Physics-Grounded Adversarial Stain Augmentation with Calibrated Coverage Guarantees

Mingi Hong MGHONG@MTSCO.CO.KR
434, Samseong-ro, Gangnam-gu, Seoul, Republic of Korea

Abstract

Stain variation across hospitals degrades histopathology models at deployment. Existing augmentation methods perturb color spaces with arbitrary hyperparameters, lacking both a principled budget and coverage guarantees for unseen centers. We propose **C**alibrated **A**dversarial **S**tain **A**ugmentation (**CASA**), which performs adversarial augmentation in the Macenko stain parameter space with a budget calibrated from multi-center statistics via the DKW inequality. On Camelyon17-WILDS (5 seeds), CASA achieves 93.9%±1.6% slide-level accuracy—outperforming HED-strong (88.4%±7.3%), RandStainNA (85.2%±6.7%), and ERM (63.9%±11.3%)—with the highest worst-group accuracy (84.9%±0.9%) among all 10 compared methods.

Keywords: Histopathology, Domain Generalization, Adversarial Augmentation, Distributionally Robust Optimization

1. Introduction

Hematoxylin and Eosin (H&E) staining varies across hospitals due to differences in reagent concentrations, protocols, and scanners. This variation causes models trained at one center to degrade at another, limiting clinical deployment (Tellez et al., 2019).

Existing approaches fall into two categories. *Stain normalization* transforms images to a reference template but introduces artifacts and depends on template selection (Macenko et al., 2009). *Stain augmentation* increases color diversity during training; Tellez et al. showed that random perturbation in HED space improves generalization (Tellez et al., 2019), and RandStainNA bridges augmentation with normalization in LAB space (Shen et al., 2022). However, random methods require manual selection of perturbation strength (σ), and this choice affects OOD performance: we observe an 8 percentage point gap between HED-light (σ=0.05, 80.4%) and HED-strong (σ=0.2, 88.4%) on Camelyon17-WILDS, with no principled criterion for selecting σ.

Recent adversarial augmentation methods—AdvST (Zheng et al., 2024) and Zhong et al. (Zhong et al., 2022)—learn worst-case perturbations via PGD but operate in generic color spaces (HSV, contrast) with arbitrary budgets unconnected to the physical staining process.

If instead the model trains on the *hardest* stain variation within a realistic range, it should generalize to any center whose stain falls within that range. We propose **CASA**, which realizes this idea: (1) adversarial augmentation in the *physics-grounded* Macenko stain parameter space, where perturbations correspond to realistic stain variations; and (2) an adversarial budget *calibrated* from multi-center data with a statistical coverage guarantee.

2. Method

Stain decomposition. The Beer-Lambert law models H&E image formation: for each pixel j, $\mathbf{I}_j = I_0 \exp(-\mathbf{W}\mathbf{h}_j)$, where $\mathbf{W} \in \mathbb{R}^{3\times 2}$ contains unit-norm stain basis vectors (Hematoxylin and Eosin columns) and $\mathbf{h}_j \in \mathbb{R}^2$ contains the corresponding stain concentrations. We extract $\mathbf{W}$ via the Macenko method (Macenko et al., 2009), building on the color deconvolution framework of Ruifrok and Johnston (Ruifrok and Johnston, 2001), and obtain $\mathbf{h}$ by least-squares inversion.

Training loop. Each iteration solves a min-max problem. The *inner maximization* finds the worst-case stain perturbation $(\boldsymbol{\delta}_W^*, \boldsymbol{\delta}_h^*)$ via K-step PGD:

$$\max_{\|\boldsymbol{\delta}_W\|\le\tau_W,\ \|\boldsymbol{\delta}_h\|\le\tau_H} \mathcal{L}\Big(f_\theta\big(I_0 e^{-(\mathbf{W}_{\text{ref}}+\boldsymbol{\delta}_W)\,(\mathbf{h}_0\odot(1+\boldsymbol{\delta}_h))}\big),\ y\Big) \tag{1}$$

where $\boldsymbol{\delta}_W$ perturbs stain vector directions (projected onto a spherical cap to preserve unit norm) and $\boldsymbol{\delta}_h$ scales concentrations element-wise. Since $\tau_H < 1$, the factor $(1 + \delta_{h,k})$ remains strictly positive for the final perturbation; intermediate PGD iterates are clamped to $\delta_{h,k} \ge -1$ to maintain non-negativity throughout. The reference stain matrix $\mathbf{W}_{\text{ref}}$ and initial concentrations $\mathbf{h}_0$ are precomputed once per batch from the Macenko decomposition. After K steps, the augmented image is reconstructed via the Beer-Lambert inverse and passed to the *outer minimization*, which updates θ by SGD on $\mathcal{L}(f_\theta(\tilde{\mathbf{I}}), y)$. The model trains on the worst-case stain perturbation within the calibrated budget at every iteration.

Calibrated budget. The budgets τ_W and τ_H are not hyperparameters—they are estimated from training data. Given n images from the training centers: (i) extract per-image stain matrices $\mathbf{W}_i$ and concentrations $\mathbf{h}_i$; (ii) compute the angular deviation $\alpha_i = \angle(\mathbf{W}_i, \bar{\mathbf{W}})$ and concentration ratio $r_i = q_{99}(\mathbf{h}_i) \,/\, \overline{q_{99}(\mathbf{h})}$, where q_{99} denotes the 99th percentile; (iii) set τ_W and τ_H as the $(1-\delta+\varepsilon_n)$-empirical quantiles of $\{\alpha_i\}$ and $\{|r_i - 1|\}$ respectively, where ε_n is the DKW correction below. On Camelyon17-WILDS with n=1,000 training images, this yields τ_W=0.353 rad (20.2°) and τ_H=0.987.

Coverage guarantee. The DKW inequality (Dvoretzky et al., 1956; Massart, 1990) bounds the deviation of the empirical CDF F_n from the true CDF F:

$$\Pr\left[\sup_t |F_n(t) - F(t)| > \varepsilon_n\right] \le 2e^{-2n\varepsilon_n^2} \tag{2}$$

Since we apply the bound to both τ_W and τ_H, a union bound requires each to hold at level $\beta/2$, giving $\varepsilon_n = \sqrt{\ln(4/\beta)/(2n)}$. We take the $(1-\delta+\varepsilon_n)$-empirical quantile as each budget, which ensures that the true $(1-\delta)$-quantile is covered with probability $\ge 1-\beta$. With δ=β=0.05 and n=1,000: ε_n=0.047, giving a quantile level of 0.997. Each stain parameter of a new center thus falls within its budget with $\ge 95\%$ confidence.

3. Experiments

Setup. We evaluate on Camelyon17-WILDS (Koh et al., 2021): binary tumor classification across 5 hospitals (train: 0,3,4; val: 1; test: 2; 302K patches). All methods use DenseNet-121 trained from scratch, SGD (lr=0.001, wd=0.01), batch size 32, 10 epochs,

5 seeds. CASA uses K=5 PGD steps. We report the WILDS official slide-level macro accuracy (`acc_avg`). Baselines include stain augmentation (HED-strong/light (Tellez et al., 2019), RandStainNA (Shen et al., 2022), Macenko-norm (Macenko et al., 2009)), adversarial methods (AdvST (Zheng et al., 2024), Zhong et al. (Zhong et al., 2022)), domain adaptation (DANN (Ganin et al., 2016)), and bilevel stain optimization in the same parameter space as CASA.

Table 1: Camelyon17-WILDS test results (DenseNet-121, 5 seeds).

Method	**acc_avg**	**acc_wg**
CASA (ours)	$\mathbf{93.9 \pm 1.6}$	$\mathbf{84.9 \pm 0.9}$
HED-strong (σ=0.2)	88.4 ± 7.3	76.1 ± 13.5
RandStainNA	85.2 ± 6.7	57.6 ± 21.3
HED-light (σ=0.05)	80.4 ± 6.6	66.1 ± 10.3
Macenko-norm	78.0 ± 14.7	46.0 ± 17.5
Bilevel-Stain	71.4 ± 12.7	43.2 ± 14.5
DANN	69.8 ± 4.1	53.1 ± 5.8
Zhong et al.	64.8 ± 9.8	46.8 ± 9.3
ERM	63.9 ± 11.3	36.1 ± 9.4
AdvST	56.4 ± 7.1	31.7 ± 14.9

Results. CASA ranks first in both accuracy (93.9%) and worst-group stability (±0.9%). Three findings emerge from Table 1:

(1) Stain-specific > generic adversarial. AdvST (56.4%) perturbs HSV, contrast, and sharpness—it scores below ERM (63.9%). Zhong et al. (64.8%) operates on feature-level style and only matches ERM. Adversarial augmentation requires stain-specific parameterization to help rather than hurt.

(2) Adversarial > random in the same space. CASA (93.9%) outperforms HED-strong (88.4%) by 5.5 points with 4× lower `acc_avg` variance. Bilevel optimization (71.4%), which minimizes rather than maximizes the inner objective in the same parameter space, scores 22.5 points lower—worst-case perturbations generalize better than "helpful" ones.

(3) Calibrated budget removes hyperparameter sensitivity. HED-strong (σ=0.2) and HED-light (σ=0.05) differ by 8 points, yet no criterion exists for choosing σ. CASA derives its budget from the data; no manual tuning is needed.

4. Conclusion

Two design choices make CASA effective where generic adversarial methods fail: grounding perturbations in the physical stain parameter space, and calibrating the budget from multi-center data instead of tuning it by hand. Because the only domain-specific components are the stain decomposition and the calibration set, the same min-max formulation transfers to any staining protocol whose forward model is known.

References

Aryeh Dvoretzky, Jack Kiefer, and Jacob Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, 27(3):642–669, 1956.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. WILDS: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning (ICML)*, pages 5637–5664, 2021.

Marc Macenko, Marc Niethammer, J Stephen Marron, David Borland, John T Woosley, Xiaojun Guan, Charles Schmitt, and Nancy E Thomas. A method for normalizing histology slides for quantitative analysis. In *IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1107–1110. IEEE, 2009.

Pascal Massart. The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *The Annals of Probability*, 18(3):1269–1283, 1990.

Arnout C Ruifrok and Dennis A Johnston. Quantification of histochemical staining by color deconvolution. *Analytical and Quantitative Cytology and Histology*, 23(4):291–299, 2001.

Yiqing Shen, Yulin Luo, Dinggang Shen, and Jing Ke. RandStainNA: Learning stain-agnostic features from histology slides by bridging stain augmentation and normalization. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 212–221. Springer, 2022.

David Tellez, Geert Litjens, Péter Bándi, Wouter Bulten, John-Melle Bokhorst, Francesco Ciompi, and Jeroen van der Laak. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. *Medical Image Analysis*, 58:101544, 2019.

Guangtao Zheng, Mengdi Huai, and Aidong Zhang. AdvST: Revisiting data augmentations for single domain generalization. In *AAAI Conference on Artificial Intelligence*, 2024.

Zhun Zhong, Yuyang Zhao, Gim Hee Lee, and Nicu Sebe. Adversarial style augmentation for domain generalized urban-scene segmentation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.