Hierarchical Self-Supervised Adversarial Training for Robust Vision Models in Histopathology

Hashmat Shadab Malik[1(✉)[0000−0001−6197−3840]], Shahina Kunhimon[1[0009−0001−6809−2285]], Muzammal Naseer[2[0000−0001−7663−7161]], Fahad Shahbaz Khan[1,3[0000−0002−4263−3143]], and Salman Khan[1,4[0000−0002−9502−1749]]

[1] Mohamed Bin Zayed University of Artificial Intelligence (MBZUAI), UAE
{hashmat.malik,shahina.kunhimon,fahad.khan,salman.khan}@mbzuai.ac.ae
[2] Khalifa University, UAE
muhammadmuzammal.naseer@ku.ac.ae
[3] Linköping University, Sweden
[4] Australian National University, Australia

Abstract. Adversarial attacks pose significant challenges for vision models in critical fields like healthcare, where reliability is essential. Although adversarial training has been well studied in natural images, its application to biomedical and microscopy data remains limited. Existing self-supervised adversarial training methods overlook the hierarchical structure of histopathology images, where patient-slide-patch relationships provide valuable discriminative signals. To address this, we propose *Hierarchical Self-Supervised Adversarial Training* (`HSAT`), which exploits these properties to craft adversarial examples using multi-level contrastive learning and integrate it into adversarial training for enhanced robustness. We evaluate `HSAT` on multiclass histopathology dataset OpenSRH and the results show that `HSAT` outperforms existing methods from both biomedical and natural image domains. `HSAT` enhances robustness, achieving an average gain of 54.31% in the white-box setting and reducing performance drops to 3-4% in the black-box setting, compared to 25-30% for the baseline. These results set a new benchmark for adversarial training in this domain, paving the way for more robust models. Code and models are available at https://github.com/HashmatShadab/HSAT.

Keywords: Adversarial Robustness · Self-Supervised Learning.

1 Introduction

Computer vision has made significant strides in recent years, improving performance across various tasks. However, deep learning models remain vulnerable to adversarial attacks, where subtle, imperceptible perturbations to images cause incorrect model responses [30,14,24,3,29,9]. These attacks are broadly classified into *white-box attacks*, where the attacker has full access to the model, and

✉Corresponding Author

black-box attacks, where access is limited, and adversarial examples are crafted using query-based [4,2,27] or transfer-based methods [10,28,26]. This vulnerability poses significant risks, especially in healthcare, where the reliability of vision models is crucial for accurate medical diagnoses and treatment planning.

Previous works in natural image domains has explored adversarial training strategies, including supervised [24,33,31] and self-supervised learning (SSL) methods [19,11,23], to enhance model robustness. However, there has been limited focus on assessing the robustness of vision models in the medical domain, where reliable performance is crucial. While SSL methods are gaining traction in medical domains, especially in histopathology [17,7,5], self-supervised adversarial training for robust representation learning of such images remains unexplored. In this work, we bridge this gap by investigating self-supervised adversarial training for learning robust visual features of histopathology images.

In self-supervised adversarial training, contrastive learning is preferred over masking-based reconstruction due to its discriminative objective, which aligns with adversarial robustness goals and effectively enhances vision model robustness [19,11,23]. Existing methods primarily focus on instance- or patch-level discriminative learning, overlooking the hierarchical structure present in biomedical data. Gigapixel slides in medical imaging are typically tiled into smaller patches, forming a patient-slide-patch hierarchy where all samples from a single patient correspond to a shared diagnosis [17]. This multi-level correlation provides a strong, underutilized discriminative signal that can be harnessed to improve adversarial training. Inspired by this, we propose *Hierarchical Self-Supervised Adversarial Training* (`HSAT`) — a novel approach that harnesses the hierarchical correlations between patches, slides, and patients to enhance adversarial training. `HSAT` first generates adversarial examples in the *maximization step* by utilizing a hierarchical contrastive objective, resulting in the generation of strong adversarial examples that confuse the model. This is followed by a *minimization step*, to reduce the loss induced by these adversarial perturbations. By jointly optimizing these multi-level objectives, `HSAT` strengthens the model's adversarial robustness, resulting in superior performance compared to current approaches [19,11,23].

We benchmark `HSAT` on the multiclass `OpenSRH` [16] microscopic cancer dataset, comparing it with state-of-the-art non-adversarial self-supervised methods and instance-level adversarial training approaches. `HSAT` significantly enhances adversarial robustness, providing stronger protection against both white-box and black-box attacks. We summarize our key contributions as follows:

- ❶ We propose a hierarchy-wise adversarial perturbation method that leverages the correlation between patient, slide, and patch levels in biomedical microscopy images to craft adversarial examples.
- ❷ Our proposed framework, `HSAT`, mitigates adversarial vulnerabilities by minimizing confusion from hierarchical adversarial perturbations, promoting robust representation learning across patient, slide, and patch levels.
- ❸ `HSAT` achieves state-of-the-art adversarial performance on the `OpenSRH` dataset, outperforming current self-supervised adversarial training approaches adapted from the natural image domain.

2 Related Work

Self-Supervised Learning in Histopathology: Advancements in microscopy have enabled the digitization of tissue imaging, resulting in large-scale digital datasets, much of which remain unannotated due to the expertise required for labeling. Therefore, SSL has become a key method for pretraining models in histopathology, with techniques such as contrastive learning [8], reconstruction from partial signals [1], and knowledge distillation [6]. Notably, contrastive learning, which leverages instance-level identity by contrasting different views of the same image, has proven effective in learning rich representations. Recent efforts in histopathology have utilized contrastive learning to design hierarchical contrastive objectives for patch representations [17]. However, these models are vulnerable to adversarial attacks, and to our knowledge, no work has explored adversarial self-supervised training to improve robustness.

Adversarial Robustness and Self-supervised Learning: Adversarial training strategies have been extensively explored in the self-supervised learning domain for natural image data to improve model robustness [19,11,23]. Several instance-level adversarial contrastive learning methods [19,11,23] leverage contrastive loss to generate adversarial examples, strengthening model resilience. However, these methods predominantly focus on instance- or patch-level adversarial training, often neglecting the hierarchical relations inherent in more complex data, such as histopathology images. As a result, directly adapting these approaches to histopathology data may lead to suboptimal robustness. Existing adversarial methods in histopathology typically address attacks and defenses at the patch or instance level [20,18,13,12,25], overlooking the multi-scale relationships within the data. To bridge this gap, we propose a novel self-supervised adversarial training framework that explicitly captures the multi-level correlation in histopathology images, offering a more effective and robust training paradigm.

3 Methodology

In this section, we introduce our approach, `HSAT` — a hierarchical self-supervised adversarial training framework designed for training vision models on histopathology images. The core idea behind `HSAT` is that adversarial examples can be generated hierarchically (see Fig. 1), not only by *pushing* an image away from its augmented versions (*Patch Positives*) but also from posiitve paired images at the slide and patient level (*Slide and Patient Positives*). This process forms the maximization step. To counter these perturbations, the model is trained to *pull* the adversarial counterparts of positive pairs closer in the feature space. Ultimately, `HSAT` strengthens model robustness across all hierarchical levels. To the best of our knowledge, we are the first to investigate self-supervised adversarial training in histopathology images and integrate hierarchy-based multiple discriminative tasks for generating adversarial examples for training.

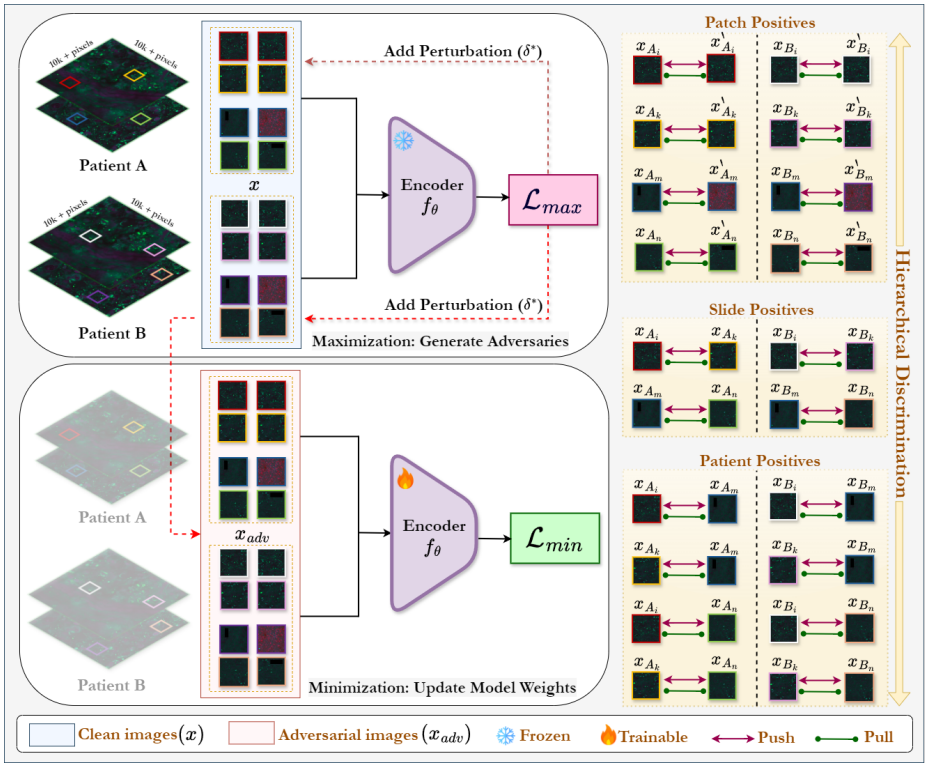

Fig. 1. `HSAT Framework`: Adversarial examples are generated by maximizing the distance between images and their corresponding positive pairs across patch, slide, and patient *(maximization)*. The model is then updated by minimizing the hierarchical loss on adversarial images *(minimization)*, enforcing robust feature learning at all levels.

3.1 Hierarchical Self-Supervised Adversarial Training

In histopathology, vision tasks follow a three-tier data hierarchy: gigapixel Whole Slide Images (WSIs or slides) are divided into sub-sampled patches, with patches linked to slides and slides to patients. Recent work [17] leverages this structure using self-supervised contrastive objectives with positive pairs sampled across all the levels. Given a batch of n patients, each with n_s slides, n_p patches per slide, and n_a augmentations per patch, the encoder f_θ extracts features from the patches, which are then projected into latent representations z. These representations are used to compute the hierarchical contrastive loss, with positive pairs defined at three levels: patch (same patch augmentations), slide (patches from the same slide), and patient (patches from the same patient). Therefore hierarchical contrastive loss on a set of images $\mathcal{I}$ is defined as follows:

$$\mathcal{L}_{con}^{l} = -\sum_{i \in \mathcal{I}} \frac{1}{|\mathcal{H}_l(i)|} \sum_{h \in \mathcal{H}_l(i)} \log \left(\frac{\exp\left((\mathbf{z}_i \cdot \mathbf{z}_h)/\tau\right)}{\sum_{p \in \mathcal{P}(i)} \exp\left((\mathbf{z}_i \cdot \mathbf{z}_p)/\tau\right)} \right) \tag{1}$$

Here, $\mathbf{z}_i$ is the feature representation of patch i, $\mathcal{H}_l(i)$ is the set of positive pairs at hierarchical level l (patch, slide, or patient), while $\mathcal{P}(i)$ denotes all the other patches in $\mathcal{I}$, except patch i. $(\mathbf{z}_i \cdot \mathbf{z}_h)$ measures cosine similarity, between the feature vectors, while τ is the temperature parameter. However, as shown in Section 4, both [17] and instance-based self-supervised adversarial training from the natural domain [19] are vulnerable adversarial attacks resulting in significant performance drop (see Table 1). To achieve optimal robustness, we propose a simple yet effective approach; adversarially training a self-supervised model using adversarial examples crafted through *hierarchy-wise* attacks. We first describe the attack, followed by our training strategy to enhance the model's robustness.

Hierarchy-wise Adversarial Attack: Given an input image x, we craft an adversarial image x_{adv} by maximizing the hierarchical contrastive loss across all levels; patch, slide, and patient. The adversarial perturbation δ^* is obtained by solving the optimization problem $\delta^* = \arg\max_{\|\delta\|_\infty \leq \epsilon} \mathcal{L}_{\max}(x+\delta, \mathcal{I})$ where:

$$\mathcal{L}_{\max}(x+\delta, \mathcal{I}) = \sum_l \mathcal{L}^l_{con}(f_\theta(x+\delta), f_\theta(\mathcal{I})), \quad \forall l \in \{\text{patch, slide, patient}\} \quad (2)$$

The perturbation δ is constrained by the ℓ_∞ norm to remain within a perturbation budget ϵ. The resulting adversarial image $x_{\text{adv}} = x + \delta$ is designed to confuse the model not only with its corresponding augmented patch x' but also with the positive paired patches at the slide and patient levels. We use Projected Gradient Descent (`PGD`) to optimize the perturbation δ iteratively, ensuring it stays within the defined ℓ_∞ bound. This multi-task maximization objective provides significantly more diversity and discriminative signal compared to current instance-level adversarial training methods [19,11,23], enabling the model to learn better robust representations.

Hierarchial Self-Supervised Adversarial Training (HSAT): We propose `HSAT`, a min-max optimization framework to learn robust representations via self-supervised contrastive learning. The inner maximization step, $\mathcal{L}_{\max}$ generates adversarial perturbations by crafting a heirarchy-wise attack. The outer minimization step, $\mathcal{L}_{\min}$, updates the encoder f_θ to minimize the heirarchial contrastive loss, encouraging robust feature representations across patch, slide, and patient levels. The overall objective is given by:

$$\min_\theta \mathcal{L}_{\min}(\theta) = \mathbb{E}_{x\sim\mathcal{D}} \sum_l \mathcal{L}^l_{con}(f_\theta(x_{adv}), f_\theta(\mathcal{I})), \ \forall l \in \{\text{patch, slide, patient}\} \quad (3)$$

By jointly optimizing across patch, slide, and patient levels, `HSAT` captures richer, more resilient representations, reducing the risk of overfitting to specific adversarial patterns as well as improving its generalization.

4 Experiments

Dataset: `OpenSRH` [16], an optical microscopy dataset of brain tumor biopsies and resections, was used for training and validation. It comprises 303 patients

across 7 classes. Data is split into 243 patients for training and 60 for validation. Whole slide images (WSIs) are patched into 300×300 patches, yielding 229K training and 56K validation patches.

Implementation details: We conduct experiments using ResNet-50 [15] and WideResNet-50 [32], each coupled with an MLP layer to project embeddings into 128-dimensional latent space, as outlined in [17]. For both non-adversarial hierarchical training [17], and our `HSAT` framework, models are trained for 40,000 epochs with an effective batch size ($n \cdot n_a \cdot n_s \cdot n_p$) of 512. We employ the AdamW optimizer [22] with an initial learning rate of 0.001 for the first 10% of iterations, followed by a cosine decay scheduler. During training, we utilize weak (horizontal and vertical flips) and strong (affine transformations, color jittering, and random erasing) augmentations. Specifically for `HSAT`, we implement a perturbation warmup strategy for the first 5,000 epochs and dynamic data augmentations [23] to stabilize the training. Adversarial examples in the maximization step are generated by optimizing $\mathcal{L}_{max}$ using (`PGD`) attack for 5 steps at $\epsilon = \frac{8}{255}$. We evaluate three `HSAT` variants: (1) `HSAT-Patient`, capturing the full patient-slide-patch hierarchy ($n_a = n_s = n_p = 2$); (2) `HSAT-Slide`, incorporating slide-level information ($n_a = n_s = 2, n_p = 1$); and (3) `HSAT-Patch`, using only patch-level information, aligning with current instance-level adversarial training methods ($n_a = 2, n_s = n_p = 1$). By keeping the effective batch size constant, we ensure that the baseline adversarial training approach, `HSAT-Patch`, operates under an equivalent computational training budget. Our results focus mainly on `HSAT-Patient`, denoted as `HSAT`, using the full term only for comparisons.

Evaluation: We use k-Nearest Neighbors (kNN) for quantitative evaluation, freezing the pretrained visual backbone to compute embeddings for the train and validation splits on `OpenSRH`. A kNN classifier predicts class labels by matching test patches to the nearest training embeddings. We report accuracy (`Acc`) and mean class accuracy (`MCA`) at patch, slide, and patient levels, following [17]. For adversarial evaluation, we generate adversarial examples across various perturbation budgets using iterative attacks (10 steps) that minimize cosine similarity between clean and adversarial feature representations.

Robustness against White-Box Attacks: In this section, we examine the robustness of vision models in white-box settings, using iterative attacks, such as `PGD` [24], `BIM` [21] and `MIFGSM` [10]. Adversarial examples are crafted on the `OpenSRH` validation set with a perturbation budget of $\epsilon = \{\frac{4}{255}, \frac{8}{255}\}$ and 10 attack iterations. Models trained with our `HSAT` framework are compared against non-adversarial hierarchical training methods for histopathology images [17] and instance-level adversarial training methods proposed for natural images ($\mathcal{L}_{max} = \mathcal{L}_{min} = \mathcal{L}_{patch}$), referred to as `HSAT-Patch`. As reported in Table 1, our approach consistently outperforms both baselines. Under the `PGD` attack at $\epsilon = \frac{8}{255}$ using a ResNet-50 backbone, we achieve gains in adversarial robustness of 43.90%, 60.70%, and 58.33% compared to [17], and 6.68%, 10.51%, and 10% on patch, slide, and patient classification, respectively compared to instance-level adversarial training (`HSAT-Patch`). Furthermore, we observe a drop in clean performance compared to [17] which is an expected trade-off between adversar-

Training	Model	Clean		PGD-4		BIM-4		MIFGSM-4		PGD-8		BIM-8		MIFGSM-8	
		Acc	MCA	Acc	MCA	Acc	MCA	Acc	MCA	Acc	MCA	Acc	MCA	Acc	MCA
Patch Classification															
Training from Scratch															
[17]	ResNet-50	83.10	82.29	10.25	9.84	10.27	9.85	11.15	10.34	5.52	5.01	5.57	5.02	5.97	5.31
HSAT-Patch	ResNet-50	44.84	39.63	44.67	39.65	44.67	39.65	44.80	39.80	44.23	39.76	44.22	39.75	44.26	39.88
HSAT	ResNet-50	66.43	65.61	64.07	61.36	64.11	61.40	65.11	62.40	57.18	53.48	57.10	53.38	58.52	54.76
Training from ImageNet weights															
[17]	ResNet-50	80.07	77.78	23.61	22.11	23.32	21.84	21.76	21.03	13.07	12.80	13.06	12.75	12.53	12.71
[17]	WideResNet-50	79.24	76.46	32.28	26.53	32.24	26.51	28.95	23.86	20.87	17.21	20.90	17.25	19.24	15.97
HSAT-Patch	ResNet-50	50.80	44.36	51.08	44.95	51.06	44.94	51.10	44.96	50.29	45.22	50.29	45.22	50.41	45.31
HSAT	ResNet-50	66.48	66.33	64.07	61.36	64.11	61.40	65.11	62.40	56.97	55.00	56.98	55.04	57.92	56.06
HSAT	WideResNet-50	68.35	67.76	60.51	59.27	60.54	59.32	61.44	60.19	53.13	51.26	53.12	51.26	54.44	52.73
Slide Classification															
Training from Scratch															
[17]	ResNet-50	89.11	88.24	9.73	8.93	9.69	8.89	12.45	11.27	3.11	2.66	4.28	3.57	4.67	3.87
HSAT-Patch	ResNet-50	58.37	53.30	57.20	52.18	57.09	52.15	57.98	52.75	56.81	52.11	56.81	52.11	57.98	53.22
HSAT	ResNet-50	76.26	75.09	77.04	73.58	77.00	73.53	77.43	73.87	75.10	69.74	75.00	69.57	74.71	69.69
Training from ImageNet weights															
[17]	ResNet-50	85.60	83.86	27.63	26.11	27.24	25.70	25.29	24.86	10.51	10.61	10.12	10.06	10.51	10.86
[17]	WideResNet-50	85.60	83.62	45.91	38.74	45.91	38.74	40.08	34.21	21.40	18.13	20.62	17.45	19.46	16.27
HSAT-Patch	ResNet-50	60.31	54.69	61.87	56.08	61.82	56.03	61.48	55.80	60.74	55.75	60.71	55.73	60.70	55.73
HSAT	ResNet-50	77.43	76.87	75.51	73.65	75.49	73.63	75.88	73.21	71.21	67.91	71.60	68.20	71.20	67.89
HSAT	WideResNet-50	80.93	79.25	77.82	74.42	77.82	74.42	76.65	73.56	71.21	66.50	70.43	65.66	70.82	66.19
Patient Classification															
Training from Scratch															
[17]	ResNet-50	90.00	91.08	10.00	8.73	10.00	8.73	13.33	11.59	3.33	2.86	3.33	3.86	5.00	4.29
HSAT-Patch	ResNet-50	55.00	51.62	55.00	51.68	54.89	51.60	54.75	50.96	56.72	54.83	56.67	54.63	56.61	54.62
HSAT	ResNet-50	76.67	73.03	78.33	79.00	78.33	79.00	76.67	76.15	76.67	77.29	76.67	73.03	75.00	71.73
Training from ImageNet weights															
[17]	ResNet-50	86.67	88.20	31.67	28.95	28.33	26.35	25.00	23.49	11.67	10.32	11.60	10.26	11.59	10.28
[17]	WideResNet-50	85.00	85.34	41.67	35.41	41.67	35.41	36.67	31.67	23.33	20.19	21.67	18.76	18.33	16.03
HSAT-Patch	ResNet-50	58.33	55.77	60.28	58.63	60.00	58.60	60.04	58.53	60.13	58.59	60.00	58.33	60.02	58.21
HSAT	ResNet-50	78.33	78.72	75.00	74.85	75.00	74.85	75.00	74.85	70.00	67.84	71.67	69.42	71.67	69.42
HSAT	WideResNet-50	80.00	80.17	76.67	73.03	78.33	75.89	75.00	73.29	66.67	64.24	66.67	64.24	68.33	65.67

Table 1. Our proposed HSAT results in significant gain in robustness against different *white box* attacks at perturbation budgets of $\epsilon = \frac{4}{255}$ and $\epsilon = \frac{8}{255}$ *(higher is better)*.

ial and clean accuracy. However, our method still surpasses current instance-level adversarial training methods on clean performance, with improvements of 15.68%, 17.12%, and 20% in patch, slide, and patient classification, respectively.

Robustness against Black-Box Attacks: In this section, we evaluate and compare the robustness of different models against transfer-based *black-box* attacks. These attacks exploit the *transferability* property of adversarial examples i.e., adversarial examples crafted on a surrogate model transfer to unknown target models. To measure transferability, all vision models are used interchangeably as both surrogate and target models. In Table 2, performance drop on adversarial examples (crafted using PGD at $\epsilon = \frac{8}{255}$) relative to clean samples is reported using Acc-D and MCA-D. We observe that target vision models trained using our HSAT framework are significantly more robust against transferability of adversarial examples crafted across different surrogate models. On average, HSAT trained target models show a performance drop of around $3-4\%$, while target models trained using [17] show a performance drop of around $25-30\%$.

Target →	Type	ResNet-50 ([17])		ResNet-50 (HSAT)		WResNet-50 ([17])		WResNet-50 (HSAT)		ResNet-101 ([17])	
Surrogate ↓		Acc-D	MCA-D	Acc-D	MCA-D	Acc-D	MCA-D	Acc-D	MCA-D	Acc-D	MCA-D
ResNet-50 ([17])	Patch	66.81	64.81	1.42	3.08	35.64	37.99	2.48	4.32	39.49	41.98
	Slide	74.71	72.94	1.55	3.56	32.68	37.34	2.72	5.26	37.35	39.60
	Patient	75.00	77.88	0	1.27	31.67	38.63	3.33	5.58	36.66	43.33
ResNet-50 (HSAT)	Patch	20.48	25.67	9.54	11.38	20.62	25.76	4.81	7.43	23.69	29.68
	Slide	15.95	22.24	6.22	8.96	18.28	24.93	5.83	9.41	12.45	18.87
	Patient	16.67	25.46	8.33	10.88	21.67	28.50	6.67	9.74	15.00	23.17
WResNet-50 ([17])	Patch	32.21	36.60	1.56	3.32	58.51	59.37	2.39	4.35	35.72	40.37
	Slide	18.67	26.83	1.94	4.20	64.20	65.73	3.50	6.84	24.51	30.75
	Patient	15.00	24.42	0	1.27	61.67	65.15	1.67	4.28	26.66	35.05
WResNet-50 (HSAT)	Patch	20.09	24.73	3.63	5.27	20.20	24.99	15.34	16.62	23.92	29.40
	Slide	12.45	18.65	3.11	5.81	15.56	22.53	10.50	13.35	12.45	18.22
	Patient	16.67	25.17	5.00	8.29	16.67	24.19	13.33	15.93	13.33	20.45
ResNet-101 ([17])	Patch	54.06	53.54	2.59	4.41	52.46	51.95	3.53	5.50	70.48	68.36
	Slide	45.91	47.43	4.28	6.76	45.91	48.13	4.67	7.19	77.82	74.36
	Patient	43.34	48.63	5.00	7.15	41.67	46.09	3.33	5.71	78.33	78.19
Average	Patch	31.71	35.13	**2.33**	**4.02**	32.23	35.18	**3.30**	**5.40**	30.68	35.36
	Slide	23.24	28.79	**2.72**	**5.08**	28.11	33.23	**4.18**	**7.17**	21.69	26.86
	Patient	22.92	30.92	**2.50**	**4.49**	27.92	34.35	**3.75**	**6.33**	22.91	30.50

Table 2. Performance against transfer-based *black-box* attacks *(lower is better)*: HSAT-trained models show strong robustness, while baseline models remain highly vulnerable.

Ablations: The ablation studies for *Hierarchical Adversarial Training* (HSAT) are presented in Tables 3 and 4. Table 3 examines the effect of increasing hierarchical discrimination levels during adversarial training, progressing from patch-level (HSAT-Patch) to slide-level (HSAT-Slide) and patient-level (HSAT-Patient). The results show a consistent improvement in adversarial robustness as the level of discrimination increases, highlighting the benefits of leveraging hierarchical structures in biomedical data. Table 4 investigates the impact of different loss terms used to generate adversarial examples in the maximization step of HSAT. Starting with patch-level loss $\mathcal{L}_{con}^{patch}$, we observe a performance boost when slide-level loss $\mathcal{L}_{con}^{slide}$ is added, with further gains upon incorporating patient-level loss $\mathcal{L}_{con}^{patient}$. These findings underscore the effectiveness of multi-level adversarial training, demonstrating that aligning robust representations across hierarchical levels enhances model resilience against adversarial attacks.

Training	Patch		Slide		Patient	
	Acc	MCA	Acc	MCA	Acc	MCA
HSAT-Patch	50.29	45.22	60.70	55.73	60.00	58.92
HSAT-Slide	55.07	50.20	69.32	64.69	68.33	66.29
HSAT-Patient	56.97	55.00	71.21	67.91	70.00	67.84

Table 3. Comparing adversarial training at different levels of discrimination.

$\mathcal{L}_{max}$	Patch		Slide		Patient	
	Acc	MCA	Acc	MCA	Acc	MCA
$\mathcal{L}_{con}^{patch}$	50.19	47.52	63.59	63.27	62.00	63.55
$\mathcal{L}_{con}^{patch} + \mathcal{L}_{con}^{slide}$	49.19	46.70	65.76	66.09	63.33	66.67
$\mathcal{L}_{con}^{p} + \mathcal{L}_{con}^{s} + \mathcal{L}_{con}^{pt}$	56.97	55.00	71.21	67.91	70.00	67.84

Table 4. Comparing effect of using different types of loss in the maximization step.

5 Conclusion

We introduce *Hierarchical Self-Supervised Adversarial Training* (`HSAT`), a novel approach designed for improving adversarial robustness in biomedical image analysis. `HSAT` integrates multi-level discriminative learning into adversarial training, capturing the hierarchical structure within biomedical data. Unlike traditional self-supervised adversarial training methods, `HSAT` forces models to learn resilient features across both fine-grained and coarse-grained levels. Our experiments on `OpenSRH` benchmark demonstrate that `HSAT` significantly enhances resistance to adversarial attacks, outperforming existing methods.

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

References

1. Boyd, J., Liashuha, M., Deutsch, E., Paragios, N., Christodoulidis, S., Vakalopoulou, M.: Self-supervised representation learning using visual field expansion on digital pathology. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 639–647 (2021)
2. Brendel, W., Rauber, J., Bethge, M.: Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. In: International Conference on Learning Representations (ICLR) (2018)
3. Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In: 2017 ieee symposium on security and privacy (sp). pp. 39–57. Ieee (2017)
4. Chen, P.Y., Zhang, H., Sharma, Y., Yi, J., Hsieh, C.J.: Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In: Proceedings of the 10th ACM workshop on artificial intelligence and security (2017)
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. Chen, R.J., Krishnan, R.G.: Self-supervised vision transformers learn visual concepts in histopathology. arXiv preprint arXiv:2203.00585 (2022)
7. Ciga, O., Xu, T., Martel, A.L.: Self supervised contrastive learning for digital histopathology. Machine Learning with Applications **7**, 100198 (2022). https://doi.org/https://doi.org/10.1016/j.mlwa.2021.100198, https://www.sciencedirect.com/science/article/pii/S2666827021000992
8. Ciga, O., Xu, T., Martel, A.L.: Self supervised contrastive learning for digital histopathology. Machine learning with applications **7**, 100198 (2022)
9. Croce, F., Hein, M.: Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In: International conference on machine learning. pp. 2206–2216. PMLR (2020)
10. Dong, Y., Liao, F., Pang, T., Su, H., Zhu, J., Hu, X., Li, J.: Boosting adversarial attacks with momentum. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
11. Fan, L., Liu, S., Chen, P.Y., Zhang, G., Gan, C.: When does contrastive learning preserve adversarial robustness from pretraining to finetuning? Advances in neural information processing systems **34**, 21480–21492 (2021)

12. Foote, A., Asif, A., Azam, A., Marshall-Cox, T., Rajpoot, N., Minhas, F.: Now you see it, now you dont: adversarial vulnerabilities in computational pathology. arXiv preprint arXiv:2106.08153 (2021)
13. Ghaffari Laleh, N., Truhn, D., Veldhuizen, G.P., Han, T., van Treeck, M., Buelow, R.D., Langer, R., Dislich, B., Boor, P., Schulz, V., et al.: Adversarial attacks and adversarial robustness in computational pathology. Nature communications **13**(1), 5711 (2022)
14. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. corr abs/1512.03385 (2015) (2015)
16. Jiang, C., Chowdury, A., Hou, X., Kondepudi, A., Freudiger, C., Conway, K., Camelo-Piragua, S., Orringer, D., Lee, H., Hollon, T.: Opensrh: optimizing brain tumor surgery using intraoperative stimulated raman histology. Advances in neural information processing systems **35**, 28502–28516 (2022)
17. Jiang, C., Hou, X., Kondepudi, A., Chowdury, A., Freudiger, C.W., Orringer, D.A., Lee, H., Hollon, T.C.: Hierarchical discriminative learning improves visual representations of biomedical microscopy. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19798–19808 (2023)
18. Kanca, E., Gulsoy, T., Ayas, S., Kablan, E.B.: Generating robust adversarial images in medical image classification using gans. In: 2024 Innovations in Intelligent Systems and Applications Conference (ASYU). pp. 1–5 (2024). https://doi.org/10.1109/ASYU62119.2024.10757024
19. Kim, M., Tack, J., Hwang, S.J.: Adversarial self-supervised contrastive learning. Advances in neural information processing systems **33**, 2983–2994 (2020)
20. Kumar, D., Sharma, A., Narayan, A.: Attacking cnns in histopathology with snap: Sporadic and naturalistic adversarial patches (student abstract). Proceedings of the AAAI Conference on Artificial Intelligence **38**(2121), 23550–23551 (Mar 2024). https://doi.org/10.1609/aaai.v38i21.30468, https://ojs.aaai.org/index.php/AAAI/article/view/30468
21. Kurakin, A., Goodfellow, I.J., Bengio, S.: Adversarial examples in the physical world. In: Artificial intelligence safety and security, pp. 99–112. Chapman and Hall/CRC (2018)
22. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
23. Luo, R., Wang, Y., Wang, Y.: Rethinking the effect of data augmentation in adversarial contrastive learning. arXiv preprint arXiv:2303.01289 (2023)
24. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083 (2017)
25. Maliamanis, T.V., Apostolidis, K.D., Papakostas, G.A.: How resilient are deep learning models in medical image analysis? the case of the moment-based adversarial attack (mb-ada). Biomedicines **10**(10), 2545 (2022)
26. Malik, H.S., Kunhimon, S.K., Naseer, M., Khan, S., Khan, F.S.: Adversarial pixel restoration as a pretext task for transferable perturbations. arXiv preprint arXiv:2207.08803 (2022)
27. Narodytska, N., Kasiviswanathan, S.: Simple black-box adversarial attacks on deep neural networks. In: Computer Vision and Pattern Recognition (CVPR) Workshop (2017)
28. Naseer, M.M., Khan, S.H., Khan, M.H., Shahbaz Khan, F., Porikli, F.: Cross-domain transferability of adversarial perturbations. Advances in Neural Information Processing Systems (NeurIPS) (2019)

29. Su, J., Vargas, D.V., Sakurai, K.: One pixel attack for fooling deep neural networks. IEEE Transactions on Evolutionary Computation **23**(5), 828–841 (2019)
30. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199 (2013)
31. Wong, E., Rice, L., Kolter, J.Z.: Fast is better than free: Revisiting adversarial training. arXiv preprint arXiv:2001.03994 (2020)
32. Zagoruyko, S., Komodakis, N.: Wide residual networks. arXiv preprint arXiv:1605.07146 (2016)
33. Zhang, H., Yu, Y., Jiao, J., Xing, E., El Ghaoui, L., Jordan, M.: Theoretically principled trade-off between robustness and accuracy. In: International conference on machine learning. pp. 7472–7482. PMLR (2019)