg}}(h,w) = \frac{1}{32}\sum_{c=1}^{32} F_0(c,h,w),\ M_{\max}(h,w) = max_c F_0(c,h,w) \tag{7}$$

The average projection captures the mean activation consensus across channels and thus measures overall feature presence at each location, whereas the maximum projection retains the single strongest channel response and is sensitive to compact salient structures such as hyperchromatic nuclei, glandular boundaries, keratinized regions and irregular tumor morphology. Their concatenation is mapped to an attention field by a single bias-free 7×7 convolution, chosen to model local and moderately broad spatial dependencies, followed by a logistic function,

$$A = \sigma(W_s * [M_{\text{avg}}; M_{\max}]),\ F_A = A \odot F_0 \tag{8}$$

where $W_s \in \mathbb{R}^{1\times2\times7\times7}$, $A \in (0,1)^{N\times1\times56\times56}$ and $\odot$ denotes broadcast element-wise multiplication, the attention coefficient being shared across all channels at a given location. Locations with weak or inconsistent channel statistics receive small coefficients and their contribution to subsequent processing is attenuated, whereas regions producing strong, spatially coherent responses are preserved. Because the mechanism performs soft reweighting rather than hard masking, diagnostically useful information near tissue boundaries or in partially degraded regions remains available downstream. Combining channel-statistical compression with a large spatial kernel makes the module computationally negligible, 98 parameters in total, while providing an explicit mechanism for suppressing non-diagnostic regions before receptive-field allocation. Fig. 3 illustrates the structure of the module.

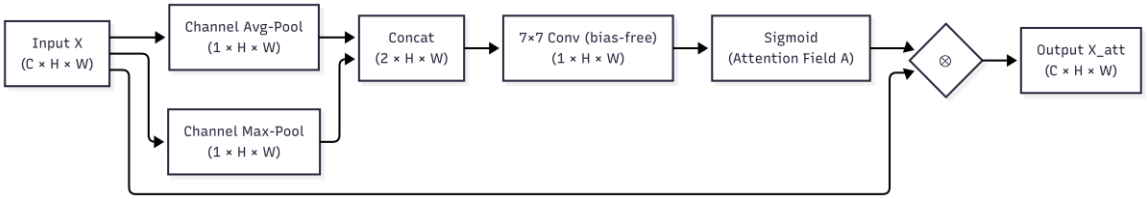


Fig. 3. Structure of the background-suppressing spatial attention module

3.3 Smoothed adaptive kernel gating block

The S-AKG block (Fig. 4) is the principal multi-scale component. Fine nuclear patterns and irregular glandular boundaries require small receptive fields, whereas larger tumor regions and stromal organization benefit from broader contextual aggregation, so a single fixed kernel is unlikely to be optimal across all locations. The block input $F_{\text{in}} \in \mathbb{R}^{N\times32\times56\times56}$ is processed by $E = 4$ parallel convolutional experts with kernel sizes $k_e \in \{1,3,5,7\}$,

$$E_e = W^{(e)} * F_{\text{in}},\ e = 1,\dots,4 \tag{9}$$

each with stride one and zero-padding $\lfloor k_e/2 \rfloor$ so that all outputs are spatially aligned. The 1×1 expert performs channel mixing with minimal spatial aggregation, while the 3×3, 5×5 and 7×7 experts incorporate progressively broader context; together they provide a basis of receptive fields ranging from localized cellular patterns to tissue-level structure. The mixture applied at each location is predicted by a lightweight spatial router: the input is reduced from $C = 32$ to $C/2 = 16$ channels by a 3×3 convolution, normalized with four-group GN, rectified, and projected to four routing channels by a second 3×3 convolution,

$$Z = W_{r2} * \text{ReLU}(\text{GN}_4(W_{r1} * F_{\text{in}})),\ Z \in \mathbb{R}^{N\times4\times56\times56} \tag{10}$$

At each location (h, w) the routing logits z_e are converted into normalized gating coefficients by a temperature-scaled softmax,

$$g_e(h,w) = \frac{\exp(z_e(h,w)/\tau)}{\sum_{e'=1}^{4}\exp(z_{e'}(h,w)/\tau)},\ \tau = 2.0 \tag{11}$$

so that $g_e(h,w) \geq 0$ and $\sum_e g_e(h,w) = 1$. The temperature governs the sharpness of the routing distribution: a low temperature produces highly selective assignments, whereas a larger temperature distributes weight more smoothly across experts. Formally, the gate sensitivity scales inversely with temperature,

$$\frac{\partial g_e}{\partial z_{e'}} = \frac{1}{\tau} g_e(\delta_{ee'} - g_{e'}) \tag{12}$$

so the setting $\tau = 2.0$, which was fixed a priori as an implementation choice rather than tuned, is intended to moderate abrupt spatial changes in expert assignment; its effect was not ablated separately. Temperature scaling alone does not remove high-frequency variation in the routing maps. Each gate map is therefore additionally filtered with a 3×3 average-pooling operator with stride one and padding one,

$$\hat{g}_e = \mathrm{AvgPool}_{3\times3}(g_e) \tag{13}$$

which acts as a low-pass filter on the routing maps; because averaging is linear and follows the softmax, the smoothed coefficients remain exactly normalized at interior locations. Neighbouring pixels are thereby encouraged to exhibit similar expert preferences, which is intended to suit spatially continuous structures such as tumour regions, glandular formations and stromal compartments. As Section 6 shows, the accuracy effect of this smoothing was not detectable on LC25000, so it is described here as an implementation choice rather than a validated component. The experts are combined by the smoothed mixture, normalized, and added residually,

$$U(h,w) = \sum_{e=1}^{4} \hat{g}_e(h,w) \cdot E_e(h,w) \tag{14}$$

$$g(F_{\mathrm{in}}) = \mathrm{ReLU}(F_{\mathrm{in}} + \mathrm{BN}(U)) \tag{15}$$

The residual pathway preserves the incoming representation and facilitates gradient propagation, while the normalized adaptive branch introduces the scale-dependent information selected by the router; Batch Normalization stabilizes the mixed expert distribution and does not replace the contrast-preserving strategy of the stem. Three blocks are cascaded at constant resolution,

$$F_1 = g_1(F_A),\ F_2 = g_2(F_1),\ F_3 = g_3(F_2) \tag{16}$$

As a consequence, the effective receptive field can vary across the feature map: regions dominated by fine cellular structure can emphasize the smaller kernels while regions requiring broader context assign greater weight to the larger ones. The design differs from conventional multi-scale feature fusion in two respects: candidate receptive fields are not concatenated or averaged with fixed coefficients but weighted by coefficients predicted from the input, and the weighting is spatially varying, so different regions within one image can receive different scale mixtures. The architecture furthermore separates where to attend from which scale to use: spatial relevance is established by the attention module before the gating blocks allocate receptive fields. Whether the trained routers actually exploit this spatial freedom is an empirical question that the present experiments do not answer (Section 5.4). The computational consequence of Eqs. (9) and (14) must also be stated plainly: all four experts are evaluated at every location before reweighting, so the gating changes how the scales are combined, not how much is computed. The expert convolutions account for about 88% of the 0.92 G multiply–accumulate operations of the network (the 7×7 experts alone for about 51%), and the block therefore offers parameter and memory economy rather than computational savings. Converting the gating into savings would require sparse or top-k evaluation of experts, which is not implemented here.

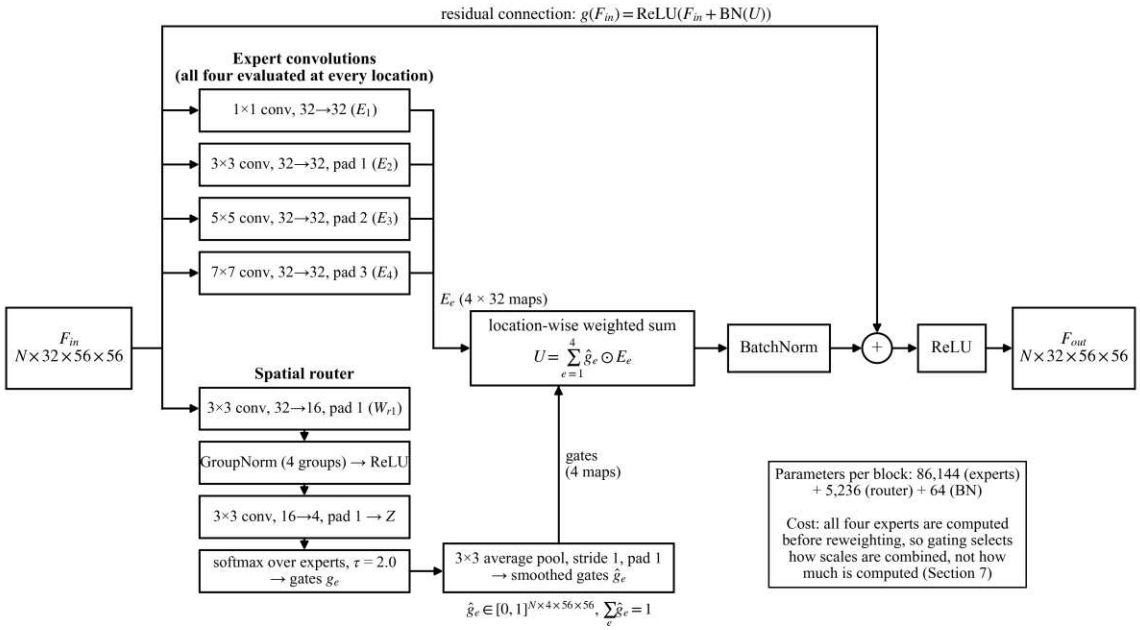


Fig. 4. Structure of the S-AKG block as implemented

3.4 Global descriptor and classification head

Diagnostic evidence resides both in the overall activation distribution, which reflects tissue-level organization, and in small, strongly responding foci such as isolated malignant structures. Global average pooling summarizes the distributed context but can dilute small regions; global max pooling preserves the strongest localized response per channel and thereby retains evidence from compact foci. The head concatenates both over F_3,

$$v = [\frac{1}{56 \cdot 56} \sum_{h,w} F_3\,(c,h,w)\ \ ;\ \ max_{h,w} F_3(c,h,w)] \in \mathbb{R}^{64} \tag{17}$$

applies dropout with probability 0.3 during training, and maps the descriptor to K class logits with a single linear layer ($K = 5$ for LC25000, $K = 2$ for BreakHis),

$$z = W_{\mathrm{cls}} \cdot \mathrm{Dropout}_{0.3}(v) + b_{\mathrm{cls}},\ p_k = \frac{\exp(z_k)}{\sum_{k'} \exp(z_{k'})},\ \hat{y} = \backslash \mathrm{arg} max_k p_k \tag{18}$$

Training minimizes the categorical cross-entropy

$$\mathcal{L} = -\sum_{k=1}^{K} y_k \log p_k \tag{19}$$

The combination of the two pooling paths provides a descriptor that reflects both distributed tissue context and focal high-response regions. The complete model comprises 279,555 parameters with the five-class LC25000 head and 279,360 with the two-class BreakHis head (1.07 MB in 32-bit precision): 258,432 in the expert convolutions (86,144 per block), 15,708 in the three routers, 192 in the block normalization layers, 4,800 in the stem, 98 in the attention module and 325 in the five-class head. The figure of 279,395 that appears in the ablation table is the parameter count of the variant without the max-pooling branch, whose head receives a 32-dimensional descriptor.

4. Experimental Setup

All models were trained under one protocol: AdamW [33] with learning rate 1e-4 and weight decay 1e-4, cosine-annealing schedule, 20 epochs, effective batch size 32 (preserved through gradient accumulation where memory required smaller physical batches), 224×224 inputs and mixed precision, on an NVIDIA GeForce RTX 2060 (6 GB) with PyTorch 2.4.1. Augmentation comprised horizontal flips, $\pm 10^{\circ}$ rotations, HED-space stain jitter ($\sigma = 0.05, p = 0.5$) and colour jitter, applied identically to every model [15]. The same training hyperparameters were used for every model and were not tuned per model; the architectural settings of AKGNet (channel width, expert kernel sizes, temperature and smoothing window) were fixed as implementation choices before any test set was used, and no test-set result informed a design or hyperparameter decision. Checkpoints were selected exclusively by validation macro-F1, and each test set was evaluated once per run. AKGNet, the size-matched Basic CNN control (0.39 M parameters; four stages of 3×3 convolution, batch normalization, ReLU and 2×2 max-pooling with 32, 64, 128 and 256 channels, followed by global average pooling and a linear layer) and all ablation variants were trained from random initialisation; ResNet-50 [10], EfficientNet-B4 [11] and ViT-Base [12] used ImageNet weights, their practical deployment mode, and ResNet-50 was additionally retrained from scratch to probe this asymmetry. On LC25000, AKGNet was trained with three seeds (42, 43 and 44) and the baselines with one or two seeds (Table 2); on BreakHis all models, and on LC25000 all ablation variants, were trained once with seed 42. Metrics include accuracy, balanced accuracy, macro-F1, per-class precision, recall and F1, ROC-AUC, expected calibration error and Brier score. Uncertainty is reported as the seed-to-seed standard deviation where several seeds exist (single runs are reported without one) and as 2,000-resample percentile bootstrap 95% confidence intervals over test images. Pairwise comparisons use exact McNemar tests with Holm correction [34]. These image-level procedures assume independent test images, an assumption weakened by augmented siblings in LC25000 and violated by within-patient correlation in BreakHis; the resulting intervals and p-values are therefore optimistic, and for BreakHis they are reported descriptively only (Section 5.3). Efficiency is measured as parameters, FLOPs (two operations per multiply–accumulate at 224×224), idle-GPU synchronized latency at batch size 32 (20 warm-up and 50 timed repetitions) and peak memory.

5. Results and Discussion

5.1 Datasets, partitions and duplicate analysis

LC25000 [2] contains 25,000 H&E images of 768×768 pixels in five balanced classes of 5,000 images each: colon adenocarcinoma, benign colon tissue, lung adenocarcinoma, benign lung tissue and lung squamous cell carcinoma (Fig. 5). All 25,000 file paths and the per-class counts were verified programmatically. The public release was produced from 750 lung and 500 colon source images (250 per class) by augmentation with left and right rotations of up to 25° and flips, about twenty images per source, and no source identifiers were released [2], [17]. To characterize the resulting duplicate structure, every image was fingerprinted with SHA-256 and with a 64-bit perceptual hash computed over the eight dihedral orientations. The analysis found 1,318 byte-identical pairs, 4,840 exact perceptual matches across orientations and 18,481 duplicate or near-duplicate edges in total at a Hamming threshold of 6 of 64 bits; under a random image-level split, 10,779 of these edges cross partition boundaries. Clustering the edge graph with union-find yields 12,594 class-pure groups, about ten times the 1,250 source images. The hash is invariant to the dihedral transformations but not to the arbitrary-angle rotations used to build the dataset, so most augmented siblings are not linked: the grouping under-merges, and undetected siblings very likely remain on both sides of any partition built from it. The analysis therefore establishes that duplicate leakage is present under random splits, but it neither measures its full extent nor provides a leakage-free partition. A rotation-tolerant grouping, such as the manually corrected LC25000-clean assignment of Batchkala et al. [17], would be required for that purpose and was not used in the experiments reported here; the LC25000 results below should accordingly be read as image-level benchmark results. The primary evaluation uses a stratified 70/15/15 split (17,499/3,751/3,750

images; Table 1), identical for every model. A secondary hash-group split was constructed by assigning whole groups to partitions; it contains no detected cross-partition edge but, for the reason above, is not leakage-free.

BreakHis [3] contains 7,909 H&E images (700×460 pixels) acquired at 40×, 100×, 200× and 400× magnification and labelled benign (2,480) or malignant (5,429) (Fig. 6). The original publication describes 82 patients; parsing the patient-identifier field of the filenames in our copy of the dataset yields 81 distinct identifiers, so at least one identifier is shared by images that the original description attributes to different patients. The parsed identifier was used as the grouping unit, which can merge but never split a patient across partitions, so the resulting split remains patient-disjoint. A single patient-disjoint 70/15/15 split was drawn with all magnifications pooled and verified so that no identifier contributes images to more than one partition; the test partition contains 1,341 images (455 benign, 886 malignant) from about twelve patients. Patient separation follows the original BreakHis protocol [3], which used five random patient-wise 70/30 partitions; the use of a single partition here is a limitation discussed in Section 5.3. The 31%/69% class imbalance makes balanced accuracy and macro-F1 the primary metrics on this dataset.

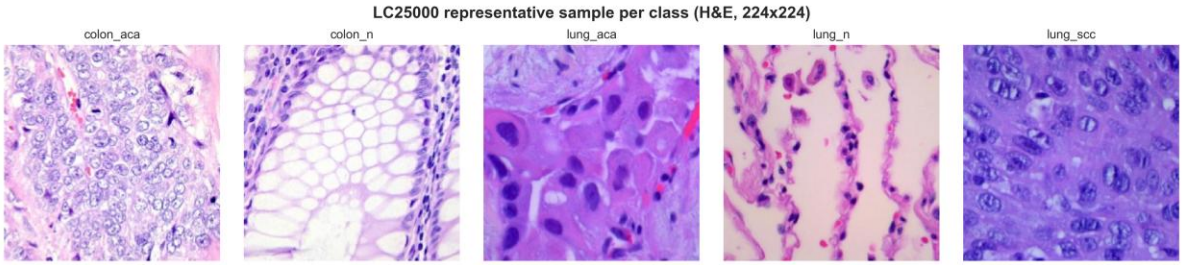


Fig. 5. Representative LC25000 sample per class

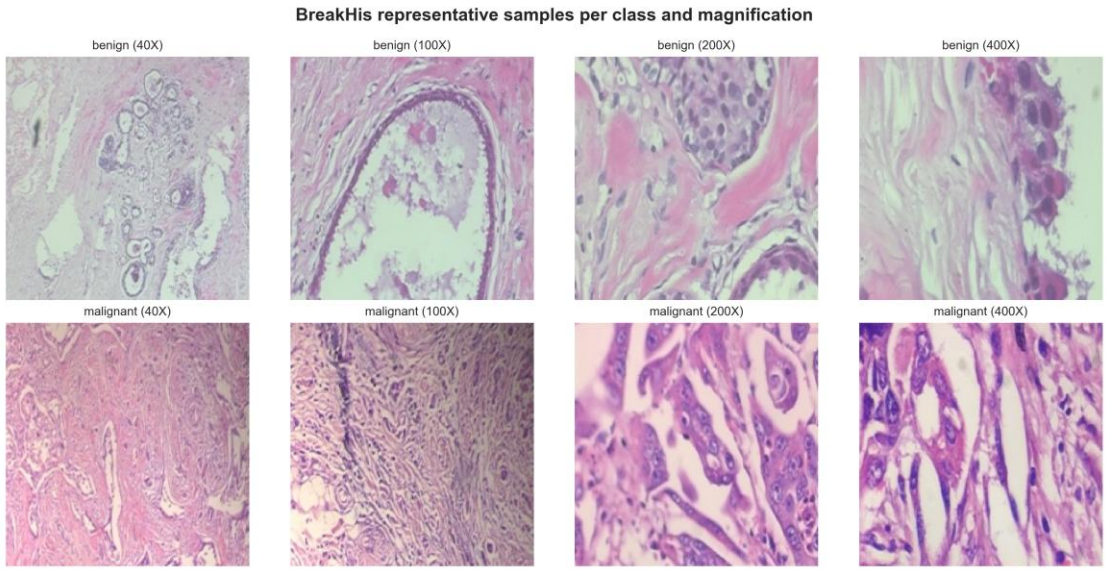


Fig. 6. Representative BreakHis samples per class and magnification

TABLE 1. LC25000 class distribution across the primary split

Class	Total	Train (70%)	Validation (15%)	Test (15%)
colon_aca	5000	3500	750	750
colon_n	5000	3500	750	750
lung_aca	5000	3500	750	750
lung_n	5000	3499	751	750
lung_scc	5000	3500	750	750

5.2 Results on LC25000

AKGNet attains 97.28%, 97.01% and 96.77% test accuracy over seeds 42, 43 and 44, i.e. $97.02 \pm 0.25\%$, with balanced accuracy and macro-F1 equal to accuracy to the third decimal on the balanced classes and one-vs-rest macro ROC-AUC of 0.9986 at seed 42 (Table 2). The training curves (Fig. 8) show stable convergence; validation accuracy was still rising when the 20-epoch budget ended, and the checkpoint rule selected the final epoch in all three seeds. The from-scratch results are therefore specific to this budget and probably underestimate what the architecture can reach; longer schedules were not evaluated, so part of the gap to the pretrained models may be a training-budget effect. Of the 102 errors committed at seed 42 (2.72% of the test set), the large majority lie between lung adenocarcinoma and lung squamous cell carcinoma, the hardest distinction in the dataset, whose resolution can depend on architectural context, keratinization patterns and glandular differentiation beyond the model's receptive field; both benign classes are recognised almost perfectly, with per-class specificity above 0.99 throughout (Fig. 9). The step-wise visualisation in Fig. 10 traces three samples through the pipeline (input, stem activation, spatial attention, Grad-CAM and class probabilities): the stem activation highlights cellular texture, the attention field suppresses empty regions, and Grad-CAM concentrates on nuclei clusters and gland borders.

The ImageNet-pretrained baselines saturate the benchmark: ResNet-50 reaches 100.00% and 99.95% over its two seeds, EfficientNet-B4 99.95% and ViT-Base 99.97% (one seed each), each significantly above AKGNet on Holm-corrected image-level McNemar tests at every seed. Retrained from random initialisation under the identical protocol, ResNet-50 still reaches 99.09%, so on this dataset capacity rather than transfer learning accounts for most of the baseline accuracy. AKGNet outperforms the size-matched Basic CNN control, which has 40% more parameters (0.39 M): 97.28% versus 94.77% at seed 42 (McNemar $p = 7.2 \times 10^{-13}$), and the margin persists across seeds (mean 97.02% versus 94.84%). Within this size class the architecture therefore matters, although the comparison involves a single control design and does not by itself identify which component of AKGNet is responsible (Section 6). Discrimination is near ceiling for all models (macro ROC-AUC of at least 0.996; Table 2), and calibration is good throughout, with AKGNet exhibiting an expected calibration error of 0.011 and the pretrained baselines being slightly over-confident. On the hash-group split, accuracy changes by at most 0.6 points for every model: AKGNet 96.70%, ResNet-50 99.92%, EfficientNet-B4 99.84%, ViT-Base 99.97% and Basic CNN 94.68% (Table 3). Because this split removes only hash-detected duplicates, the small change shows that removing those duplicates matters little; it does not show that the near-saturation of LC25000 is free of duplicate effects, and it does not establish that the accuracies reflect generalisation to unseen source tiles.

TABLE 2. LC25000 test performance on the image-level split

Model	Accuracy	Balanced accuracy	Macro-precision	Macro-recall	Macro-F1	ROC-AUC (OvR)	PR-AUC (OvR)	Acc 95% CI (bootstrap)	Seeds
AKGNet	0.9702 ± 0.0025	0.9702 ± 0.0025	0.9702 ± 0.0025	0.9702 ± 0.0025	0.9702 ± 0.0026	0.9983 ± 0.0002	0.9936 ± 0.0007	[0.9658, 0.9748]	3
ResNet-50	0.9997 ± 0.0004	0.9997 ± 0.0004	0.9997 ± 0.0004	0.9997 ± 0.0004	0.9997 ± 0.0004	1.0000 ± 0.0000	1.0000 ± 0.0000	[0.9993, 1.0000]	2
EfficientNet-B4	0.9995	0.9995	0.9995	0.9995	0.9995	1.0000	1.0000	[0.9987, 1.0000]	1
ViT-Base	0.9997	0.9997	0.9997	0.9997	0.9997	1.0000	1.0000	[0.9992, 1.0000]	1
Basic CNN	0.9484 ± 0.0009	0.9484 ± 0.0009	0.9487 ± 0.0010	0.9484 ± 0.0009	0.9483 ± 0.0009	0.9964 ± 0.0001	0.9864 ± 0.0003	[0.9413, 0.9549]	2

Note: Mean ± SD over seeds where more than one seed was run; single runs are given without SD. Accuracy 95% CIs are percentile bootstrap intervals over test images.

TABLE 3. Image-level versus hash-group LC25000 split (seed 42)

Model	Image-level split acc	Hash-group split acc	Acc change	Image-level macro-F1	Hash-group macro-F1
AKGNet	0.9728	0.9670	0.0058	0.9728	0.9668
ResNet-50	1.0000	0.9992	0.0008	1.0000	0.9992
EfficientNet-B4	0.9995	0.9984	0.0011	0.9995	0.9984
ViT-Base	0.9997	0.9997	0.0000	0.9997	0.9997
Basic CNN	0.9477	0.9468	0.0010	0.9476	0.9464

Note: The hash-group split removes only duplicates detected by dihedral perceptual hashing and is not leakage-free.

Fig. 7 summarizes the two LC25000 duplicate-control comparisons. In the figure, (1) is the image-level split; (2a) and (2b) evaluate the same checkpoints on test images with and without a hash-detected sibling in the training or validation partitions; and (3a) and (3b) are seed-42 models trained on the image-level and hash-group splits. Error bars show the seed-to-seed standard deviation; because hash detection is not rotation-invariant, (2b) and (3b) are not leakage-free. Under the image-level split, 2,502 of the 3,750 test images (66.7%) have at least one detected duplicate or near-duplicate sibling in the training or validation partitions. Evaluating the same checkpoints separately on these images and on the remaining 1,248 gives $97.31 \pm 0.27\%$ and $96.45 \pm 0.38\%$, respectively; the gap is positive at every seed (+0.61, +1.29 and +0.69 points) and for every baseline model. Retraining on the hash-group split gives 96.70% versus 97.28% at seed 42. Both comparisons indicate that detected near-duplicates confer a measurable advantage, but neither yields an upper or a lower bound on the total inflation: the subset without detected siblings and the hash-group split still contain undetected rotated siblings, and the subset comparison is also confounded by any difference in difficulty between images with and without detected siblings. Quantifying the full effect requires a rotation-tolerant source grouping [17]. BreakHis, which concerns patient-level rather than duplicate-level separation, is a different dataset and a different question and is reported separately in Section 5.3.

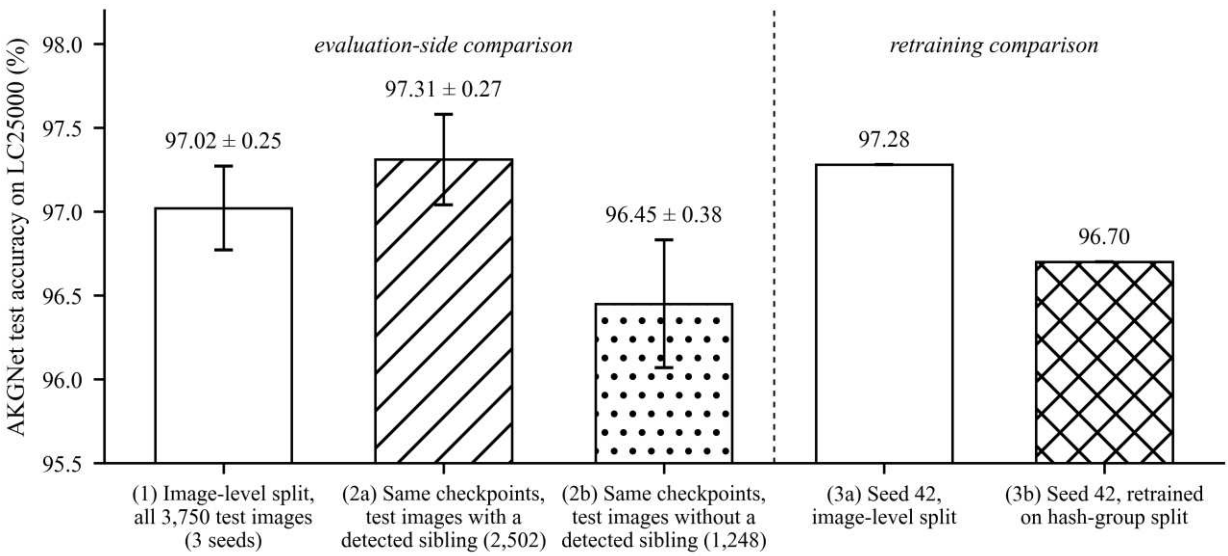


Fig. 7. LC25000 duplicate-control comparisons for AKGNet

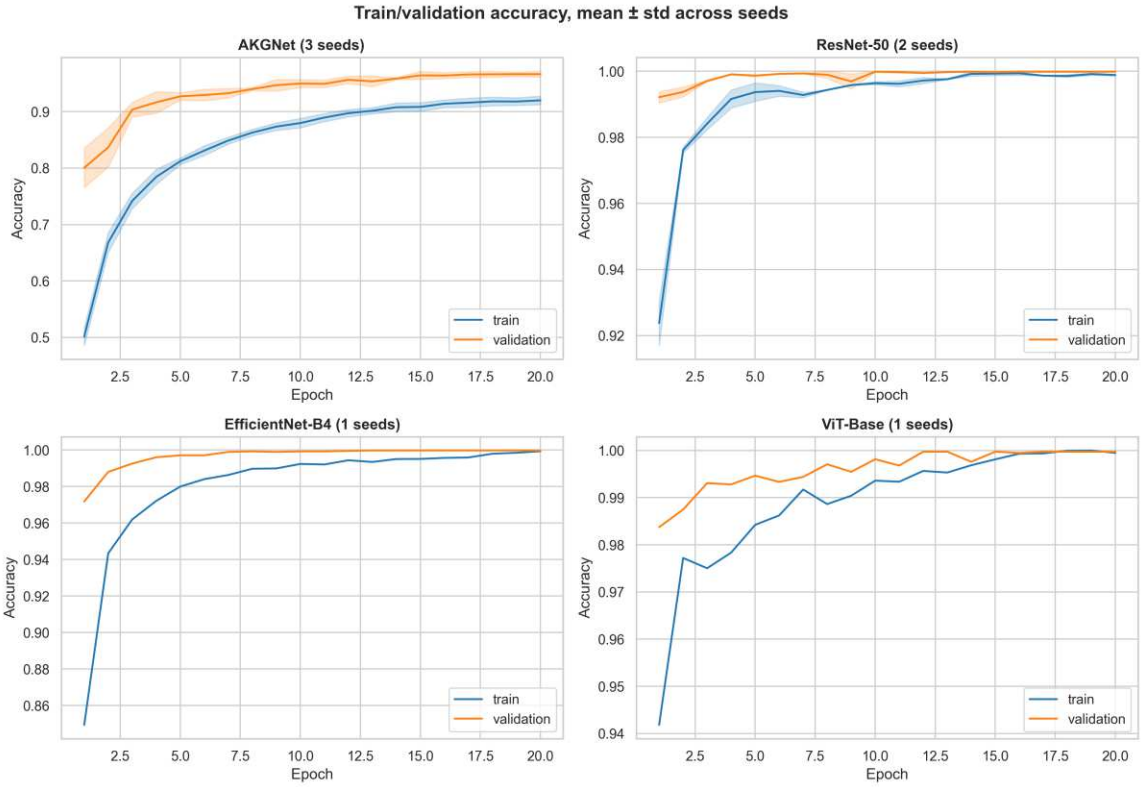


Fig. 8. LC25000 training and validation accuracy (mean ± SD across seeds)

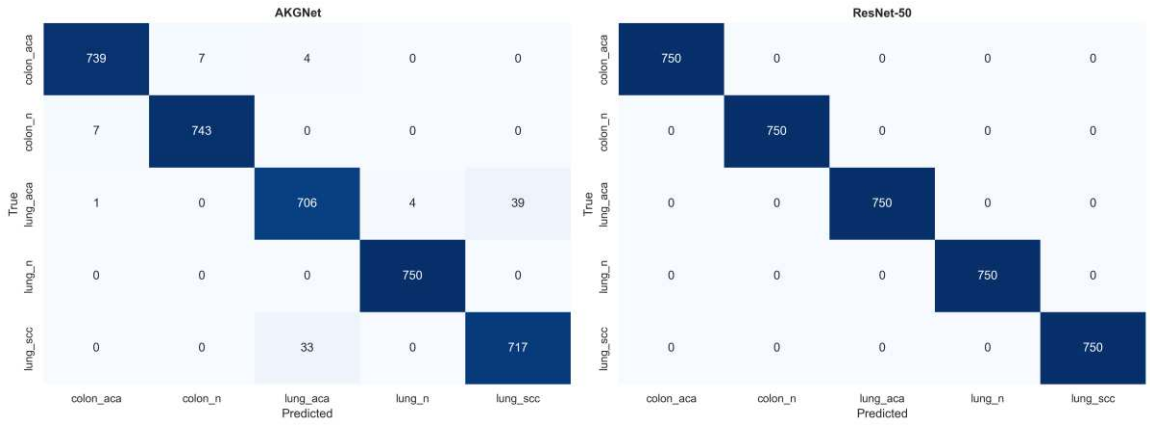


Fig. 9. LC25000 confusion matrices of AKGNet and ResNet-50

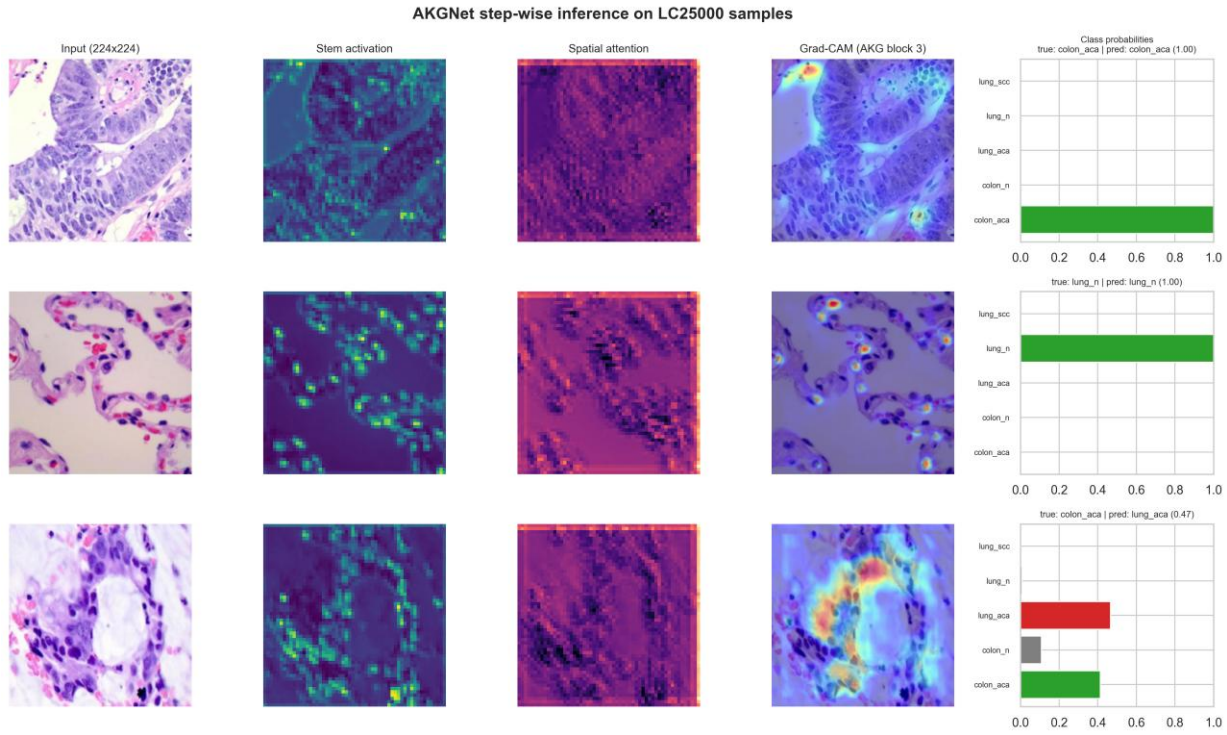


Fig. 10. Step-wise AKGNet inference on LC25000 samples

5.3 Results on BreakHis under patient-disjoint evaluation

The patient-disjoint protocol changes the character of the task: the test patients are unseen individuals, and the evaluation measures cross-patient generalisation rather than within-slide recognition. All models score far below the image-level accuracies common in the BreakHis literature (Table 4). AKGNet reaches 82.03% accuracy, 78.27% balanced accuracy and 0.792 macro-F1, compared with 80.39%, 77.41% and 0.778 for the size-matched Basic CNN, 83.45%, 80.95% and 0.813 for EfficientNet-B4, 84.34%, 81.41% and 0.821 for ViT-Base and 86.88%, 85.31% and 0.854 for ResNet-50. The threshold-free metric does not follow this order: AKGNet has the lowest ROC-AUC of all models (0.743), below the Basic CNN (0.798) as well as ViT-Base (0.833), EfficientNet-B4 (0.892) and ResNet-50 (0.913). Its accuracy advantage over the Basic CNN therefore reflects the operating point at the default 0.5 threshold rather than a better ranking of malignant above benign images, and on this dataset AKGNet does not offer a better accuracy–efficiency

trade-off than the larger models. The training curves (Fig. 11) show that the pretrained models fit the training set within a few epochs, whereas the training accuracy of the compact models was still rising after 20 epochs; the validation curves of all models fluctuate strongly from epoch to epoch, reflecting the small number of validation patients. For AKGNet, benign images classified as malignant are the dominant error (152 of 455 benign test images, against 89 of 886 malignant images missed), whereas ResNet-50 errs about equally in both directions (89 and 87; Fig. 12).

The statistical evidence on BreakHis is limited. The test partition contains about twelve patients from a single split, and images from one patient are strongly correlated; the image-level McNemar tests in Table 4 treat 1,341 dependent images as independent and therefore overstate the evidence. Their p-values are reported descriptively only: no superiority over the Basic CNN is claimed from them, and the non-significant comparison with EfficientNet-B4 is not interpreted as equivalence, which the AUC difference of 0.149 (0.743 versus 0.892) contradicts. Patient-clustered inference, for example a patient-level bootstrap or repeated patient-grouped cross-validation with paired comparisons across folds, is required and was not performed. Likewise, the difference between the best validation accuracy (98.8%) and the test accuracy should not be read as an isolated effect of patient separation: the validation partition was itself patient-disjoint and was used for checkpoint selection, so with only a few patients per partition the difference also reflects selection on a small, noisy validation set and heterogeneity between patients.

The ROC curve of AKGNet (Fig. 13) has an unusual, step-like shape, which was checked against the confusion matrix (Fig. 12). The operating point at the 0.5 threshold (false-positive rate 152/455 = 0.334, true-positive rate 797/886 = 0.900) lies on the plotted curve, so the curve and the reported accuracy are consistent. Up to a false-positive rate of about 0.26, i.e. about 118 of the 455 benign test images, the curve rises only slowly before climbing steeply: a large block of benign images receives malignant scores as high as those of most malignant images. Such a shape is expected when errors are concentrated in a few patients, because all images of a confidently misclassified benign patient then occupy the top of the score ranking together. Since the per-patient breakdown was not computed, this explanation remains a hypothesis; the shape nevertheless illustrates why an image-level ROC-AUC computed on about a dozen test patients is an unstable summary.

TABLE 4. BreakHis benign versus malignant results, patient-disjoint test partition

Model	Accuracy	Balanced acc	Macro-F1	ROC-AUC	Image-level McNemar p vs AKGNet (descriptive)
AKGNet	0.8203	0.7827	0.7921	0.7434	N/A
Basic CNN	0.8039	0.7741	0.7780	0.7977	0.0087
ResNet-50	0.8688	0.8531	0.8535	0.9132	2.9×10^{-6}
EfficientNet-B4	0.8345	0.8095	0.8131	0.8922	0.20
ViT-Base	0.8434	0.8141	0.8211	0.8330	0.0044

Note: Single split, seed 42; 1,341 images from about twelve patients. Image-level McNemar p-values against AKGNet are descriptive only: they ignore within-patient correlation and support neither significance nor equivalence claims. N/A, not applicable.

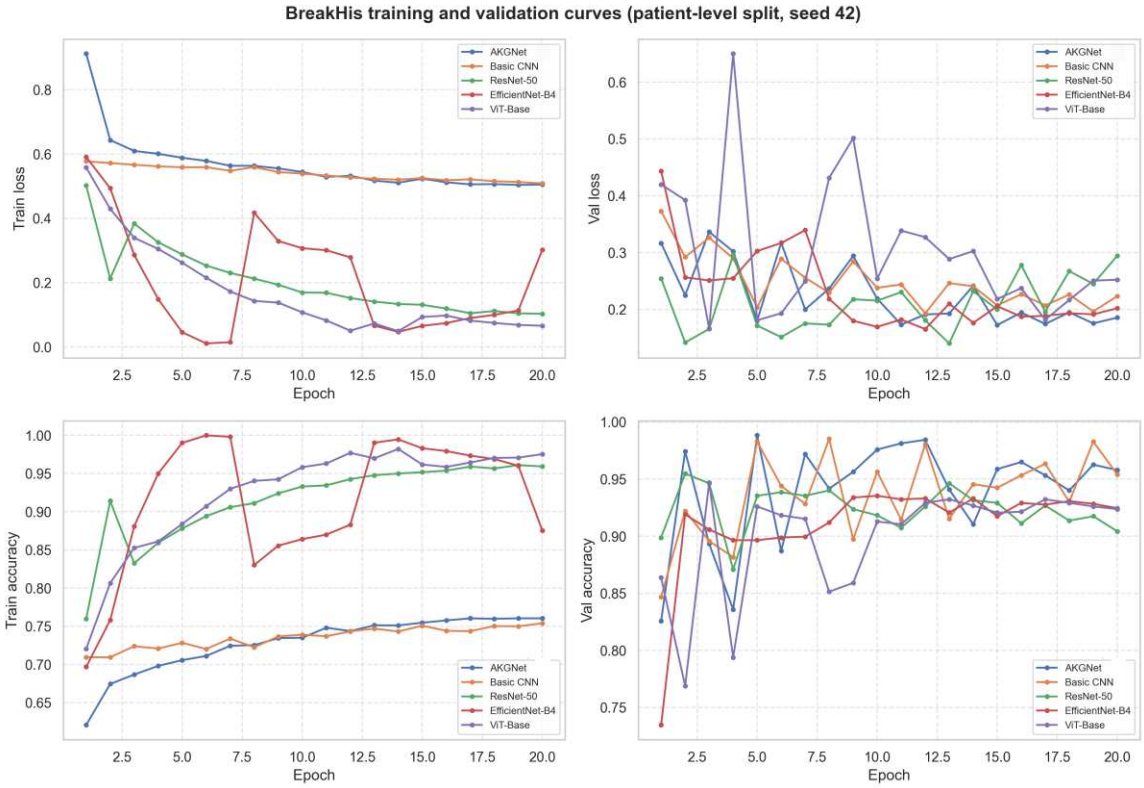


Fig. 11. BreakHis training and validation curves (patient-disjoint split, seed 42)

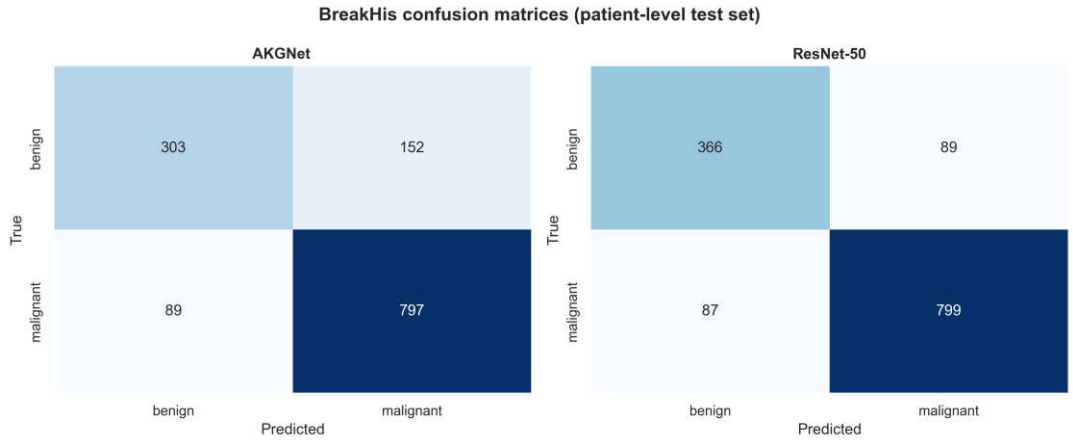


Fig. 12. BreakHis confusion matrices of AKGNet and ResNet-50 (patient-disjoint test partition)

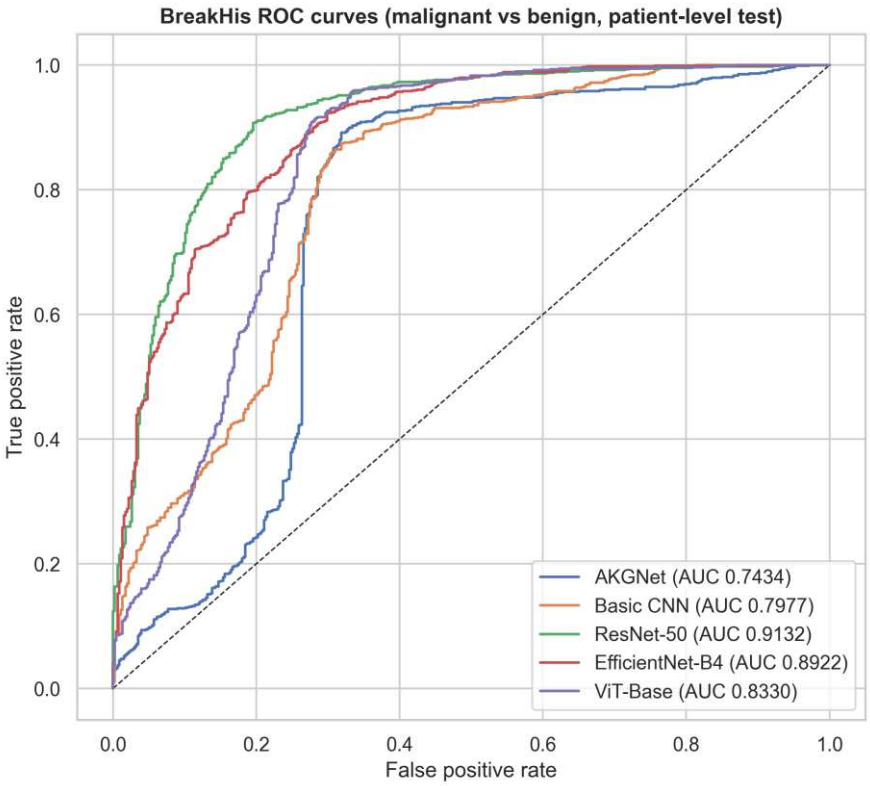


Fig. 13. BreakHis image-level ROC curves of all models, patient-disjoint test partition

5.4 Robustness, cross-dataset behaviour and qualitative analysis

Robustness was probed along three axes, and the failure cases are collected in Table 5. Under nine common image corruptions (JPEG recompression at two qualities, additive Gaussian noise at two amplitudes, Gaussian blur at two radii, and brightness and contrast shifts), robustness broadly scales with capacity and pretraining: ViT-Base retains 98.6% of its clean accuracy on average and ResNet-50 95.9%, against 85.2% for AKGNet and 82.9% for EfficientNet-B4. AKGNet exceeds the size-matched Basic CNN in all nine conditions and degrades gradually under moderate Gaussian noise (88.4% against 67.8% for EfficientNet-B4, which falls to 2.4% under strong noise), while its own weakest conditions are strong noise and strong blur, consistent with a reliance on fine local texture (Fig. 14). Under stain-colour perturbation beyond the training-jitter envelope (HED-space shifts of twice the training amplitude, and grayscale conversion), AKGNet fails: its accuracy falls to near chance under both strong shifts and to 48.2% on grayscale, whereas the large pretrained models mostly retain 81–94%, although ResNet-50 also collapses (23.8%) under the negative shift. Within the training-jitter envelope all models are stable. This is the most important practical limitation of the compact model; deployment behind stain normalization, or a wider stain-augmentation envelope, would be required. Finally, in a supplementary small-data experiment on the nine-class NCT-CRC-7K colorectal collection [35] (about 5,000 training images), the size-class ordering reversed (Basic CNN 89.7% versus AKGNet 85.1%), so the advantage of AKGNet over the plain control did not hold in this few-thousand-image, fine-grained setting.

TABLE 5. Robustness and failure cases on LC25000 unless stated otherwise

Setting	AKGNet	Comparison models	Observation
Nine common corruptions, mean accuracy retained (% of clean)	85.2%	ViT-Base 98.6%, ResNet-50 95.9%, EfficientNet-B4 82.9%; Basic CNN below AKGNet in all nine conditions	Robustness broadly scales with capacity and pretraining
Moderate Gaussian noise	88.4%	EfficientNet-B4 67.8%	Compact model degrades gradually
Strong Gaussian noise and strong blur	Weakest conditions of AKGNet (Fig. 14)	EfficientNet-B4 2.4% under strong noise	Reliance on fine local texture
HED stain shift, twice the training amplitude (positive and negative)	Near chance under both shifts	Large pretrained models mostly 81–94%; ResNet-50 23.8% under the negative shift	Failure outside the stain-augmentation envelope; stain normalization required for deployment
Grayscale conversion	48.2%	Large pretrained models mostly 81–94%	Strong dependence on colour cues
NCT-CRC-7K, nine classes, about 5,000 training images	85.1%	Basic CNN 89.7%	Size-class advantage did not hold in a small, fine-grained setting

Note: One evaluation per condition.

Qualitatively, Grad-CAM analysis (Fig. 15) shows AKGNet attending to fine cytological structure, nuclei clusters, gland borders and crypt edges, whereas ResNet-50 attends to broader tissue-architecture regions. Neither model attends to background, corners or slide artefacts, which argues against trivial shortcut learning; in misclassified cases the attention of the compact model is visibly more scattered. These visualisations support interpretation only and are not treated as evidence of accuracy.

Grad-CAM does not show the routing maps themselves. Visualizing the gate maps, quantifying their spatial variation, entropy and per-expert utilization, and comparing with a control that uses one mixture per image would be needed to show that the trained routers exploit spatial variation. These analyses were not performed, so the statement that different regions of one image receive different scale mixtures describes what the design permits, not a verified behaviour of the trained models.

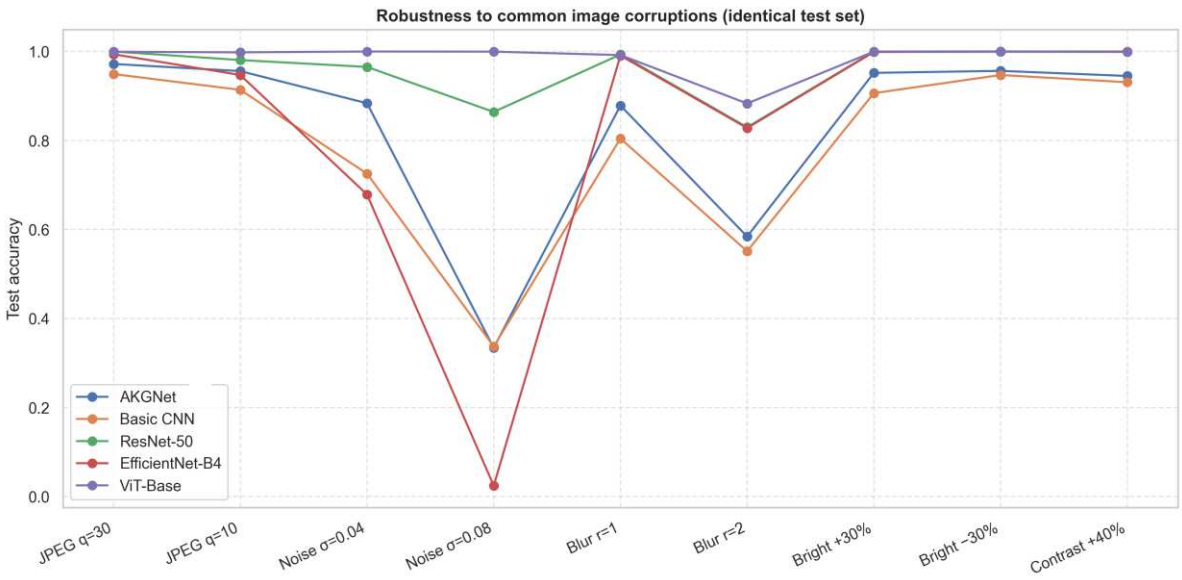


Fig. 14. Accuracy under nine image corruptions on the LC25000 test set

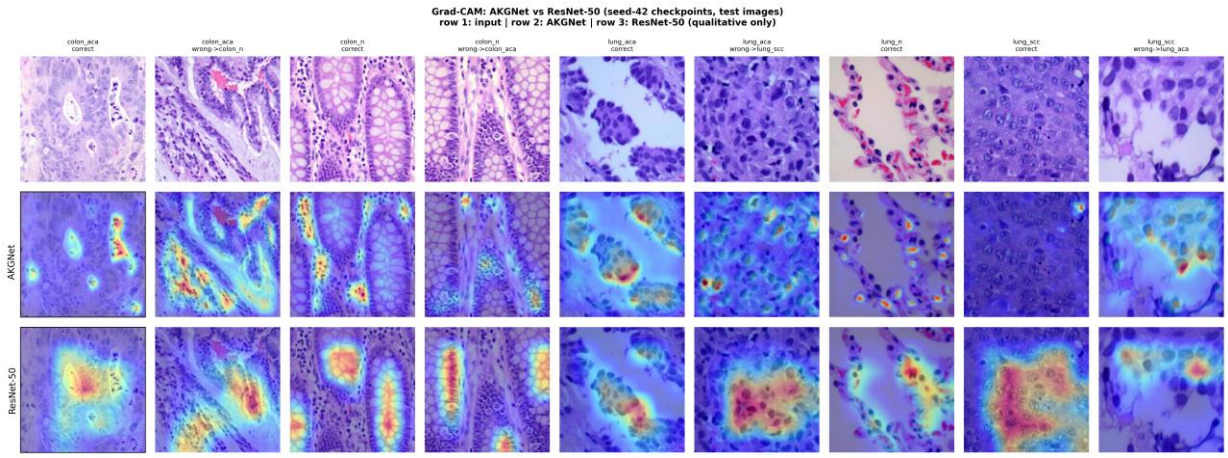


Fig. 15. Grad-CAM on LC25000: input (top), AKGNet (middle), ResNet-50 (bottom)

6. Ablation Study

Five single-component ablations were trained once (seed 42) under the identical protocol and evaluated on the same test images (Table 6); the full model is therefore compared at the same seed (97.28%) rather than with its three-seed mean. Because each ablation is a single run, differences of a few tenths of a point are within the seed-to-seed variation of the full model (SD 0.25 points), and repeating the ablations over the same three seeds would be required for firm conclusions. Replacing the learned router with uniform averaging of the four experts costs 1.36 points ($p = 1.7 \times 10^{-6}$), indicating that the learned weighting carries signal beyond the presence of parallel kernels. Removing spatial attention (−0.77 points; $p = 0.0019$) or the max-pooling branch of the head (−0.72; $p = 0.005$) gives smaller decreases of about three seed standard deviations. Removing gate smoothing changes accuracy by −0.08 points ($p = 0.8$), an effect that is not detectable; smoothing and the temperature are therefore implementation choices without an accuracy claim. The single-expert variant requires a precise description: its parameter count (22,986) corresponds to retaining only the 1×1 expert in each block, with a one-channel router, so its accuracy of 86.03% (−11.25 points; $p = 5.7 \times 10^{-83}$) reflects the removal of all spatial aggregation inside the gating blocks, not the removal of multi-scale processing as such. Single-spatial-kernel controls (for example, a 3×3-only variant with 47,562 parameters) and a parameter-matched fixed multi-scale block were not trained, so the contribution of multi-scale selection per se is not isolated by these experiments. The McNemar p-values in this section are image-level and are affected by the dependence among augmented LC25000 siblings noted in Section 5.1.

TABLE 6. Ablation study on LC25000 (one run per variant, seed 42)

Variant	Accuracy	Balanced accuracy	Macro-F1	Parameters (M)	FLOPs (G)
AKGNet (full)	0.9728	0.9728	0.9728	0.280	1.837
w/o spatial attention	0.9651	0.9651	0.9651	0.279	1.836
w/o gate smoothing	0.9720	0.9720	0.9720	0.280	1.837
single 1×1 expert (one-channel router)	0.8603	0.8603	0.8594	0.023	0.230
w/o max-pooling branch (average pooling only)	0.9656	0.9656	0.9656	0.279	1.837
uniform expert averaging (no router)	0.9592	0.9592	0.9591	0.264	1.740

Note: The full model is shown at the same seed. Parameters refer to the five-class head; FLOPs are two operations per multiply–accumulate at 224×224.

7. Comparison Study

Table 7 situates AKGNet among recent LC25000 and BreakHis studies, and Table 8 reports the measured computational comparison. Published image-level LC25000 accuracies concentrate near or above 99% [19]–[22], and the pretrained baselines tested here reach the same range; the duplicate analysis of Section 5.1 shows that image-level splits of this benchmark contain many cross-partition near-duplicates, but it does not determine how much of this accuracy depends on them, and the LC25000-clean grouping [17] offers a better basis for such comparisons. On BreakHis, reported "patient-level" scores of 99% [23], [24] derive from image-level splits with per-patient aggregation; under patient-disjoint evaluation the accuracies of all architectures tested here fall to the mid-80s or below, consistent with [25]. In accuracy, AKGNet is the strongest model of its size class on LC25000 (Fig. 16) but is clearly below the larger pretrained models on both datasets.

Efficiency claims require distinguishing parameters from computation. AKGNet has the fewest parameters (0.28 M, 1.07 MB) and the lowest peak inference memory (158 MB at batch size 32) of the models in Table 8, but because the four experts are evaluated densely, its 1.84 GFLOPs exceed those of the Basic CNN (1.50 GFLOPs), and at batch size 32 on the GPU it is slower than the Basic CNN (29.6 versus 26.1 ms); relative to the large baselines it is 3.1×, 4.2× and 10.1× faster than ResNet-50, EfficientNet-B4 and ViT-Base. Batch-size-1 latency on a CPU, which is closer to interactive or edge deployment, and comparisons with standard lightweight CNNs such as MobileNet and ShuffleNet were not measured in this study and are needed before deployment claims can be made. Overall, AKGNet is parameter- and memory-efficient, but its gating does not reduce arithmetic, and the measurements do not establish a better accuracy–efficiency trade-off than the larger pretrained models, which are more accurate on both datasets and, on BreakHis, substantially better in ROC-AUC.

TABLE 7. Comparison with recent published studies (protocols stated per row)

Work	Method	Dataset	Protocol	Reported accuracy (%)	Parameters
Masud et al. [18], 2021	Compact CNN, spatial+spectral features	LC25000	image-level split	96.33	~1 M (compact)
Attallah et al. [19], 2022	Lightweight CNNs + FWHT/PCA + classical classifier	LC25000	image-level split	99.6	multi-model
Ji et al. [21], 2024	Swin Transformer V2	LC25000	5-fold, image-level	≥ 99 (reported)	~88 M
EfficientNet-B3 TL [22], 2025	Transfer learning	LC25000	image-level split	99.39	12 M
HoRFNet [24], 2025	High-order multi-branch receptive fields	BreakHis	image-level split, per-patient aggregation	99.23	> 20 M
AKGNet (proposed)	Spatially adaptive kernel gating, from scratch	LC25000	image-level split, single-use test; hash-group split (seed 42): 96.70	97.02 ± 0.25	0.28 M
AKGNet (proposed)	as above	BreakHis	patient-disjoint split (single split)	82.03	0.28 M

TABLE 8. Measured computational comparison at batch size 32

Model	Parameters (M)	FLOPs (G)	Latency B32 (ms)	Throughput (img/s)	Peak GPU mem (MB)
AKGNet	0.280	1.837	29.63 ± 3.39	1,080	158
ResNet-50	23.518	8.263	92.08 ± 8.36	348	427
EfficientNet-B4	17.558	3.154	125.78 ± 11.95	254	565
ViT-Base	85.803	22.571	297.94 ± 15.35	107	594
Basic CNN	0.391	1.498	26.13 ± 2.61	1,225	499

Note: Idle-GPU synchronized timing on an RTX 2060 with PyTorch 2.4.1; latency is the mean ± SD over 50 timed runs.

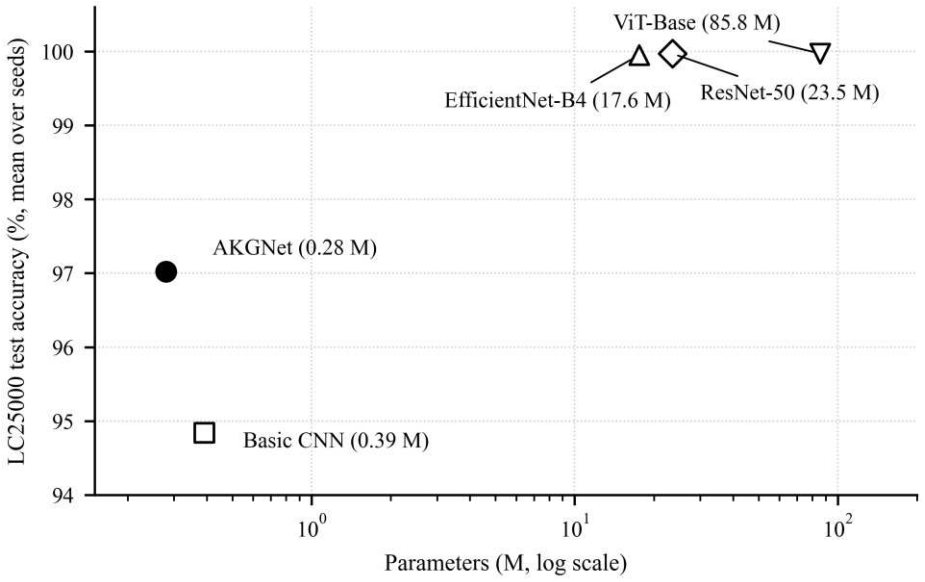


Fig. 16. Measured LC25000 accuracy versus parameter count (logarithmic parameter axis)

8. Conclusion

This paper introduced AKGNet, a 0.28 M-parameter histopathology classifier built on spatially adaptive kernel gating with temperature-scaled and smoothed routing, and evaluated it on LC25000 and on a patient-disjoint BreakHis split. On LC25000 the model reaches $97.02 \pm 0.25\%$ accuracy, above a size-matched convolutional control with 40% more parameters but below ImageNet-pretrained baselines that approach 100%. On BreakHis it reaches 82.03% accuracy but the lowest ROC-AUC of the models tested (0.743), and with about twelve test patients from a single split the BreakHis comparisons are not conclusive. The model is parameter- and memory-efficient and infers 3–10× faster than the large baselines on a GPU, but because its experts are computed densely it does not reduce arithmetic relative to a plain CNN, and the results do not support a better accuracy–efficiency trade-off than larger models. Single-seed ablations indicate that learned routing adds 1.36 points over uniform expert averaging on LC25000, whereas gate smoothing has no detectable effect. The perceptual-hash analysis shows that random image-level splits of LC25000 place detected near-duplicates across partitions, but it is not rotation-invariant and does not bound the resulting inflation.

The limitations of the study define the next steps: (i) the BreakHis results come from one patient-disjoint split and image-level tests, so repeated patient-grouped cross-validation with patient-clustered confidence intervals is required; (ii) the LC25000 comparison should be repeated on a rotation-tolerant source grouping such as LC25000-clean [17]; (iii) the ablations should be repeated over the same seeds as the full model and

complemented by single-spatial-kernel controls; (iv) a per-image dynamic-convolution baseline and lightweight baselines evaluated for accuracy and batch-size-1 CPU latency are needed to attribute the gains specifically to spatial gating and to support deployment claims; (v) the router's behaviour should be verified with gate maps and with entropy and expert-utilization statistics; (vi) the from-scratch models were still improving at 20 epochs, so longer schedules should be evaluated; (vii) the compact model is not robust to stain shifts beyond its augmentation envelope and did not keep its advantage on a small nine-class colorectal dataset (Table 5); and (viii) sparse or top-k expert evaluation would be needed to turn the gating into computational savings.

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Competing Interests: The authors declare that they have no competing interests.

Author Contributions: Chandrapal Singh: Conceptualization, methodology, software, implementation, experimentation, data analysis, visualization, and writing original draft. All authors contributed to the interpretation of the results, reviewed the manuscript, and approved the final version of the manuscript.

Ethics Approval: Ethics approval was not required for this study because the research used only publicly available, de-identified datasets (LC25000 and BreakHis) and did not involve the collection or processing of new data from human participants.

Consent to Participate: Not applicable. This study used publicly available, de-identified datasets and did not involve direct participation of human subjects.

Consent to Publish: Not applicable. This study does not contain personally identifiable information or individual-level information that requires consent for publication.

Data and code availability: LC25000 and BreakHis are publicly available from their original sources [2], [3]. The evaluated code, training configurations, model checkpoints, the BreakHis patient-to-partition assignment and the LC25000 hash-group assignment will be released in a public repository upon publication.

References

[1] H. Sung *et al.*, "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians*, vol. 71, no. 3, pp. 209–249, 2021, doi: 10.3322/caac.21660.

[2] A. A. Borkowski, M. M. Bui, L. B. Thomas, C. P. Wilson, L. A. DeLand, and S. M. Mastorides, "Lung and colon cancer histopathological image dataset (LC25000)," 2019, *arXiv:1912.12142*, doi: 10.48550/arXiv.1912.12142.

[3] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, "A dataset for breast cancer histopathological image classification," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 7, pp. 1455–1462, 2016, doi: 10.1109/TBME.2015.2496264.

[4] R. Schmitz *et al.*, "Multi-scale fully convolutional neural networks for histopathology image segmentation: From nuclear aberrations to the global tissue architecture," *Medical Image Analysis*, vol. 70, Art. no. 101996, 2021, doi: 10.1016/j.media.2021.101996.

[5] N. Hashimoto *et al.*, "Multi-scale domain-adversarial multiple-instance CNN for cancer subtype classification with unannotated histopathological images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 3851–3860, doi: 10.1109/CVPR42600.2020.00391.

[6] J. Li *et al.*, "A multi-resolution model for histopathology image classification and localization with multiple instance learning," *Computers in Biology and Medicine*, vol. 131, Art. no. 104253, 2021, doi: 10.1016/j.compbiomed.2021.104253.

[7] B. Yang, G. Bender, Q. V. Le, and J. Ngiam, "CondConv: Conditionally parameterized convolutions for efficient inference," in *Advances in Neural Information Processing Systems*, vol. 32, 2019, pp. 1307–1318, doi: 10.48550/arXiv.1904.04971.

[8] Y. Chen, X. Dai, M. Liu, D. Chen, L. Yuan, and Z. Liu, "Dynamic convolution: Attention over convolution kernels," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 11027–11036, doi: 10.1109/CVPR42600.2020.01104.

[9] J. Dai *et al.*, "Deformable convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 764–773, doi: 10.1109/ICCV.2017.89.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[11] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proceedings of the 36th International Conference on Machine Learning (ICML)*, Proceedings of Machine Learning Research, vol. 97, 2019, pp. 6105–6114, doi: 10.48550/arXiv.1905.11946.

[12] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021, doi: 10.48550/arXiv.2010.11929.

[13] O. N. Manzari, H. Ahmadabadi, H. Kashiani, S. B. Shokouhi, and A. Ayatollahi, "MedViT: A robust vision transformer for generalized medical image classification," *Computers in Biology and Medicine*, vol. 157, Art. no. 106791, 2023, doi: 10.1016/j.compbiomed.2023.106791.

[14] M. Ferdous, S. Mahmud, and M. E. Z. Shimul, "MedNet: A lightweight attention-augmented CNN for medical image classification," *Scientific Reports*, vol. 15, no. 1, Art. no. 41936, 2025, doi: 10.1038/s41598-025-25857-w.

[15] D. Tellez *et al.*, "Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology," *Medical Image Analysis*, vol. 58, Art. no. 101544, 2019, doi: 10.1016/j.media.2019.101544.

[16] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," *Patterns*, vol. 4, no. 9, Art. no. 100804, 2023, doi: 10.1016/j.patter.2023.100804.

[17] G. Batchkala, B. Li, and J. Rittscher, "Evaluating histopathology foundation models for few-shot tissue clustering: An application to LC25000 augmented dataset cleaning," in *Data Engineering in Medical Imaging (DEMI 2024)*, Lecture Notes in Computer Science, vol. 15265. Cham, Switzerland: Springer, 2025, pp. 11–21, doi: 10.1007/978-3-031-73748-0_2.

[18] M. Masud, N. Sikder, A.-A. Nahid, A. K. Bairagi, and M. A. AlZain, "A machine learning approach to diagnosing lung and colon cancer using a deep learning-based classification framework," *Sensors*, vol. 21, no. 3, Art. no. 748, 2021, doi: 10.3390/s21030748.

[19] O. Attallah, M. F. Aslan, and K. Sabanci, "A framework for lung and colon cancer diagnosis via lightweight deep learning models and transformation methods," *Diagnostics*, vol. 12, no. 12, Art. no. 2926, 2022, doi: 10.3390/diagnostics12122926.

[20] M. M. R. Said *et al.*, "Innovative deep learning architecture for the classification of lung and colon cancer from histopathology images," *Applied Computational Intelligence and Soft Computing*, vol. 2024, no. 1, Art. no. 5562890, 2024, doi: 10.1155/2024/5562890.

[21] J. Ji *et al.*, "Automated lung and colon cancer classification using histopathological images," *Biomedical Engineering and Computational Biology*, vol. 15, Art. no. 11795972241271569, 2024, doi: 10.1177/11795972241271569.

[22] R. Ochoa-Ornelas, A. Gudiño-Ochoa, J. A. García-Rodríguez, and S. Uribe-Toscano, "A robust transfer learning approach with histopathological images for lung and colon cancer detection using EfficientNetB3," *Healthcare Analytics*, vol. 7, Art. no. 100391, 2025, doi: 10.1016/j.health.2025.100391.

[23] L. Cao, K. Pan, Y. Ren, R. Lu, and J. Zhang, "Multi-branch spectral channel attention network for breast cancer histopathology image classification," *Electronics*, vol. 13, no. 2, Art. no. 459, 2024, doi: 10.3390/electronics13020459.

[24] M. Zhao, C. Hou, L. Cao, and J. Zhang, "Breast cancer histopathological image classification based on high-order modeling and multi-branch receptive fields," *Applied Sciences*, vol. 15, no. 11, Art. no. 6085, 2025, doi: 10.3390/app15116085.

[25] P. Veeram, N. Modigari, and D. R. Tharasi, “A patient-aware benchmarking of CNN and transformer architectures for breast cancer histopathology classification,” *Frontiers in Digital Health*, vol. 8, Art. no. 1752938, 2026, doi: 10.3389/fdgth.2026.1752938.

[26] X. Li, W. Wang, X. Hu, and J. Yang, “Selective kernel networks,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 510–519, doi: 10.1109/CVPR.2019.00060.

[27] C. Li and A. Yao, “KernelWarehouse: Rethinking the design of dynamic convolution,” in *Proceedings of the 41st International Conference on Machine Learning (ICML)*, Proceedings of Machine Learning Research, vol. 235, 2024, pp. 29201–29221, doi: 10.48550/arXiv.2406.07879.

[28] T. Zhang, F. Wan, H. Duan, K. W. Tong, J. Deng, and Y. Long, “FMDConv: Fast multi-attention dynamic convolution via speed-accuracy trade-off,” *Knowledge-Based Systems*, vol. 317, Art. no. 113393, 2025, doi: 10.1016/j.knosys.2025.113393.

[29] S. Mu and S. Lin, “A comprehensive survey of mixture-of-experts: Algorithms, theory, and applications,” 2025, *arXiv:2503.07137*, doi: 10.48550/arXiv.2503.07137.

[30] Y. Wu and K. He, “Group normalization,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, Lecture Notes in Computer Science, vol. 11217. Cham, Switzerland: Springer, 2018, pp. 3–19, doi: 10.1007/978-3-030-01261-8_1.

[31] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, “CBAM: Convolutional block attention module,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, Lecture Notes in Computer Science, vol. 11211. Cham, Switzerland: Springer, 2018, pp. 3–19, doi: 10.1007/978-3-030-01234-2_1.

[32] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual explanations from deep networks via gradient-based localization,” in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626, doi: 10.1109/ICCV.2017.74.

[33] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019, doi: 10.48550/arXiv.1711.05101.

[34] S. Holm, “A simple sequentially rejective multiple test procedure,” *Scandinavian Journal of Statistics*, vol. 6, no. 2, pp. 65–70, 1979.

[35] J. N. Kather *et al.*, “Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study,” *PLOS Medicine*, vol. 16, no. 1, Art. no. e1002730, 2019, doi: 10.1371/journal.pmed.1002730.