Compact and Structurally Transparent Cervical Cytology with Geometry-Driven Features and Closed-Form Attention

Dichao Liu[1]

College of Artificial Intelligence, Dalian Maritime University, Dalian, China
dichao_liu@dlmu.edu.cn

Abstract. Cervical cytology is practiced as a sequence of roles: screening decides *where* to look, careful reading determines *what* is present, and reporting accounts for the decision. We translate this workflow into a compact and structurally transparent model. Screening is cast as a free–energy allocation over four explicit, cytology–native concepts (nuclear–envelope roughness, orientation disorder, local N/C ratio, chromatin heterogeneity), yielding a closed-form spatial attention as a softmax of negative energy. Reading is handled by an analytic backbone built from fixed geometric responses, nonnegative near–diagonal channel mixing, and convex residual updates, producing stable and auditable features. A probabilistic readout with a linear head then provides an intrinsic evidence–to–logit accounting without post–hoc explainers. Despite only 2.2M parameters, the model achieves state-of-the-art accuracy on DSCC (93.2% ACC), competitive results on SIPaKMeD (98.5% ACC) and Herlev (76.8% ACC), and favorable latency. Source code is publicly available at `https://github.com/Dichao-Liu/GeoCEAN`.

Keywords: Cervical cytology · Transparent AI · White-box model

1 Introduction

Cervical cancer is a leading cause of mortality in women, yet early detection through cytological screening makes it preventable [27]. Automated analysis is essential to scale screening and reduce workload. Most current models either transfer generic backbones or build dataset-specific systems; both often lack interpretability and show uneven generalization. The challenge is to design a compact model that is auditable and clinically aligned.

Clinical diagnosis follows a clear workflow: a screening technologist decides *where* to look, an attending pathologist examines *what* is present, and the final report explains the decision [1]. We adopt this structure as our design principle. The resulting system, termed GeoCEAN (Geometry-driven Concept–Energy–Attention Network), is shown in Fig. 1.

Where to look. GeoCEAN computes four concept maps reflecting diagnostic cues: nuclear-envelope irregularity, epithelial orientation disorder, local

nucleus-to-cytoplasm ratio, and chromatin heterogeneity. After monotone calibration, these maps form a nonnegative diagnostic potential field. Minimizing a free-energy objective driven by this potential yields a closed-form spatial attention, mirroring the screening phase by allocating a finite reading budget analytically before feature extraction.

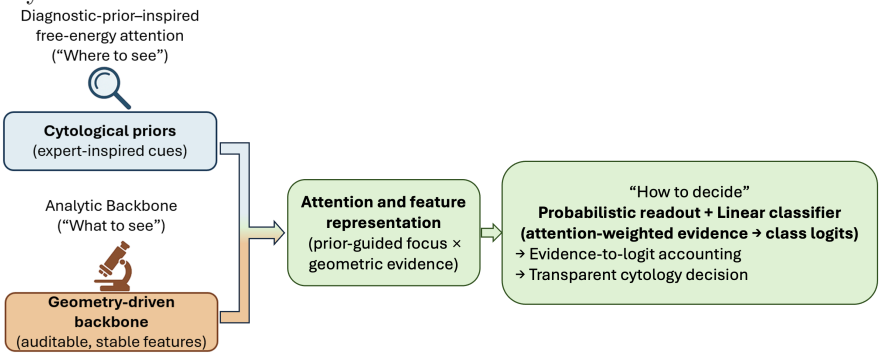

Fig. 1: GeoCEAN at a glance. Echoing cytology practice, a prior-driven free-energy attention marks where to look, while a geometry-driven backbone reads what is present. Both parts follow a structurally transparent design: the attention is the closed-form optimum of a free-energy objective, and the backbone relies on fixed geometric operators with constrained (nonnegative, near-diagonal) mixing and convex updates, producing auditable, stable features. A probabilistic readout + linear head then aggregates attention-weighted features into logits, yielding clear evidence-to-logit accounting from pixels to decision.

What to read. A geometry-driven analytic backbone extracts features using fixed gradient and Laplacian operators, nonnegative near-diagonal channel mixing, and convex residual updates. These constraints preserve correspondence between computation and cellular structures and keep feature evolution stable and auditable through depth. The backbone is co-designed with the solved attention so that nonnegative near-diagonal mixing preserves positive geometric evidence, while convex residual updates keep feature evolution controlled and auditable through depth, distinguishing GeoCEAN from generic white-box designs.

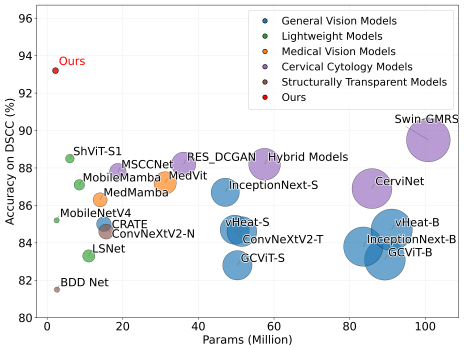

Fig. 2: Accuracy–parameter trade-off on DSCC. Our model attains the highest accuracy (93.2%) with only 2.2M parameters, outperforming larger general-purpose, lightweight, medical, and cervix-specific baselines.

How to decide. An attention-weighted probabilistic readout aggregates features into a compact descriptor, followed by a linear head that produces class

logits. This yields an intrinsic concept-to-logit mapping without post-hoc explanation.

Despite its small size, GeoCEAN achieves state-of-the-art accuracy with only 2.2M parameters (Fig. 2), outperforming larger general, medical, and Pap-specific models on DSCC, SIPaKMeD, and Herlev. Embedding domain structure improves interpretability, robustness, and efficiency.

Contributions. (i) Cytology screening is formulated as free-energy allocation with a closed-form spatial attention, where clinical monotonicity refers to strictly increasing concept calibration and nonnegative simplex fusion from concepts to diagnostic potential. (ii) A geometry-driven analytic backbone enforces nonnegative, locally band-limited, and convex updates for stable, auditable computation. (iii) A probabilistic readout and linear head provide end-to-end evidence-to-logit accounting within a compact model. (iv) The backbone is jointly constructed with the closed-form attention so that nonnegative mixing preserves positive geometric evidence and the convex residual update controls feature evolution, providing structural novelty beyond prior white-box designs while remaining efficient and interpretable.

2 Related Work

General-purpose and compact backbones. Recent CNN and ViT families target accuracy and efficiency but remain task-agnostic. ConvNeXtV2 [25], GCViT [7], and InceptionNeXt [29] expand receptive fields and hierarchical design; SHViT [33] and vHeat [24] streamline attention or replace it with physical operators; MobileNetV4 [17] and LSNet [22] optimize under hardware constraints. These models provide strong accuracy–latency baselines but do not encode medical priors or clinical accountability.

Medical image backbones. Domain-oriented designs combine convolutional locality with long-range modeling. MedViT [14] builds hybrid pyramids of convolutional and transformer blocks, and MedMamba [32] uses state-space models for efficient context. They improve robustness under shift, yet attention and feature mixing remain implicit and unconstrained by cytological concepts such as nuclear rims, N/C ratio, or chromatin pattern.

Task-specific cervical cytology. Pap-smear systems range from early transfer-learning pipelines on Herlev [9] and SIPaKMeD [20] to structured designs. MSC-CNet [2] uses multi-scale capsules for spatial relations; CerviNet [10] builds a cervical feature hierarchy; RES_DCGAN [26] adds generative augmentation with self-attention. Recent work fuses multiple backbones or introduces auxiliary reconstruction and distillation to address label ambiguity [4]. These methods can be accurate but rely on post-hoc interpretation and lack explicit alignment to clinical semantics, limiting cross-dataset stability. *Other lines* explore knowledge-guided descriptors, multi-task fusion, contextual relations, and ensemble/lightweight ViTs for Pap-smear classification [5, 11, 13, 15, 18, 21, 27, 28].

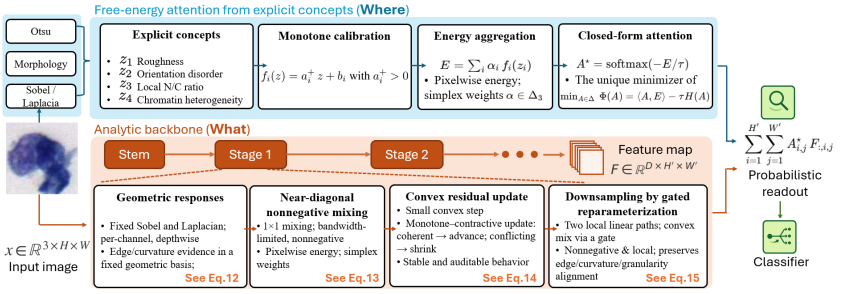

Fig. 3: Overview of the proposed architecture. Prior cytological knowledge guides a free-energy attention to locate **where** diagnostic evidence lies, while a geometry-driven analytic backbone maintains explicit geometric correspondence between operations and interpretable cellular structures, encoding **what** is present in an auditable way. A separable probabilistic readout combines attention and features with a linear head to produce logits, yielding a clear evidence-to-logit accounting from pixels to decision.

Structurally transparent vision. Post-hoc tools such as Grad-CAM [19] visualize saliency without changing how a model reasons, whereas White-box approaches embed interpretability into computation. Dendrite Net [12] and BDD Net [6] craft analyzable dendritic modules; CRATE [30] derives an interpretable encoder–decoder through sparse rate reduction and unrolled optimization; follow-up studies [31] show emergent grouping from analyzable updates. We refer to structural transparency as architectures where selection, interaction, and decision are explicitly solvable, mathematically constrained, and decomposable. Prior work supports this principle but rarely incorporates domain-native clinical semantics or an interpretable spatial allocation process. Dynamic routing attention and cytology-grade precision diagnosis offer alternative allocation paradigms [23, 34].

Our positioning. GeoCEAN sits at the intersection of compact design, medical adaptation, and white-box interpretability. Its distinction is that cytology evidence is formed, spatially allocated, morphologically read, and accumulated into logits within a single forward computation, rather than being imposed as a generic white-box constraint or explained post hoc after a medical classifier. Compared with general backbones [7,17,22,24,25,29,33], it keeps efficiency while replacing generic token mixing with task-aligned attention derived in closed form. Relative to medical backbones [14,32], it constrains channel interaction through nonnegative near-diagonal mixing and convex residual updates, preserving interpretable geometric features. Compared with Pap-specific systems [2,4,10,20,26], it encodes cytological reasoning directly in the computation graph rather than relying on multi-branch fusion. In line with white-box modeling [6,12,30,31], it is analyzable from concept pixels to logits while keeping semantics domain-native and attention analytically derived.

3 Method

3.1 Overview

In cervical cytology, diagnostic evidence concentrates in a few structures repeatedly queried by experts: (i) irregularities of the nuclear envelope, (ii) loss of epithelial orientation, (iii) deviations of the nucleus-to-cytoplasm (N/C) ratio, and (iv) chromatin granularity. As shown in Fig. 3, we encode this prior by first constructing four pixelwise maps on the input image $x \in \mathbb{R}^{3\times H\times W}$. Here $H\times W$ is the image plane and we denote its discrete lattice by $\Omega = \{1,\dots,H\}\times\{1,\dots,W\}$. A pixel/location is written as $u \in \Omega$. For each concept $i = 1,\dots,4$, we compute

$$\{z_i(u)\}_{i=1}^{4}, \qquad u \in \Omega. \tag{1}$$

We index i over the ordered set $\{\text{rough}, \text{orient}, \text{nc}, \text{ent}\}$, i.e., $z_1 = z_{\text{rough}},\ z_2 = z_{\text{orient}},\ z_3 = z_{\text{nc}},\ z_4 = z_{\text{ent}}$.

A monotone calibration f_i transforms each concept into a common scale and a nonnegative simplex weight $\alpha \in \Delta_3$ aggregates them into a *diagnostic potential field*

$$V(u) \;=\; \sum_{i=1}^{4} \alpha_i\, f_i\big(z_i(u)\big), \qquad \Delta_3 := \big\{\alpha \in \mathbb{R}_+^4 : \sum_{i=1}^{4} \alpha_i = 1\big\}. \tag{2}$$

Thus $V : \Omega \to \mathbb{R}_+$ acts as a spatial accumulation of cytological abnormality. To model the screening phase as a thermodynamic process, we define the energy landscape as the negative of this potential:

$$E(u) \;=\; -V(u). \tag{3}$$

This formulation ensures that regions with the strongest clinical evidence correspond to the lowest energy states (deepest potential wells).

The last feature map $F \in \mathbb{R}^{D\times H'\times W'}$ lives on a (possibly downsampled) lattice $\Omega' = \{1,\dots,H'\}\times\{1,\dots,W'\}$. We ask "how to spend a finite spatial budget on Ω'", and pose a free-energy principle: among all spatial probability distributions A on Ω' (nonnegative and summing to 1), choose the one that minimizes expected energy penalized by negative entropy. This yields a closed-form attention

$$A^{\star} \;=\; \operatorname{softmax}\big(-E/\tau\big), \tag{4}$$

where $\tau > 0$ is a temperature and E is sampled/aligned on Ω' (by standard resizing) when needed. The attention $(A^{\star}_u)_{u\in\Omega'}$ plays a single role: it selects where to read from F. The image representation is the probabilistic readout

$$m \;=\; \sum_{u\in\Omega'} A^{\star}_u\, F_{:,u} \in \mathbb{R}^D, \qquad y \;=\; W_{\text{cls}}\, m, \quad W_{\text{cls}} \in \mathbb{R}^{C\times D}, \tag{5}$$

followed by a linear classifier that produces logits $y \in \mathbb{R}^C$.

Every term in y is additively decomposable over *locations* $u \in \Omega'$ and *channels* $d \in \{1,\dots,D\}$, and its upstream dependence on the four clinical concepts is explicit through (2)–(5). The structural terms used below are therefore interpreted within this formulation, describing computation-level properties of GeoCEAN's auditable computation for cytology-structured visual classification under the stated design. The pipeline mirrors the pathologist's workflow: delineate concept evidence on the image plane, allocate attention as a principled trade-off between focus and coverage, then linearly tally the gathered evidence.

3.2 Explicit Concepts and Monotone Calibration

We construct four pixelwise maps, each tied to a cytology cue.

Nuclear envelope irregularity (roughness). Let $n : \Omega \to \{0,1\}$ be a nucleus mask (obtained by thresholding with a data-adaptive criterion and morphological cleaning). Denote by ∂n its one-pixel boundary indicator. With a small odd kernel k and the discrete Laplacian Δ,

$$z_{\text{rough}}(u) \;=\; \frac{\big(|\Delta n| \cdot \mathbf{1}_{\partial n}\big) \star k}{\big(\mathbf{1}_{\partial n}\big) \star k + \varepsilon}, \tag{6}$$

i.e., boundary Laplacian flux normalized by local boundary support. Here "$\star$" denotes spatial convolution on Ω, and $\varepsilon > 0$ avoids division by zero.

Orientation disorder. Let $\theta : \Omega \to (-\pi, \pi]$ be the gradient orientation of the grayscale image and $W_u \subset \Omega$ an odd-size window centered at u. The circular variance

$$\mu_s(u) = \frac{1}{|W_u|} \sum_{v \in W_u} \sin\theta(v),\ \mu_c(u) = \frac{1}{|W_u|} \sum_{v \in W_u} \cos\theta(v),\ z_{\text{orient}}(u) = 1 - \sqrt{\mu_s(u)^2 + \mu_c(u)^2}. \tag{7}$$

quantifies the loss of layer-wise alignment (larger means more disordered).

Local N/C ratio. Let $A_n(u)$ and $A_c(u)$ be the nuclear and cytoplasmic areas estimated within W_u (the latter from a controlled dilation of n),

$$z_{\text{nc}}(u) \;=\; \frac{A_n(u)}{A_c(u)}. \tag{8}$$

Chromatin heterogeneity. A stable surrogate of granularity uses the log of local intensity variance,

$$z_{\text{ent}}(u) \;=\; \log\big(1 + \text{Var}_{v \in W_u} I(v)\big). \tag{9}$$

Monotone calibration and simplex fusion. Each concept is passed through a strictly increasing calibration $f_i(z) = a_i^+ z + b_i$ with $a_i^+ > 0$. The nonnegativity and simplex constraint on α in (2) prevent rank inversions: more irregular nuclei or coarser chromatin cannot reduce the local diagnostic potential $V(u)$ (i.e., they cannot increase the local energy $E(u)$). This yields a pointwise monotone link from each calibrated concept to the diagnostic potential. We use clinical monotonicity in this formulation-bounded sense: when other cues are fixed, increasing a calibrated abnormality cue cannot decrease its contribution to $V(u)$.

3.3 Free-Energy Attention and Its Closed Form

Let $\Delta = \{A \in \mathbb{R}_+^{|\Omega'|} : \sum_{u \in \Omega'} A_u = 1\}$ be the probability simplex on Ω'. We define the free-energy functional

$$\Phi(A) \;=\; \langle A, E \rangle \;-\; \tau\, H(A), \quad H(A) = -\sum_{u \in \Omega'} A_u \log A_u, \tag{10}$$

where $\langle A, E\rangle = \sum_{u\in\Omega'} A_u E_u$ and E is the energy field resampled onto Ω' when resolutions differ. Minimizing this free energy drives the system toward the lowest energy states—naturally gravitating the "reading budget" A toward the deepest potential wells where diagnostic evidence $V(u)$ is most concentrated. Meanwhile, the temperature $\tau > 0$ balances this localized focus with spatial coverage.

Proposition 1 (Closed-form optimum). *The unique minimizer of* (10) *over* Δ *is* $A^\star = \mathrm{softmax}(-E/\tau) = \mathrm{softmax}(V/\tau)$.

Proof. Form the Lagrangian $\mathcal{L}(A,\lambda) = \sum_u A_u E_u + \tau \sum_u A_u \log A_u + \lambda(\sum_u A_u - 1)$. First-order optimality gives $E_u + \tau(1+\log A_u) + \lambda = 0$, hence $A_u \propto \exp(-E_u/\tau)$, normalized on Δ. Strict convexity of Φ on Δ implies uniqueness.

Concept-to-logit sensitivity. For class c with weight vector $w_c \in \mathbb{R}^D$, the logit equals

$$y_c = \sum_{u\in\Omega'} A_u^\star \,\langle w_c, F_u\rangle. \tag{11}$$

Since $A^\star = \mathrm{softmax}(V/\tau)$, using $\partial A_u^\star/\partial V_v = (1/\tau)\big(A_u^\star \mathbf{1}_{u=v} - A_u^\star A_v^\star\big)$ and $\partial V_v/\partial z_i(v) = \alpha_i f_i'(z_i(v))$, the marginal effect of concept i at location v reads

$$\frac{\partial y_c}{\partial z_i(v)} = \frac{\alpha_i f_i'(z_i(v))}{\tau}\left(A_v^\star \,\langle w_c, F_v\rangle - \sum_u A_u^\star A_v^\star \,\langle w_c, F_u\rangle\right). \tag{12}$$

Equation (12) provides a closed-form link from each concept pixel to each class logit, supporting fine-grained accountability: "which concept, where, pushes which class."

3.4 Analytic Backbone: Geometry-Driven, Stable Representation

The final feature map F is not produced by an unconstrained stack. Each block first applies fixed geometric operators per channel, then performs nonnegative, bandwidth-limited channel coupling, and finally updates via a convex residual step.

Geometric responses. For an input $X \in \mathbb{R}^{D\times h\times w}$, obtained from the input image through a simple stem layer for initial downsampling, we compute per-channel gradient magnitude and Laplacian responses:

$$M_{\mathrm{gm}} = \sqrt{(G_x * X)^2 + (G_y * X)^2}, \qquad M_{\mathrm{lap}} = |\Delta * X|, \tag{13}$$

where G_x, G_y are fixed derivative kernels and Δ is the discrete Laplacian (all convolutions over the spatial lattice).

Nonnegative near-diagonal mixing. Let $\mathcal{M}$ be a 1×1 mixing with a near-diagonal, nonnegative weight matrix (bandwidth-limited around the main diagonal). With positive coefficients $a_{\mathrm{gm}}, a_{\mathrm{lap}} > 0$,

$$\widehat{X} \;=\; \mathrm{BN}\Big(\mathcal{M}\big(a_{\mathrm{gm}} M_{\mathrm{gm}} + a_{\mathrm{lap}} M_{\mathrm{lap}}\big)\Big). \tag{14}$$

Bandwidth-limited coupling enforces that channels explaining similar geometric phenomena interact, while forbidding far-field, sign-flipping interference.

Convex residual update. The block output is a convex step toward the mixed geometry,

$$X^{\mathrm{out}} \;=\; X + \gamma\,(\widehat{X} - X), \qquad \gamma \in (0,1). \tag{15}$$

When edge and curvature evidence coherently increase, the update moves in the same direction; when they conflict, $\gamma < 1$ limits the residual step size, giving the block a local contractive effect on each residual update.

Downsampling by gated reparameterization. Resolution changes use a two-branch convex combination

$$Y \;=\; (1-\sigma(g)) \cdot \mathrm{Conv}_{3\times 3}(X) \;+\; \sigma(g) \cdot \mathcal{M}_2\big(\mathrm{DW}\,\mathcal{M}_1(X)\big), \tag{16}$$

followed by normalization. Both branches are nonnegative and local, preserving the alignment of channels to edge/curvature/granularity families. Stacking such blocks makes the terminal channels naturally co-linear with the four concept families, so that "what to look for" (channels) and "where to look" (attention) live in the same geometric coordinates.

3.5 Separable Decision and Multi-Granular Attribution

Combining (4)–(5), class c admits a fully separable logit:

$$y_c \;=\; \sum_{u\in\Omega'} \sum_{d=1}^{D} A_u^{\star}\, W_{cd}\, F_d(u). \tag{17}$$

Hence three levels of attribution cohere by construction: *location-level* maps $u \mapsto A_u^{\star}\langle w_c, F_u\rangle$; *channel-level* importances $\{W_{cd}\}$; and *concept-level* marginal effects via (12). No post-hoc explainer is needed; the decomposition is intrinsic to the computation.

3.6 Learning Objective and Differentiable Closed Loop

Training uses only a standard discriminative loss $\mathcal{L}_{\mathrm{cls}}(y,t)$. Monotone calibrations are parameterized with strictly positive slopes, α lives on the simplex, and the attention is a closed-form solution of a convex subproblem. The forward pass thus already solves the selection subproblem; the backward pass propagates stable and traceable gradients along an explicit concept-to-attention-to-logit path (e.g., (12)) without learning extra attention parameters.

3.7 Alignment with Cervical Cytology

Each concept map corresponds to a principal diagnostic thread: rough nuclear envelopes (notches/indentations), layer-wise disarray in epithelium, abnormal N/C ratio, and chromatin coarsening. The diagnostic potential V stacks these evidences in space; the free-energy solution allocates the reading budget to nuclear rims and chromatin-heterogeneous interfaces where information density is highest, while avoiding mucus/debris/overlap artifacts. On the representation

side, geometry-driven blocks with nonnegative, local channel coupling keep features anchored to edge/curvature/granularity families, preventing irrelevant textures from dominating. The final decision is a linear tally of evidence with known provenance, from concept to location to channel, forming a coherent chain from pixels to logits under the semantics of cervical cytology.

4 Experiments

4.1 Datasets and Implementation Details

Datasets. We evaluated on three public cervical cytology benchmarks: DSCC [3], SIPaKMeD [16], and Herlev [9]. DSCC contains 15,509 single-cell images labeled into three diagnostic categories: normal, SIL (suggesting further examination), and ASC (requiring cytologist confirmation). SIPaKMeD includes 4,049 isolated cells across five morphological types: superficial–intermediate, parabasal, koilocytotic, dyskeratotic, and metaplastic. Herlev includes 917 Pap–smear images in seven classes: normal superficial, normal intermediate, normal columnar, light dysplastic, moderate dysplastic, severe dysplastic, and carcinoma in situ. All datasets were used for single-cell classification under 5-fold cross-validation, reporting mean±standard deviation for ACC, PRE, REC, F1, and AUC.

Implementation details. Unless otherwise stated, models were trained from scratch for 300 epochs using SGD (momentum 0.9, weight decay 5×10^{-4}, batch size 16) with a cosine-annealed learning rate of 0.004. Images were resized to 256×256, randomly cropped to 224×224 with horizontal flip for training, and center-cropped for inference. To ensure reproducibility of the explicit concepts, the nucleus mask was derived via 32-bin Otsu thresholding on the grayscale image, followed by morphological opening and closing (both using 3×3 kernels). The local N/C ratio estimated the cytoplasm area using a dilation radius of 10. Nuclear roughness normalization used a 5×5 boundary support window, while orientation disorder, local N/C ratio, and chromatin heterogeneity were computed over a 9×9 local window W_u. By default, four analytic backbone blocks produced the terminal feature map F in (5); the attention $A^\star = \mathrm{softmax}(-E/\tau)$ in (4) and the readout were computed on the downsampled lattice Ω' at 1/32 resolution of the input (with V aligned by resizing); and the convex residual update (15) started with $\gamma = 0.5$. These constitute the reference configuration used in the ablation.

4.2 Comparison with State-of-the-Art Models

Competitor groups. We benchmarked against five representative families spanning general-purpose, lightweight, medical-imaging, cervix-specific, and structurally transparent designs: (a) ConvNeXtV2-N/T [25], GCViT-B/S [7], InceptionNext-B/S [29], vHeat-B/S [24]; (b) MobileNetV4 [17], SHViT-S1 [33], LSNet [22], MobileMamba [8]; (c) MedViT [14], MedMamba [32]; (d) MSCCNet [2], RES_DCGAN [26], Hybrid Models [20], CerviNet [10], Swin-GMRS [4]; (e) CRATE [30], BDD Net [6].

Table 1: Comparison with state-of-the-art models on **DSCC**.

		Venue	ACC	F1	AUC	PRE	REC	Param	GFLOPs	LAT
(a)	CNXtV2-N	CVPR'23	85.0±1.6	83.2±1.2	94.7±1.3	83.4±1.7	83.0±1.1	15.0	2.5	5.2
	CNXtV2-T		84.7±0.6	82.6±1.0	94.6±0.2	83.2±0.6	82.4±1.2	49.6	8.7	11.6
	GCViT-B	ICML'23	83.1±1.2	80.8±1.7	93.8±1.0	81.3±1.3	80.5±1.2	89.3	14.9	20.5
	GCViT-S		82.8±1.1	80.2±1.2	93.8±0.3	81.0±0.8	79.9±1.3	50.3	8.6	20.3
	IncNeXt-B	CVPR'24	83.8±0.4	81.4±0.3	94.1±0.2	82.1±0.5	81.1±0.3	83.6	14.9	11.0
	IncNeXt-S		86.7±0.5	85.0±0.5	96.1±0.3	85.3±0.4	84.8±0.3	47.1	8.4	11.0
	vHeat-B	CVPR'25	84.7±1.4	82.7±1.3	94.6±1.1	83.1±1.5	82.5±1.4	91.1	16.2	21.3
	vHeat-S		84.6±1.2	82.4±1.0	94.6±1.3	83.1±1.0	82.1±1.1	51.5	9.2	21.1
(b)	MobileNetV4	ECCV'24	85.2±1.5	83.4±1.9	94.1±1.0	83.7±1.7	83.2±2.0	2.5	**0.2**	4.2
	SHViT-S1	CVPR'24	88.5±1.6	87.0±1.7	96.0±1.4	87.4±1.7	86.7±1.7	6.0	**0.2**	8.8
	LSNet	CVPR'25	83.3±1.5	81.1±0.8	93.8±1.6	81.8±1.6	80.8±1.2	11.0	0.3	11.7
	MobileMamba	CVPR'25	87.1±1.8	85.5±0.8	95.3±1.4	86.0±1.0	85.1±1.7	8.5	0.4	38.4
(c)	MedViT	CIBM'23	87.2±2.3	85.7±2.6	95.7±1.4	85.9±2.6	85.6±2.6	31.2	5.9	17.4
	MedMamba	arXiv'24	86.3±0.8	84.6±1.0	94.3±0.7	84.9±0.9	84.5±1.0	14.0	2.0	10.1
(d)	MSCCNet	BIBM'23	87.8±1.2	86.1±1.1	95.2±0.4	86.2±1.2	85.9±1.4	18.7	3.7	4.7
	RES_DCGAN	NCA'24	88.2±1.1	88.0±1.1	96.6±0.5	88.1±0.9	88.0±1.2	16.2	4.5	5.7
	Hybrid Models	BSPC'25	88.2±1.6	86.8±1.8	96.0±0.5	87.3±1.9	86.4±1.8	57.5	23.6	6.1
	CerviNet	JCB'25	86.9±1.4	85.1±1.9	95.6±1.1	85.8±1.1	84.7±1.2	85.9	17.0	5.3
	Swin-GMRS	J-BHI'25	89.5±2.1	91.3±1.8	96.9±1.0	88.4±2.4	88.2±2.4	100.8	8.8	60.0
(e)	CRATE	JMLR'24	84.6±0.5	82.4±0.7	94.5±0.3	83.1±0.6	82.0±0.7	15.6	2.9	5.0
	BDD Net	TIM'25	81.5±0.4	78.8±0.4	92.5±0.2	79.5±0.3	78.5±0.5	2.6	1.9	**0.2**
Ours		ECCV'26	**93.2±0.2**	**92.6±0.2**	**98.7±0.1**	**92.7±0.3**	**92.4±0.3**	**2.2**	0.4	5.4

Results are summarized in Tabs. 1 and 2; complexity and latency are reported in Tab. 1 (NVIDIA Tesla V100, Intel Core i9–10980XE).

Baseline training protocol. All baselines and our model were trained from scratch using identical 5-fold splits and the same recipe. ImageNet pre-training is standard for natural images, but cervical cytology has a large domain gap (e.g., cell morphologies, overlapping clusters, staining artifacts), and from-scratch training can match or outperform fine-tuning due to negative transfer [21]. This unified protocol isolates intrinsic architectural suitability rather than reliance on external data or disparate tuning budgets.

Classification performance. Across DSCC, Herlev, and SIPaKMeD, GeoCEAN achieves state-of-the-art or tied state-of-the-art results (Tabs. 1 and 2). It consistently ranks at the top across datasets; the strongest baseline is typically Swin-GMRS, while other competitive models trail on at least one benchmark.

Across-dataset stability. Several baselines vary sharply across datasets (e.g., InceptionNext-S: 86.7 ACC on DSCC vs. 52.9 on Herlev), whereas GeoCEAN ranks at the top on all three, indicating robustness to staining, label definitions, and scale.

Model size and efficiency. GeoCEAN uses 2.2M parameters with 0.4 GFLOPs and 5.4 ms latency (Tab. 1), delivering top accuracy at orders-of-magnitude lower complexity than large general backbones and remaining competitive with lightweight models. BDD Net is faster (0.2 ms) but with a large accuracy gap on DSCC/Herlev/SIPaKMeD.

Table 2: Comparison with state-of-the-art models on **Herlev** and **SIPaKMeD**.

		Herlev					SIPaKMeD				
		ACC	F1	AUC	PRE	REC	ACC	F1	AUC	PRE	REC
(a)	CNXtV2-N	67.8±5.3	70.2±4.6	91.8±2.5	70.8±5.4	70.9±4.9	96.0±0.9	96.0±0.9	99.5±0.2	96.0±0.9	96.0±0.9
	CNXtV2-T	69.7±6.5	71.8±6.0	92.3±2.3	73.3±6.2	71.8±5.8	96.1±0.7	96.1±0.7	99.5±0.2	96.1±0.7	96.1±0.7
	GCViT-B	67.4±3.5	70.5±3.2	91.2±2.8	71.0±5.4	70.4±3.9	96.7±0.9	96.7±0.9	99.6±0.2	96.7±0.7	96.7±0.8
	GCViT-S	66.8±2.9	69.6±2.5	90.5±2.5	70.5±3.3	69.7±4.3	96.0±0.8	96.0±0.9	99.4±0.2	96.0±0.6	96.0±0.8
	IncNeXt-B	53.1±2.9	54.4±2.4	83.8±2.4	55.7±1.7	55.1±2.3	88.7±3.5	88.7±3.5	98.2±0.8	88.8±3.5	88.8±3.4
	IncNeXt-S	52.9±3.0	54.4±2.4	84.0±2.1	55.7±2.9	54.6±1.9	78.3±0.3	78.2±0.3	94.9±0.3	78.8±0.2	78.4±0.3
	vHeat-B	73.9±4.1	75.9±4.2	93.9±2.7	77.0±5.3	75.6±4.2	96.7±0.5	96.7±0.5	99.6±0.1	96.7±0.5	96.7±0.5
	vHeat-S	71.9±3.5	74.1±3.9	92.3±2.5	75.1±5.2	73.8±3.9	96.1±0.9	96.2±1.0	99.5±0.1	96.2±1.0	96.2±0.9
(b)	MobileNetV4	65.9±3.3	68.6±3.0	90.4±2.1	69.9±4.4	68.5±3.8	96.0±1.0	96.0±1.0	99.6±0.1	96.0±1.0	96.0±1.0
	SHViT-S1	71.4±4.0	74.2±3.5	92.8±2.6	75.4±4.6	74.0±3.8	97.3±0.9	97.3±0.9	99.7±0.1	97.3±0.9	97.3±0.9
	LSNet	65.9±5.8	69.1±4.4	89.6±3.1	69.7±5.8	69.2±4.3	96.0±0.7	96.0±0.7	99.6±0.1	96.0±0.7	96.0±0.7
	MobileMamba	72.4±5.1	75.6±4.7	93.0±2.8	76.6±5.3	75.5±4.1	97.3±1.0	97.3±1.0	99.7±0.1	97.3±1.0	97.3±1.0
(c)	MedViT	69.5±3.9	72.3±4.5	92.4±2.3	73.4±5.1	72.0±4.8	97.9±0.4	98.0±0.4	99.7±0.1	98.0±0.4	98.0±0.4
	MedMamba	73.5±3.0	75.5±3.8	93.8±1.3	75.9±3.9	75.9±4.4	97.5±0.8	97.5±0.8	99.7±0.1	97.5±0.8	97.5±0.8
(d)	MSCCNet	71.0±4.1	72.3±4.3	92.0±1.0	73.4±3.9	71.8±4.0	97.3±0.6	97.3±0.7	99.7±0.1	97.3±0.5	97.3±0.4
	RES_DCGAN	73.1±4.3	75.4±4.1	93.4±1.8	76.2±2.1	75.4±4.7	97.1±0.6	97.1±0.6	99.7±0.1	97.1±0.4	97.1±0.6
	Hybrid Models	68.6±5.7	71.0±5.8	91.5±2.8	72.7±5.9	70.9±5.6	96.5±0.7	96.5±0.7	99.6±0.1	96.5±0.7	96.5±0.6
	CerviNet	70.5±4.7	73.3±5.0	93.0±2.7	73.6±4.9	73.6±5.0	96.9±0.4	96.9±0.6	99.6±0.1	96.9±0.9	97.0±1.0
	Swin-GMRS	73.5±3.1	75.5±3.1	93.3±0.9	76.7±2.8	76.2±3.6	98.3±0.5	98.9±0.3	99.9±0.0	98.3±0.4	98.3±0.5
(e)	CRATE	66.3±3.2	68.4±3.1	90.8±0.9	68.2±3.4	69.3±2.7	96.5±1.0	96.4±1.1	99.5±0.2	96.6±0.9	96.4±1.0
	BDD Net	43.4±2.8	38.8±3.0	78.8±4.2	48.9±6.9	39.5±2.9	90.0±0.8	90.0±0.8	97.4±0.5	90.0±0.8	90.1±0.8
Ours		**76.8±2.4**	**79.1±3.5**	**94.3±1.3**	**79.7±3.4**	**79.1±3.2**	**98.5±0.5**	**98.5±0.4**	**99.9±0.0**	**98.5±0.4**	**98.5±0.4**

Table 3: Concept-to-potential monotonicity across datasets. For each clinical concept i (Sec. 3), we report (i) the sign agreement rate $\mathrm{SignRate}\big(z_i \nearrow\!\!\!\!\!\!\Rightarrow V \nearrow\big)$ computed on aligned lattices and (ii) Spearman's rank correlation $\rho(z_i, V)$. Larger values indicate stronger monotonic consistency.

Concept	DSCC		Herlev		SIPaKMeD	
	SignRate $\big(z_i \nearrow\!\!\!\!\!\!\Rightarrow V \nearrow\big)$	$\rho(z_i, V)$	SignRate $\big(z_i \nearrow\!\!\!\!\!\!\Rightarrow V \nearrow\big)$	$\rho(z_i, V)$	SignRate $\big(z_i \nearrow\!\!\!\!\!\!\Rightarrow V \nearrow\big)$	$\rho(z_i, V)$
Roughness (nuclear envelope)	0.425	0.260	0.871	0.835	0.902	0.772
Orientation disorder	1.000	0.998	0.996	0.992	0.998	0.994
Local N/C ratio	0.765	0.695	0.757	0.683	0.778	0.702
Chromatin heterogeneity	0.992	0.743	0.971	0.729	0.983	0.736

4.3 Task-Aligned Transparency: From Concepts to Energy, Attention, and Decisions

Intrinsic panels. Fig. 4 visualizes quantities already computed along the forward path in Sec. 3, one row per backbone stage (last AnalyticBlock), with the input shown once (upsampled to 2×, vertically centered). The right columns show: (i) attention–update coupling $A^\star \cdot U$, where $A^\star = \mathrm{softmax}(-E/\tau)$ from Eq. (4) and $U = \|X^{\mathrm{out}} - X\|_1$; we render

$$\widetilde{A \cdot U}(u) = \frac{A^\star_u}{\max_v A^\star_v} \cdot \frac{U(u)}{\max_v U(v)}, \tag{18}$$

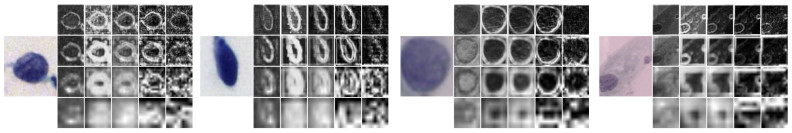

Fig. 4: Intrinsic evidence panels produced by the model. Each row corresponds to the last AnalyticBlock of a backbone stage. Columns are: input (shown once at 2×, vertically centered), $A^\star \cdot U$ (attention–update coupling), $\cos(F, \Delta X)$ (alignment between geometry and residual motion), S_{gm} (finite-difference sensitivity to the gradient-magnitude calibration a_{gm}), and top-k averages of M_{gm} and M_{lap}. All tiles are grayscale and independently min–max normalized. Panels reveal that reading budget and update magnitude concentrate on the deepest potential wells (negative energy) at nuclear rims and peri-nuclear halos, with positive cone alignment and high geometric sensitivity at those loci, matching routine cytological reading.

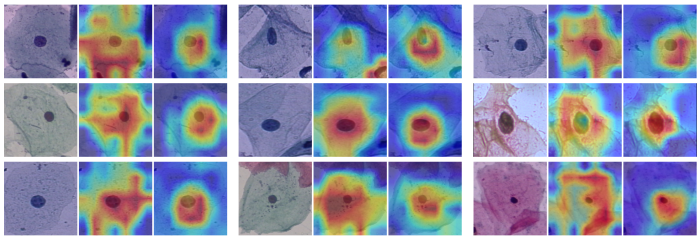

Fig. 5: Our potential-guided attention vs. Swin-GMRS Grad-CAM. Both peak near nuclei, but Grad-CAM often diffuses into cytoplasm/nearby stain, whereas ours concentrates on the deepest potential wells with smooth peri-nuclear falloff.

to show where reading budget is spent and how much the state moves; (ii) alignment between geometric responses and residual motion,

$$\cos(F, \Delta X)(u) = \frac{\langle F_{\mathrm{geo}}(u),\, \Delta X(u)\rangle}{\|F_{\mathrm{geo}}(u)\|_2 \, \|\Delta X(u)\|_2} \in [-1, 1], \tag{19}$$

with F_{geo} from Eqs. (13)–(14) and $\Delta X = \widehat{X} - X$ used in Eq. (15); (iii) finite-difference sensitivity to the gradient-magnitude calibration $a_{\mathrm{gm}} > 0$,

$$S_{\mathrm{gm}}(u) \approx \frac{\left\| \widehat{X}(a_{\mathrm{gm}}+\epsilon)(u) - \widehat{X}(a_{\mathrm{gm}}-\epsilon)(u) \right\|_1}{2\epsilon}; \tag{20}$$

and (iv–v) top-k channel averages of the gradient-magnitude and Laplacian maps from Eq. (13),

$$\overline{M}_\bullet(u) = \frac{1}{k} \sum_{d \in \mathrm{Top}k} M_{\bullet,d}(u), \qquad \bullet \in \{\mathrm{gm}, \mathrm{lap}\}. \tag{21}$$

Panels concentrate on nuclear rims and peri-nuclear halos, where envelope roughness and chromatin heterogeneity appear, consistent with the analytic backbone: nonnegative near-diagonal mixing and convex residual updates yield positive $\cos(F, \Delta X)$ at informative rims; shallow stages emphasize fine edge/curvature,

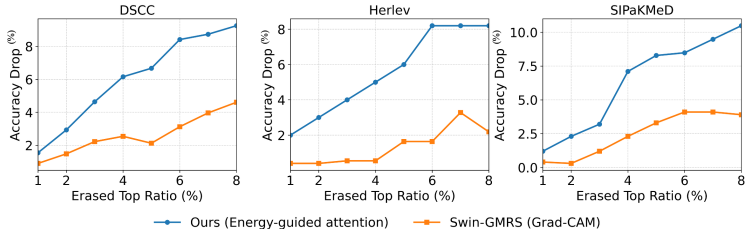

Fig. 6: Quantitative consistency of potential-guided attention (from V). Overlays highlight diagnostic regions; on a frozen RES_DCGAN evaluator, erasing our top-ranked pixels causes larger accuracy drops than erasing Swin-GMRS Grad-CAM pixels.

while deeper stages form stable ring-like patterns at lower resolution due to gated downsampling.

Comparison with External Post-hoc Explainers. Fig. 5 contrasts Swin-GMRS Grad-CAM [19] with our intrinsic potential-guided attention under a matched protocol. While both highlight nuclei, Grad-CAM frequently spreads into cytoplasm or staining, whereas our closed-form attention concentrates on the deepest diagnostic wells with a nucleus-centered, peri-nuclear decay pattern. Fig. 6 corroborates this: with a frozen RES_DCGAN evaluator, erasing our top-ranked pixels induces a steeper accuracy drop than erasing Grad-CAM pixels, indicating more faithful capture of critical evidence.

Concept-to-energy monotonicity. Given

$$V(u) = \sum_{i=1}^{4} \alpha_i \, f_i\big(z_i(u)\big), \qquad \alpha_i \geq 0, \tag{22}$$

$$f_i'(z) > 0.$$

the model implies $\partial V(u)/\partial z_i(u) \geq 0$ pointwise. We tested this implication on an aligned lattice (resizing V to Ω' if needed) using two complementary statistics.
Local sign agreement. Let $\mathcal{S}$ be short-range pixel pairs on Ω' (4-neighbors). We compute

$$\mathrm{SignRate}_i \;=\; \frac{1}{|\mathcal{S}|} \sum_{(u,v)\in\mathcal{S}} \mathbf{1}\Big((z_i(u) - z_i(v)) \cdot (V(u) - V(v)) \geq 0\Big). \tag{23}$$

which tests whether increases in roughness, orientation disorder, local N/C ratio, or chromatin heterogeneity tend not to decrease local evidence.
Rank-level agreement. Because f_i can rescale z_i, we also compute Spearman's rank correlation on the same lattice,

$$\rho_i \;=\; \rho\big(z_i,\, V\big) \;=\; \mathrm{corr}\big(\mathrm{rank}(z_i(u)),\, \mathrm{rank}(V(u))\big), \tag{24}$$

capturing monotone consistency independent of global scaling.

Tab. 3 reports $\mathrm{SignRate}_i$ and ρ_i on DSCC/Herlev/SIPaKMeD for the four concepts in Sec. 3. Orientation disorder and chromatin heterogeneity are near-ideal across datasets, matching the nucleus-centered patterns in Fig. 4; local

Table 4: Cross-dataset generalization (DSCC→Herlev).

Method	Acc (%)	F1 (%)	Rec (%)	AUC (%)	Pre (%)
Ours	74.0±3.8	82.2±3.2	82.8±5.4	74.1±3.5	82.4±1.1
Swin-GMRS	70.2±4.7	80.3±4.5	82.1±5.5	66.8±4.3	78.0±1.7
RES_DCGAN	59.8±5.3	65.5±5.1	71.6±9.3	58.2±5.4	63.9±3.4

Table 5: Robustness under stain/contrast shifts and parameter perturbations.

Dataset	ΔAcc(M)	Corr(M)	ΔAcc(S)	Corr(S)	DCorr
DSCC	−0.08±0.48%	0.86±0.00	−7.13±1.35%	0.77±0.00	0.98±0.00
Herlev	−3.04±1.87%	0.81±0.01	−9.92±2.09%	0.79±0.01	0.97±0.00
SIP	−2.59±0.42%	0.75±0.00	−8.89±1.52%	0.67±0.01	0.98±0.00

N/C ratio is also consistently high. DSCC shows lower roughness consistency (0.425/0.260), likely due to boundary-mask quality and stain noise that destabilize the boundary-Laplacian flux in (6); Herlev and SIPaKMeD approach ideal values. All metrics are computed after resolution alignment and per-image min–max normalization, reflecting sign/rank relations rather than absolute magnitudes. Together, intrinsic panels, external CAM references, and monotonicity checks support the same reading mechanism: evidence accumulates at nuclear rims and peri-nuclear interfaces, attention allocates budget via Eq. (4), and clinical concepts act nondecreasingly on local diagnostic potential via (22).

Cross-dataset generalization. We performed a zero-shot test where models trained on DSCC were evaluated on Herlev. Due to differing taxonomies, labels were merged into a binary task (normal vs. non-normal). As shown in Tab. 4, our model outperforms Swin-GMRS and RES_DCGAN (74.0% vs. 70.2% and 59.8% ACC), indicating improved robustness to dataset-specific biases by anchoring spatial allocation to explicit, domain-invariant diagnostic potentials.

Robustness of concept extraction. We evaluated stability under imaging shifts and hyperparameter perturbations. Under Mild (M) and Strong (S) stain/contrast shifts (Tab. 5), the model is resilient to M-shifts (e.g., ΔAcc $\approx -0.08\%$ on DSCC). For S-shifts, applying M-shifts as training augmentation with an L_1 consistency loss largely suppresses degradation ($\Delta\text{Acc} \geq -0.5\%$). Perturbing core morphological hyperparameters (dilation and window size) by discrete integer steps $(\pm 1, \pm 2, \pm 3)$ yields near-perfect map correlation (DCorr ≥ 0.97), suggesting the diagnostic potentials capture stable cytological structure rather than brittle algorithmic artifacts.

4.4 Ablation Studies

We ablate five design choices on SIPaKMeD (Tab. 6) using the same protocol as the main experiments: (i) number of backbone blocks, (ii) attention/readout lattice resolution Ω', (iii) residual step γ, (iv) local window W_u, and (v) the four explicit concepts with the nonnegative constraint. Performance saturates at four blocks. A 1/32 lattice is optimal; overly small strides collapse because an overly fine lattice provides insufficient local support per cell, yielding unreliable

Table 6: Results of ablation studies on SIPaKMeD.

Setting	Value	ACC	F1	AUC	PRE	REC
Stages	1	82.4±0.6	82.1±0.5	96.3±0.1	82.1±0.5	82.1±0.5
	2	97.3±0.7	97.3±0.6	99.1±0.2	97.3±0.6	97.3±0.6
	3	96.5±0.8	96.6±0.7	98.9±0.3	96.6±0.7	96.6±0.7
	4	**98.5±0.5**	**98.5±0.4**	**99.9±0.0**	**98.5±0.4**	**98.5±0.4**
	5	97.1±0.8	97.0±0.7	98.7±0.3	97.0±0.7	97.0±0.7
Attention stride	4	25.8±0.6	15.3±0.5	83.6±0.1	15.3±0.5	15.3±0.5
	8	46.2±0.7	35.7±0.6	85.1±0.2	35.7±0.6	35.7±0.6
	16	97.0±0.8	97.0±0.7	99.0±0.3	97.0±0.7	97.0±0.7
	32	**98.5±0.5**	**98.5±0.4**	**99.9±0.0**	**98.5±0.4**	**98.5±0.4**
	64	97.2±0.7	97.1±0.6	98.8±0.2	97.1±0.6	97.1±0.6
Residual step (γ)	0.2	97.8±0.6	97.7±0.5	99.4±0.1	97.7±0.5	97.7±0.5
	0.5	**98.5±0.5**	**98.5±0.4**	**99.9±0.0**	**98.5±0.4**	**98.5±0.4**
	0.8	97.0±0.7	97.0±0.6	99.1±0.2	97.0±0.6	97.0±0.6
Local window size	5	97.8±0.6	97.7±0.5	99.4±0.1	97.7±0.5	97.7±0.5
	7	97.4±0.7	97.4±0.6	99.2±0.2	97.4±0.6	97.4±0.6
	9	**98.5±0.5**	**98.5±0.4**	**99.9±0.0**	**98.5±0.4**	**98.5±0.4**
	11	96.8±0.8	96.8±0.7	98.9±0.3	96.8±0.7	96.8±0.7
Concepts & Design	w/o nonneg.	95.9±0.7	95.6±0.8	98.7±0.4	96.1±0.6	95.8±0.9
	w/o N/C	95.3±0.8	94.9±0.9	98.3±0.5	95.5±0.7	95.1±1.0
	w/o rough.	96.4±0.6	96.2±0.7	98.9±0.3	96.6±0.6	96.3±0.8
	w/o orient.	96.9±0.5	96.7±0.6	99.1±0.3	97.1±0.5	96.8±0.7
	w/o chrom.	97.8±0.3	97.6±0.4	99.5±0.2	97.8±0.3	97.7±0.4

potentials and unstable softmax aggregation. A 9×9 window best balances optimization stability and cue estimation. Concept ablations confirm necessity, with local N/C ratio most critical, and the nonnegative calibration essential to prevent rank inversion (higher abnormality monotonically deepens the diagnostic potential well without negative cancellation) for reliable attention routing.

5 Conclusion

We present a compact and structurally transparent framework for cervical cytology classification that mirrors clinical workflow: a free–energy attention determines *where to look* and a geometry–driven analytic backbone explains *what is present*. Attention is the closed-form optimum of a free-energy objective, while the backbone enforces nonnegative near-diagonal mixing, monotone calibration, and convex residual updates for stable and auditable computation. Despite only 2.2M parameters, the model matches or exceeds larger general-purpose and medical backbones, indicating that structural transparency can improve both interpretability and efficiency.

Limitations and future work. GeoCEAN makes its evidence chain inspectable through concept maps, potential-guided attention, and evidence-to-logit attributions. While we evaluate these outputs by internal consistency, cytology-relevant localization, erasure, monotonicity, and cross-dataset tests, systematic cytopathologist validation remains future work. We will use these outputs as expert-assessment targets and extend GeoCEAN to whole-slide screening with additional cytological priors.

References

1. Bychkov, A., Yoshikawa, A., Munkhdelger, J., Hori, T., Fukuoka, J.: Integrating cytology into routine digital pathology workflow: A 5-year journey. Virchows Archiv **483**(4), 555–559 (2023)
2. Chen, H., Liu, J., Jiang, P., Jin, Y., Feng, J., Pang, B., Cao, D., Li, C.: Msccnet: Multi-scale convolution-capsule network for cervical cell classification. In: 2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). pp. 1842–1845 (2023)
3. Chen, H., Liu, J., Jin, Y., Pang, B., Cao, D., Xiao, D.: Dscc: a data set of cervical cell images for cervical cytology screening. International Journal of Data Mining and Bioinformatics **27**(1-3), 171–186 (2022)
4. Chen, Y., Jiang, H., Pang, W., Diao, Z., Yang, J.: Multi-task collaborative assisted training method for grouping fuzzy categories classification of cervical cancer cells. IEEE Journal of Biomedical and Health Informatics **29**(9), 6699–6712 (2025)
5. Fei, M., Shen, Z., Liu, M., Song, Z., Sun, Y., Han, X., Liu, Z., Jiang, H., Bai, L., Wang, Q., et al.: Refining cervical cell classification with cytological knowledge and optimal attribute descriptor matching. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 519–529 (2025)
6. Guan, C., Wang, W., Ai, H., Singh, R.P.: A white-box module for medical image multi-classification via bias dendrite net. IEEE Transactions on Instrumentation and Measurement **74**, 1–12 (2025)
7. Hatamizadeh, A., Yin, H., Heinrich, G., Kautz, J., Molchanov, P.: Global context vision transformers. In: International Conference on Machine Learning. pp. 12633–12646 (2023)
8. He, H., Zhang, J., Cai, Y., Chen, H., Hu, X., Gan, Z., Wang, Y., Wang, C., Wu, Y., Xie, L.: Mobilemamba: Lightweight multi-receptive visual mamba network. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 4497–4507 (2025)
9. Jantzen, J., Norup, J., Dounias, G., Bjerregaard, B.: Pap-smear benchmark data for pattern classification. Nature inspired smart information systems pp. 1–9 (2005)
10. Khowaja, A., Beiji, Z., Kui, X.: Cervinet: A novel approach for cervical cancer classification using pap-smear images. Journal of Computational Biology **32**(10), 974–985 (2025)
11. Liang, Y., Feng, S., Liu, Q., Kuang, H., Liu, J., Liao, L., Du, Y., Wang, J.: Exploring contextual relationships for cervical abnormal cell detection. IEEE Journal of Biomedical and Health Informatics **27**(8), 4086–4097 (2023)
12. Liu, G., Wang, J.: Dendrite net: A white-box module for classification, regression, and system identification. IEEE Transactions on Cybernetics **52**(12), 13774–13787 (2022)
13. Liu, Y., Wang, Z., Dai, A., Gu, W.: A novel ensemble framework based on cnn models and swin transformer for cervical cytology image classification. In: Seventh International Conference on Computer Graphics and Virtuality. vol. 13158, pp. 150–161 (2024)
14. Manzari, O.N., Ahmadabadi, H., Kashiani, H., Shokouhi, S.B., Ayatollahi, A.: Medvit: a robust vision transformer for generalized medical image classification. Computers in Biology and Medicine **157**, 106791 (2023)
15. Pacal, I.: Maxcervixt: A novel lightweight vision transformer-based approach for precise cervical cancer detection. Knowledge-Based Systems **289**, 111482 (2024)

16. Plissiti, M.E., Dimitrakopoulos, P., Sfikas, G., Nikou, C., Krikoni, O., Charchanti, A.: Sipakmed: A new dataset for feature and image based classification of normal and pathological cervical cells in pap smear images. In: Proceedings of the 25th ieee international conference on image processing. pp. 3144–3148 (2018)
17. Qin, D., Leichner, C., Delakis, M., Fornoni, M., Luo, S., Yang, F., Wang, W., Banbury, C., Ye, C., Akin, B., et al.: Mobilenetv4: Universal models for the mobile ecosystem. In: Proceedings of the european conference on computer vision. pp. 78–96 (2024)
18. Qin, J., He, Y., Ge, J., Liang, Y.: A multi-task feature fusion model for cervical cell classification. IEEE journal of biomedical and health informatics **26**(9), 4668–4678 (2022)
19. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)
20. Sharma, A.K., Nandal, A., Dhaka, A., Alhudhaif, A., Polat, K., Sharma, A.: Diagnosis of cervical cancer using cnn deep learning model with transfer learning approaches. Biomedical Signal Processing and Control **105**, 107639 (2025)
21. Tan, S.L., Selvachandran, G., Ding, W., Paramesran, R., Kotecha, K.: Cervical cancer classification from pap smear images using deep convolutional neural network models. Interdisciplinary Sciences: Computational Life Sciences **16**(1), 16–38 (2024)
22. Wang, A., Chen, H., Lin, Z., Han, J., Ding, G.: Lsnet: See large, focus small. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 9718–9729 (2025)
23. Wang, J., Yu, Y., Tan, Y., Wan, H., Zheng, N., He, Z., Mao, L., Ren, W., Chen, K., Lin, Z., et al.: Artificial intelligence enables precision diagnosis of cervical cytology grades and cervical cancer. Nature Communications **15**(1), 4369 (2024)
24. Wang, Z., Liu, Y., Tian, Y., Liu, Y., Wang, Y., Ye, Q.: Building vision models upon heat conduction. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 9707–9717 (2025)
25. Woo, S., Debnath, S., Hu, R., Chen, X., Liu, Z., Kweon, I.S., Xie, S.: Convnext v2: Co-designing and scaling convnets with masked autoencoders. In: Proceedings of the ieee/cvf conference on computer vision and pattern recognition. pp. 16133–16142 (2023)
26. Wubineh, B.Z., Rusiecki, A., Halawa, K.: Classification of cervical cells from the pap smear image using the res_dcgan data augmentation and resnet50v2 with self-attention architecture. Neural Computing and Applications **36**(34), 21801–21815 (2024)
27. Xu, L., Cai, F., Fu, Y., Liu, Q.: Cervical cell classification with deep-learning algorithms. Medical & Biological Engineering & Computing **61**(3), 821–833 (2023)
28. Yaman, O., Tuncer, T.: Exemplar pyramid deep feature extraction based cervical cancer image classification model using pap-smear images. Biomedical Signal Processing and Control **73**, 103428 (2022)
29. Yu, W., Zhou, P., Yan, S., Wang, X.: Inceptionnext: When inception meets convnext. In: Proceedings of the ieee/cvf conference on computer vision and pattern recognition. pp. 5672–5683 (2024)
30. Yu, Y., Buchanan, S., Pai, D., Chu, T., Wu, Z., Tong, S., Bai, H., Zhai, Y., Haeffele, B.D., Ma, Y.: White-box transformers via sparse rate reduction: Compression is all there is? Journal of Machine Learning Research **25**(300), 1–128 (2024)

31. Yu, Y., Chu, T., Tong, S., Wu, Z., Pai, D., Buchanan, S., Ma, Y.: Emergence of segmentation with minimalistic white-box transformers. In: Conference on Parsimony and Learning. pp. 72–93 (2024)
32. Yue, Y., Li, Z.: Medmamba: Vision mamba for medical image classification. arXiv preprint arXiv:2403.03849 pp. 1–26 (2024)
33. Yun, S., Ro, Y.: Shvit: Single-head vision transformer with memory efficient macro design. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5756–5767 (2024)
34. Zhu, L., Wang, X., Ke, Z., Zhang, W., Lau, R.W.: Biformer: Vision transformer with bi-level routing attention. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10323–10333 (2023)