Learning Class Difficulty via Dynamic Focal Attention for Histopathology Segmentation

Lakmali Nadeesha Kumari Rathukohe Mudiyanselage and Sen-Ching Samson Cheung

University of Kentucky, Lexington, KY 40506, USA
{lakmali.nadeesha, sen-ching.cheung}@uky.edu

Abstract. Frequency-based loss reweighting, the standard remedy for imbalanced histopathology segmentation, implicitly assumes that rare classes are difficult. Yet difficulty also arises from morphological variability, boundary ambiguity, and contextual similarity, all largely orthogonal to class frequency. We propose *Dynamic Focal Attention* (DFA), a simple, efficient mechanism that learns class-specific difficulty directly within the cross-attention of query-based mask decoders. DFA adds a learnable per-class bias to the attention logits, reweighting representations *before* prediction rather than gradients *after* it. Initialised from a centred log-frequency prior to prevent gradient starvation and then optimised end-to-end, the bias adapts to difficulty signals as they emerge during training, unifying frequency- and difficulty-aware reweighting in a single attention-bias framework. On three benchmarks (BCSS, BDSA, CRAG), DFA consistently improves Dice and IoU, matching or exceeding a two-stage difficulty-aware baseline without a separate estimator or extra training stage. This shows that encoding difficulty at the representation level is a principled alternative to conventional loss reweighting.

Keywords: Class difficulty · Attention mechanism · Pathology segmentation · Transformer · Difficulty-aware learning

1 Introduction

Semantic segmentation of histopathology images underpins tissue characterisation, tumour delineation, and biomarker extraction in computational pathology [3,16,12]. Transformer-based architectures [7,6] perform strongly overall yet underperform on clinically critical classes, a shortfall typically attributed to class imbalance. But imbalance is only part of the story: class imbalance is a dataset-level frequency property, whereas class difficulty is governed by morphological variability, boundary ambiguity, and contextual similarity.

Limitations of existing approaches. Data-level strategies [2,14] alter training distributions without adding semantic information, while loss-level methods [15,21,8,20] reweight gradient magnitudes *after* prediction, once class-specific representations are already formed. All treat rarity as a proxy for difficulty, yet the two are largely orthogonal [2,11] (Section 3.2). Complementary pathology

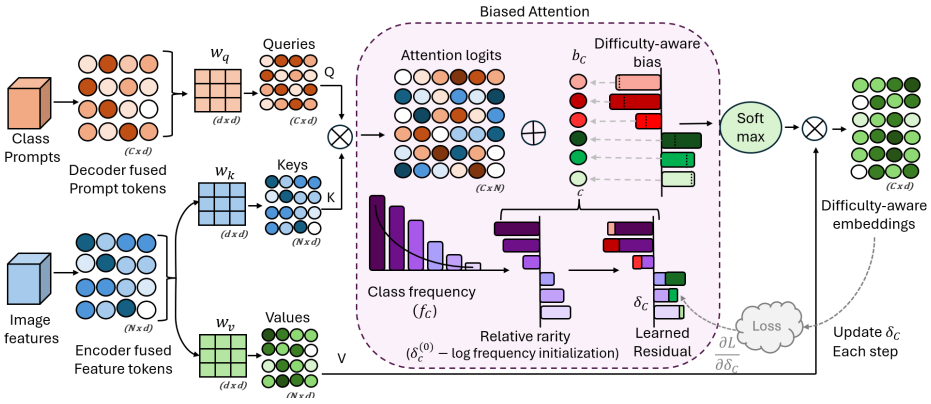

Fig. 1: **DFA integrated into cross-attention.** A learnable class-specific bias b_c is added to attention logits before softmax, initialised from a log-frequency prior and optimised end-to-end via $\mathcal{L}$ (Eq. 7). Here $\otimes$ denotes matrix multiplication and $\oplus$ broadcast addition of $\mathbf{b} \in \mathbb{R}^C$ across N spatial positions.

methods target *spatial* difficulty, with HoVer-Net [12] exploiting distance maps and MILD-Net [10] employing multi-scale dilated context, yet neither supplies an explicit *class-level* difficulty signal during feature aggregation.

Attention as a principled intervention point. SAM [13] and SAM-Path [22,5] recast segmentation as mask classification, with learnable class queries interacting with image features through cross-attention [4,6]—directly shaping per-class feature aggregation and providing a natural site for difficulty-aware intervention. Yet standard attention treats all queries uniformly. Structured logit biases [19,18,23] help when task-specific priors are available, but none encode *class-level semantic difficulty*, which is crucial for the diverse challenges of histopathology segmentation.

Our approach. We propose **Dynamic Focal Attention (DFA)**, which injects a learnable class-specific scalar bias b_c into cross-attention logits (Fig. 1), enabling difficulty-aware feature aggregation at the representation level. Rather than reweighting gradients after prediction, DFA adjusts logits class-conditionally within attention, modulating how strongly features are aggregated per class. To prevent early gradient starvation on rare classes, b_c is initialised from a centred log-frequency prior and thereafter updated by the segmentation objective to emphasise difficult-to-segment classes. We also formalise two variants, class-frequency (CFFA) and hard-class (HCFA), giving a progression from static to dynamic difficulty estimation. DFA runs in a single stage, adds one scalar per class, and needs no pre-trained baseline.

Contributions. (*1*) We show class frequency is an unreliable difficulty proxy, with direct rarity–difficulty inversions across three pathology benchmarks. (*2*) We propose the *Focal Attention Bias* framework, unifying CFFA, HCFA, and DFA as a progressive family of class-difficulty-aware attention mechanisms. (*3*) DFA

achieves up to 15.2 Dice points improvement on difficult classes and ∼40% training-time reduction versus two-stage approaches. The code is publicly available at `https://github.com/lakmali240/DFA-Imbalance`.

2 Method

2.1 Preliminaries: SAM-Path Decoder

We build on SAM-Path [22], which extends SAM [13] with two encoders, a frozen SAM encoder and a pathology foundation model, whose outputs align into a unified feature $\mathbf{H} \in \mathbb{R}^{N \times d}$ (N spatial tokens, d feature dimension). The decoder keeps C learnable class prompt tokens $\mathcal{P} \in \mathbb{R}^{C \times d}$ that interact with $\mathbf{H}$ via cross-attention to produce per-class masks $\hat{y}_c$ and IoU predictions $\widehat{\text{iou}}_c$, $c = 1, \ldots, C$. The attention weights $\alpha_{c,i} = \exp(s_{c,i}) / \sum_{c'} \exp(s_{c',i})$ with $s_{c,i} = (\mathbf{Q}\mathbf{K}^\top)_{c,i}/\sqrt{d}$, where $\mathbf{Q} \in \mathbb{R}^{C \times d}$ and $\mathbf{K} \in \mathbb{R}^{N \times d}$ derive from $\mathcal{P}$ and $\mathbf{H}$, depend only on feature–prompt similarity, with no mechanism to modulate aggregation by class-level complexity, which is the limitation we address. Our formulation applies broadly to cross-attention in transformer-based segmentation models beyond SAM-Path.

2.2 Dynamic Focal Attention

Focal Attention Bias framework. We introduce a class-specific scalar $b_c \in \mathbb{R}$, broadcast across all N spatial positions and added to the cross-attention logit for class c before a class-wise softmax at each spatial location:

$$\tilde{\alpha}_{c,i} = \frac{\exp(s_{c,i} + b_c)}{\sum_{c'=1}^{C} \exp(s_{c',i} + b_{c'})}, \quad \widetilde{\text{Attn}}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\Big(\tfrac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}} \oplus \mathbf{b} \Big) \mathbf{V}, \quad (1)$$

where $\mathbf{b} = [b_1, \ldots, b_C]^\top \in \mathbb{R}^C$ and $\mathbf{V} \in \mathbb{R}^{N \times d}$ are values from $\mathcal{P}$. The operator $\oplus$ denotes broadcast addition of $\mathbf{b}$ across N spatial positions, following structured attention biases such as ALiBi [18], T5 [19], and Deformable DETR [23]. Setting $b_c = \log \omega_c$ ($\omega_c > 0$) multiplicatively rescales the c-th score by ω_c uniformly across all aggregations; parameterising b_c yields a progressive family of strategies.

(S1) Class Frequency Focal Attention (CFFA). CFFA sets b_c inversely proportional to pixel frequency:

$$b_c = \gamma \, \log(1 - f_c), \qquad \gamma > 0, \quad (2)$$

where $f_c \in (0, 1)$ is the normalised training-set pixel frequency of class c and γ controls magnitude. Unlike output-level logit adjustment [17], this prior acts on feature aggregation directly, but still conflates rarity with difficulty (Section 3.2).

(S2) Hard Class Focal Attention (HCFA). HCFA replaces frequency with empirical difficulty estimated from baseline performance:

$$b_c = \gamma \, \log(1 - p_c), \quad p_c = \frac{\text{Dice}_c^{\text{base}}}{\sum_{c'=1}^{C} \text{Dice}_{c'}^{\text{base}}}, \quad (3)$$

where $\mathrm{Dice}_c^{\mathrm{base}}$ is the per-class validation Dice of the pre-trained baseline and p_c its normalised difficulty proxy; HCFA targets hard classes but requires a full pre-training run.

(S3) Dynamic Focal Attention (DFA)—Proposed. CFFA and HCFA fix b_c before training and cannot adapt to difficulty signals that emerge during optimisation. DFA instead treats b_c as a learnable scalar $\delta_c \in \mathbb{R}$ optimised end-to-end:

$$b_c = \delta_c, \qquad \boldsymbol{\delta} \in \mathbb{R}^C, \tag{4}$$

$$\left.\frac{\partial \tilde{\alpha}_{c,i}}{\partial \delta_c}\right|_{\text{on-diag.}} = \underbrace{\tilde{\alpha}_{c,i}(1-\tilde{\alpha}_{c,i})}_{\text{self-gating factor}}, \quad \frac{\partial \mathcal{L}}{\partial \delta_c} = \sum_{k=1}^{C}\sum_{i=1}^{N} \frac{\partial \mathcal{L}}{\partial \tilde{\alpha}_{k,i}}\, \tilde{\alpha}_{k,i}\big(\mathbf{1}[k=c] - \tilde{\alpha}_{c,i}\big), \tag{5}$$

driving δ_c positive for poorly-segmented classes and suppressing it for well-segmented ones. The on-diagonal self-gating factor $\tilde{\alpha}_{c,i}(1-\tilde{\alpha}_{c,i})$ peaks when attention weights are uncertain and vanishes at saturation, yielding an implicit curriculum effect.

Warm-start initialisation. Zero-initialising $\boldsymbol{\delta}$ starves rare classes of gradient ($\tilde{\alpha}_{c,i} \approx 0 \Rightarrow$ self-gating ≈ 0, regardless of loss). We therefore initialise $\boldsymbol{\delta}_c$ from a *centred log-frequency prior*:

$$\delta_c^{(0)} = \beta\left(\log \pi_c - \tfrac{1}{C}\sum_{c'=1}^{C} \log \pi_{c'}\right), \quad \log \pi_c = \gamma\, \log(1-f_c), \tag{6}$$

giving minority classes a positive initial bias and majority classes a negative one at zero mean shift (in practice $\gamma = 2$, $\beta = 0.8$). Thereafter δ_c is learned end-to-end with no explicit frequency or hardness constraints.

Training objective. Following SAM-Path [22], the objective sums per-class soft Dice loss $\mathcal{L}_{\mathrm{dice}}$, focal loss [15] $\mathcal{L}_{\mathrm{focal}}$, and MSE-IoU regression $\mathcal{L}_{\mathrm{mse}}$ against the realised $\mathrm{IoU}(\hat{y}_c, y_c) = |\hat{y}_c \cap y_c| / |\hat{y}_c \cup y_c|$, plus an ℓ_2 regulariser on the learnable biases:

$$\mathcal{L} = \sum_{c=1}^{C}\Big[(1-\alpha)\mathcal{L}_{\mathrm{dice}}(\hat{y}_c, y_c) + \alpha\, \mathcal{L}_{\mathrm{focal}}(\hat{y}_c, y_c) + \mu\, \mathcal{L}_{\mathrm{mse}}\big(\widehat{\mathrm{iou}}_c, \mathrm{IoU}(\hat{y}_c, y_c)\big)\Big] + \lambda \|\boldsymbol{\delta}\|_2^2, \tag{7}$$

where $\alpha \in [0,1]$ balances Dice and focal terms, $\mu > 0$ weights IoU regression, and $\lambda = 10^{-4}$ prevents unbounded bias growth. We use a $100\times$ higher learning rate η_b for δ_c for rapid adaptation. DFA adds negligible cost: one scalar per class and one broadcast logit shift per cross-attention layer.

3 Experiments and Results

3.1 Experimental Setup

Datasets and splits. We evaluate on three public histopathology benchmarks, each using the official train/test split with 20% of training held out for

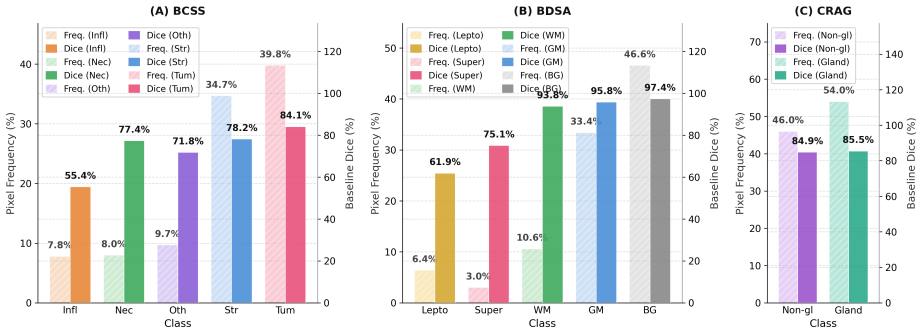

Fig. 2: **Frequency–difficulty disconnect.** Pixel frequency (hatched bars) vs. Dice of the focal-loss baseline (solid bars) per class across three datasets.

validation (SAM-Path protocol [22]); a single fixed split (no cross-validation) is held identical across all methods. (**i**) **BCSS** [1]: breast cancer, five classes (Tumor, Stroma, Inflammatory, Necrosis, Other), 40×, 1024^2. (**ii**) **BDSA** [9]: brain tissue, five classes (Background, Gray/White Matter, Leptomeninges, Superficial Cortex), 10×, 1024^2; 10 WSIs yield ∼82k patches, split at the slide level. (**iii**) **CRAG** [10]: colorectal adenocarcinoma, two classes (Non-gland, Gland), 20×, 1536^2.

Architecture and training. All models use SAM-Path with ViT-B (SAM encoder) and HIPT (pathology encoder), trained with AdamW ($\eta = 10^{-4}$, batch 4, cosine annealing) on NVIDIA A100 GPUs. Epoch counts (30/20/60 for BCSS /BDSA/CRAG) follow each dataset's validation-loss plateau and are identical across all five methods, so any residual schedule effect is shared. **Hyperparameters:** $\gamma = 2$ for all focal variants; $\alpha = 0.25$, $\mu = 0.0625$ on BCSS/BDSA and $\alpha = 0.125$, $\mu = 0$ on CRAG (SAM-Path defaults); $\beta = 0.8$, $\lambda = 10^{-4}$, $\eta_b = 10^{-2}$ for DFA, all fixed across datasets after a small grid search on BCSS, with stable performance and no extensive tuning.

Baselines. Five configurations differing *only* in attention bias: (**i**) *Base w/o FL* (Eq. 7 without focal term); (**ii**) *Base w/ FL* (SAM-Path default); (**iii**) *CFFA* (Eq. 2); (**iv**) *HCFA* (two-stage, Eq. 3); (**v**) *DFA (Ours)* (single-stage, Eq. 4).

3.2 The Frequency–Difficulty Disconnect

We first test whether pixel frequency predicts segmentation difficulty—the assumption behind CFFA and frequency-based reweighting. The focal-loss baseline contradicts it across all three benchmarks (Fig. 2). In BCSS, Inflammatory and Necrosis share near-identical frequency yet differ by 22.0 Dice points; Other (the most frequent minority class) scores below the rarer Necrosis (71.8% vs. 77.4%), inverting the rarity–difficulty ranking. In BDSA, the rarest class (Superficial Cortex, 3.0%) beats the more frequent Leptomeninges (6.4%) by 13.2 points. Only CRAG, near-balanced, shows frequency loosely tracking difficulty.

Thus true difficulty is governed by morphological variability, boundary ambiguity, and inter-class contextual confusion [2], largely independent of pixel frequency. This is the regime in which we evaluate CFFA, HCFA, and DFA.

3.3 Main Results

Table 1 reports per-class Dice (F1 overlap) and IoU (Jaccard); IoU mirrors Dice throughout, so discussion focuses on Dice, with hard classes (baseline Dice below the dataset mean) of primary interest. Since CFFA already captures the frequency-related component of difficulty, the CFFA→DFA gain concentrates on the non-frequency residual (i.e., hard classes) rather than on global mean Dice.

BCSS. DFA reaches the highest mean Dice (0.7826), +4.9 over the focal-loss baseline and +0.5 over HCFA. Inflammatory improves +15.2 over Base w/o FL and +2.7 over CFFA, confirming DFA targets true difficulty rather than frequency; it also gives the best Other (0.7581), whose difficulty frequency-based methods underestimate.

BDSA. Without focal loss the base model collapses on Leptomeninges and Superficial Cortex (Dice =0.000); focal loss recovers them to 0.6191 and 0.7508. DFA further raises both to 0.6908 and 0.7808 and achieves the best White Matter (0.9556).

CRAG. Even with minimal imbalance, DFA attains the highest mean Dice (0.8858), +0.5 over HCFA and +1.5 over CFFA, capturing difficulty structure beyond frequency.

Efficiency. As a single-stage method, DFA cuts wall-clock training time by ∼40% versus two-stage HCFA while matching or exceeding its Dice everywhere.

3.4 Analysis of Learned Attention Biases

At convergence (Fig. 3a–c), $\boldsymbol{\delta}$ provides a model-derived difficulty signal: hard classes acquire positive bias (e.g., Inflammatory +0.410; Leptomeninges +0.434), easy classes negative (e.g., Tumor −0.524; Background −0.720). The scatter plots (Fig. 3d–f), plotting δ_c against *baseline Dice* rather than frequency, show a strong negative correlation between difficulty and δ_c, complementary to, not contradicting, the frequency–difficulty disconnect in Fig. 2. Crucially, biases diverge from their frequency-based init: Inflammatory ($f_c = 7.8\%$) gets +0.410 vs. Necrosis +0.068 at near-identical frequency (8.0%), a 6× gap mirroring their 22-point Dice difference, which no frequency-driven scheme could produce. Thus, δ_c captures difficulty sources, such as boundary complexity and contextual confusion, that are invisible to frequency alone.

Ablation: warm-start initialisation. Table 2 isolates the centred log-frequency prior (Eq. 6) on BCSS. Cold-start converges to a near-flat δ profile ($|\delta_c| \leq 0.02$) as self-gating suppression ($\tilde{\alpha}_{c,i}(1-\tilde{\alpha}_{c,i}) \approx 0$) stalls updates. Warm-start breaks this degeneracy: ∼20× wider spread and +0.75 Dice. Beyond the Dice gain, the warm prior is what lets DFA learn the intended difficulty-aware profile (hard +, easy −); without it δ stays uninformative even when overall Dice is close.

Table 1: **Segmentation performance across three histopathology benchmarks.** Each method occupies two columns: Dice (left) and IoU (right). Means and stds are computed over three independent training runs with different random seeds. Hard classes (baseline Dice below dataset mean) are shaded amber. **Bold**: best result per row.

		Base w/o FL [22]		**Base w/ FL** [22]		**CFFA**[17]		**HCFA**(Ours)		**DFA (Ours)**	
Dataset	**Class**	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU
BCSS	*Mean*	0.7032	0.5535	0.7336	0.5882	0.7737	0.6382	0.7780	0.6434	**0.7826**	**0.6488**
	$\pm std$	±0.003	±0.004	±0.003	±0.003	±0.002	±0.003	±0.002	±0.003	±0.002	±0.002
	Tumor	0.8298	0.7090	0.8410	0.7257	0.8662	0.7640	0.8674	0.7659	**0.8684**	**0.7674**
	Stroma	0.7521	0.6027	0.7816	0.6415	0.7968	0.6622	0.7944	0.6590	**0.7969**	**0.6624**
	Inflamm.	0.4974	0.3310	0.5537	0.3828	0.6228	0.4522	0.6392	0.4697	**0.6493**	**0.4807**
	Necrosis	0.7615	0.6148	0.7739	0.6312	0.8311	0.7110	0.8398	**0.7238**	**0.8402**	0.7230
	Other	0.6754	0.5099	0.7178	0.5598	0.7514	0.6018	0.7490	0.5987	**0.7581**	**0.6104**
BDSA	*Mean*	0.5589	0.5241	0.8480	0.7644	0.8703	0.7931	0.8712	0.7944	**0.8739**	**0.7953**
	$\pm std$	±0.003	±0.003	±0.002	±0.003	±0.002	±0.002	±0.002	±0.002	±0.001	±0.002
	BG	0.9530	0.9121	0.9742	0.9501	0.9761	0.9537	**0.9763**	**0.9539**	**0.9763**	0.9538
	GM	0.9270	0.8644	0.9585	0.9205	0.9648	0.9322	0.9656	**0.9337**	**0.9658**	0.9335
	WM	0.9147	0.8441	0.9375	0.8834	0.9504	0.9060	0.9514	0.9079	**0.9556**	**0.9090**
	Lepto.	0.0000	0.0000	0.6191	0.4660	0.6840	0.5384	0.6851	0.5395	**0.6908**	**0.5405**
	Super.	0.0000	0.0000	0.7508	0.6022	0.7761	0.6351	0.7773	0.6368	**0.7808**	**0.6395**
CRAG	*Mean*	0.8423	0.7276	0.8520	0.7422	0.8707	0.7711	0.8806	0.7868	**0.8858**	**0.7951**
	$\pm std$	±0.003	±0.004	±0.002	±0.003	±0.002	±0.003	±0.002	±0.002	±0.002	±0.002
	Non-gland	0.8392	0.7229	0.8487	0.7372	0.8643	0.7611	0.8712	0.7717	**0.8774**	**0.7816**
	Gland	0.8454	0.7322	0.8553	0.7472	0.8771	0.7812	0.8901	0.8020	**0.8942**	**0.8087**

Table 2: Warm- vs. cold-start ablation on BCSS.

Init	$\delta_{\text{Infl.}}$	$\delta_{\text{Oth.}}$	$\delta_{\text{Nec.}}$	$\delta_{\text{Str.}}$	$\delta_{\text{Tum.}}$	**Dice (±std)**
Cold ($\delta_c^{(0)}=0$)	+0.010	+0.002	+0.001	−0.012	−0.019	$0.7751_{\pm.005}$
Warm (Eq. 6)	+0.410	+0.253	+0.068	−0.308	−0.524	$0.7826_{\pm.002}$

3.5 Qualitative Results

Fig. 4 shows consistent gains: misclassified Inflammatory regions in BCSS are recovered with coherent boundaries; Leptomeninges boundaries in BDSA sharpen progressively from CFFA to HCFA to DFA; and CRAG glands show markedly reduced fragmentation.

4 Limitations and Future Work

DFA applies a *single global* bias per class. Classes such as Tumor or Stroma can be harder in some regions than others, so a class-level scalar captures only *average* difficulty, acting as a prior, while the token-dependent $\mathbf{QK}^\top$ term preserves spatial selectivity. Spatially adaptive or region-conditioned biases capturing instance-level and intra-class difficulty are a natural extension. Broader validation across architectures (CNN–transformer hybrids, other query-based

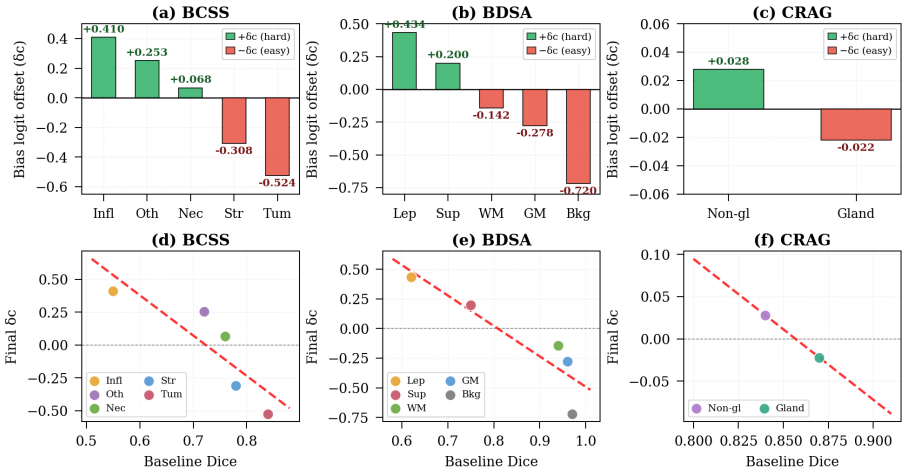

Fig. 3: **Converged biases δ_c and difficulty correlation.** Top (a–c): δ_c per class (**positive** ⇔ hard, **negative** ⇔ easy). Bottom (d–f): δ_c vs. baseline Dice with fitted trend line; the strong negative correlation indicates δ_c tracks empirical difficulty rather than frequency.

decoders) and modalities (MRI, CT), explicit attention-map comparison, and a convergence analysis remain future work.

5 Conclusion

Focal loss alone is insufficient and frequency-based reweighting unreliable: CFFA fails when rare classes are not inherently hard, and HCFA needs a costly pre-training stage. DFA resolves both by learning class-specific difficulty continuously within cross-attention, outperforming all baselines across three benchmarks with only C extra scalar parameters. The warm-start prior prevents gradient starvation, the self-gating factor gives implicit curriculum regulation, and encoding difficulty at the representation level yields model-wide gains over post-hoc gradient corrections.

Acknowledgments. The work was supported by NIH grants P30 AG072946 and U24 NS133945. We are grateful to the research participants, clinicians, and staff for making this work possible. This work used Delta advanced computing and data resource at the University of Illinois Urbana-Champaign through allocation CIS230383 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by U.S. National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

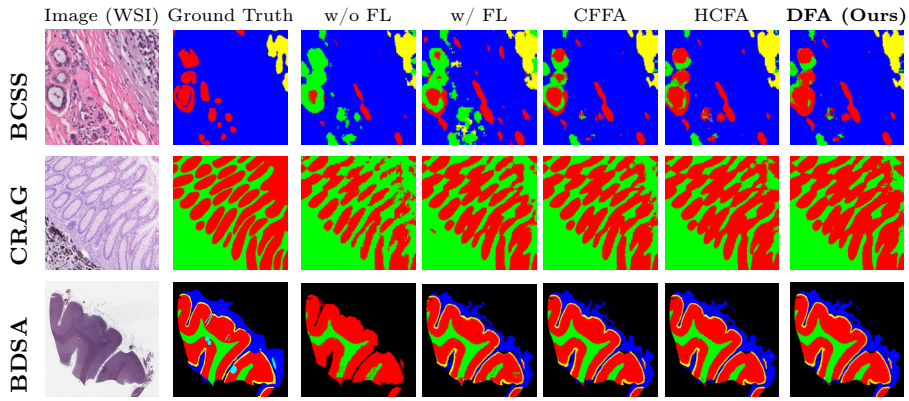

Fig. 4: **Qualitative segmentation comparison across three pathology benchmarks.** Each row shows a WSI patch with ground truth and predictions from all five methods.

BCSS: Inflamm. Tumor Stroma Other Necrosis **CRAG:** Non-Gland Gland **BDSA:** BG Gray M. White M. Leptom. Superficial.

References

1. Amgad, M., Elfandy, H., Hussein, H., Atteya, L.A., Elsebaie, M.A., Abo Elnasr, L.S., Sakr, R.A., Salem, H.S., Ismail, A.F., Saad, A.M., et al.: Structured crowdsourcing enables convolutional segmentation of histology images. Bioinformatics **35**(18), 3461–3467 (2019)
2. Buda, M., Maki, A., Mazurowski, M.A.: A systematic study of the class imbalance problem in convolutional neural networks. Neural networks **106**, 249–259 (2018)
3. Campanella, G., Hanna, M.G., Geneslaw, L., Miraflor, A., Werneck Krauss Silva, V., Busam, K.J., Brogi, E., Reuter, V.E., Klimstra, D.S., Fuchs, T.J.: Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nature medicine **25**(8), 1301–1309 (2019)
4. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European conference on computer vision. pp. 213–229. Springer (2020)
5. Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F., Jaume, G., Song, A.H., Chen, B., Zhang, A., Shao, D., Shaban, M., et al.: Towards a general-purpose foundation model for computational pathology. Nature medicine **30**(3), 850–862 (2024)
6. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1290–1299 (2022)
7. Cheng, B., Schwing, A., Kirillov, A.: Per-pixel classification is not all you need for semantic segmentation. Advances in neural information processing systems **34**, 17864–17875 (2021)
8. Cui, Y., Jia, M., Lin, T.Y., Song, Y., Belongie, S.: Class-balanced loss based on effective number of samples. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9260–9269 (2019)

9. Flanagan, M.E., Gutman, D., Dugger, B.N., Cooper, L., Pearce, T.M., Kovacs, G.G., Kukull, W.W., Crary, J.F., Manthey, D., Biber, S., et al.: Brain digital slide archive: An open source whole slide image sharing platform for ad/adrd research and diagnostics. Alzheimer's & Dementia **21**, e103720 (2025)
10. Graham, S., Chen, H., Gamper, J., Dou, Q., Heng, P.A., Snead, D., Tsang, Y.W., Rajpoot, N.: Mild-net: Minimal information loss dilated network for gland instance segmentation in colon histology images. Medical image analysis **52**, 199–211 (2019)
11. Graham, S., Vu, Q.D., Jahanifar, M., Weigert, M., Schmidt, U., Zhang, W., Zhang, J., Yang, S., Xiang, J., Wang, X., Rumberger, J.L., Baumann, E., Hirsch, P., Liu, L., Hong, C., Aviles-Rivero, A.I., Jain, A., Ahn, H., Hong, Y., Azzuni, H., Xu, M., Yaqub, M., Blache, M.C., Piégu, B., Vernay, B., Scherr, T., Böhland, M., Löffler, K., Li, J., Ying, W., Wang, C., Snead, D., Raza, S.E.A., Minhas, F., Rajpoot, N.M.: Conic challenge: Pushing the frontiers of nuclear detection, segmentation, classification and counting. Medical Image Analysis **92**, 103047 (2024). `https://doi.org/https://doi.org/10.1016/j.media.2023.103047`
12. Graham, S., Vu, Q.D., Raza, S.E.A., Azam, A., Tsang, Y.W., Kwak, J.T., Rajpoot, N.: Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. Medical image analysis **58**, 101563 (2019)
13. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023)
14. Kumari, L.N., Bandara, C.T., Chuah, C.N., Cheung, S.C.S.: A warmer start to active learning with adaptive gaussian mixture models for skin lesion segmentation. In: 2025 IEEE International Conference on Image Processing (ICIP). pp. 2247–2252 (2025)
15. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence **42**(2), 318–327 (2020)
16. Lu, M.Y., Williamson, D.F., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F.: Data-efficient and weakly supervised computational pathology on whole-slide images. Nature biomedical engineering **5**(6), 555–570 (2021)
17. Menon, A.K., Jayasumana, S., Rawat, A.S., Jain, H., Veit, A., Kumar, S.: Long-tail learning via logit adjustment. arXiv preprint arXiv:2007.07314 (2020)
18. Press, O., Smith, N.A., Lewis, M.: Train short, test long: Attention with linear biases enables input length extrapolation. arXiv preprint arXiv:2108.12409 (2021)
19. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of machine learning research **21**(140), 1–67 (2020)
20. Shrivastava, A., Gupta, A., Girshick, R.: Training region-based object detectors with online hard example mining. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 761–769 (2016)
21. Yeung, M., Sala, E., Schönlieb, C.B., Rundo, L.: Unified focal loss: Generalising dice and cross entropy-based losses to handle class imbalanced medical image segmentation. Computerized Medical Imaging and Graphics **95**, 102026 (2022)
22. Zhang, J., Ma, K., Kapse, S., Saltz, J., Vakalopoulou, M., Prasanna, P., Samaras, D.: Sam-path: A segment anything model for semantic segmentation in digital pathology. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 161–170. Springer (2023)
23. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159 (2020)