DGHFA: Dynamic Gradient and Hierarchical Feature Alignment for Robust Distillation of Medical VLMs

Boyi Xiao[1,3], Jianghao Wu[2], Lanfeng Zhong[2], Xiaoguang Zou[4], Yuanquan Wu[4], Guotai Wang[2,3], and Shaoting Zhang[2,3]

[1] Department of Automation, University of Science and Technology of China, Hefei, China
[2] School of Mechanical and Electrical Engineering, University of Electronic Science and Technology of China, Chengdu, China
guotai.wang@uestc.edu.cn
[3] Shanghai Artificial Intelligence Laboratory, Shanghai,China
[4] Clinical Medical Research Center, The First People's Hospital of Kashi (Kashgar) Prefecture, Kashi (Kashgar), China

Abstract. Recent advancements in Medical Vision-Language Models (VLMs) have significantly improved medical cross-modal task performance through large-scale contrastive pre-training. However, deploying these large models in clinical settings is hindered by their computational complexity and vulnerability to adversarial attacks. While knowledge distillation offers a solution by transferring knowledge to efficient student models, traditional methods usually ignore the robustness problem, leaving models susceptible to adversarial attacks. To address these challenges, we propose a novel Dynamic Gradient and Hierarchical Feature Alignment framework (DGHFA) for robust knowledge distillation. Our approach introduces a dynamic gradient calibration mechanism for balanced knowledge transfer and a hierarchical adversarial feature alignment framework to enhance robustness under adversarial attacks. Extensive experiments on two medical VLMs and downstream pathology and X-Ray datasets demonstrate that our method outperforms state-of-the-art approaches across multiple attack scenarios, achieving improvements of 2.3 and 1.7 percentage points in robust accuracy, respectively.

Keywords: Fine-tuning · Adversarial robustness · Distillation.

1 Introduction

In recent years, large pre-trained Vision-Language Models (VLMs) have made significant strides in cross-modal semantic understanding by aligning visual and textual data into a shared embedding space [9, 21]. Building on these advancements, Medical VLMs have also achieved notable progress in fields such as pathology and X-ray imaging [24, 33]. Leveraging large-scale contrastive pre-training, these models capture intricate patterns in both medical visual and textual data, enabling effective adaptation to downstream medical tasks. To enable

practical deployment of such models in clinical settings, Knowledge Distillation (KD) provides a viable solution by transferring knowledge from large teacher models to efficient models, substantially reducing computational burdens [8,20].

Nevertheless, deploying such models in safety-critical applications remains challenging due to their vulnerability to adversarial attacks, where malicious perturbations can disrupt the alignment between images and text, posing significant risks [10,22]. Traditional KD methods focus on preserving teacher model accuracy by relying solely on natural samples, neglecting adversarial robustness. This oversight renders student models susceptible to adversarial samples [6], undermining their reliability. When distilling knowledge from foundation models, this issue is further exacerbated by their complex, high-dimensional feature spaces [26]. Although Adversarial Training (AT) has proven effective in enhancing robustness [17,31], the limited capacity of student models often necessitates a trade-off between accuracy and robustness, leading to suboptimal performance in both aspects [25].

Recently, Adversarial Distillation (AD) has emerged as a promising approach to enhance the robustness of distilled models by transferring robust knowledge from well-trained teachers [4,18]. Methods such as Adversarial Robust Distillation (ARD) [6] and RSLAD [34] integrate adversarial training or leverage robust soft labels to guide student learning. However, these conventional approaches rely on pre-trained robust teachers, which may compromise the generalization capacity of large foundation models [13], and focus solely on output-level alignment, thereby neglecting the teacher's intricate decision-making signals. This limitation hinders the student models' ability to capture high-level semantic nuances and undermines their overall robustness [11,28]. Consequently, distilling knowledge from adversarial-sensitive VLMs into lightweight models with robustness remains an open challenge for safe clinical deployment.

In this paper, we propose a novel Dynamic Gradient and Hierarchical Feature Alignment (DGHFA) framework for robust distillation of medical VLMs, which is inspired by the strong correlation between human perceptual gradients and model robustness shown in recent works [5,23]. Specifically, Perceptually Aligned Gradients (PAG) [1] proposed to enhance robustness by aligning model gradients with human perceptual priors. However, obtaining such perceptual priors usually requires additional model training or costly annotations. Differently from them, our method avoids manual annotations of human perception, and instead leverages the loss gradients of pre-trained models, which naturally emphasize clinically salient features such as edges, textures, and other diagnostic cues and align with human perception, to guide the distillation process. As a result, the student model would focus on the most informative cues, thereby enhancing its predictive accuracy and adversarial robustness. Our contributions are threefold: (i) A dynamic gradient calibration mechanism that leverages perceptually guided gradients to eliminate the need for pre-trained robust teachers while preserving generalization capacity; (ii) A dual weighting strategy, combining sample-adaptive weighting with class-aware gradient harmonization, to ensure balanced and consistent knowledge transfer across heterogeneous architectures;

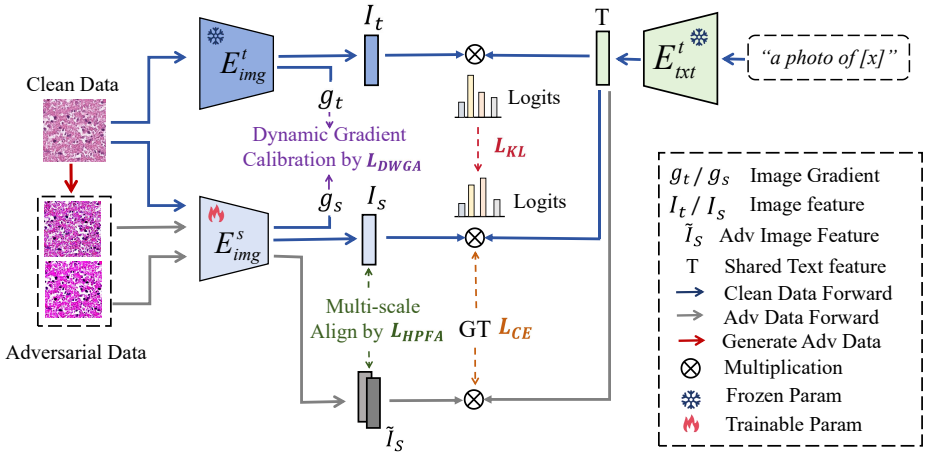

Fig. 1: Overview of our robust knowledge distillation framework DGHFA that combines dynamic gradient and hierarchical feature alignment to enhance robustness of distilled VLMs on downstream datasets.

and (iii) A hierarchical adversarial feature alignment strategy that exploits intermediate adversarial samples to optimize feature distributions between clean and perturbed inputs, thereby enhancing decision boundary robustness. Extensive experiments on pathology and X-ray datasets demonstrate that our approach, leveraging domain-specific pre-trained VLMs, not only maintains competitive performance on clean samples, but also significantly improves robust accuracy over state-of-the-art adversarial distillation and defense methods across multiple attack scenarios, with gains of 2.3 and 1.7 percentage points, respectively.

2 Method

2.1 Problem Settings and Method Overview

Consider a pre-trained medical VLM teacher $\mathcal{T}$, which comprises a large image encoder E_{img} and a text encoder E_{txt} for visual-text alignment. Although $\mathcal{T}$ achieves high accuracy on classification tasks, it is susceptible to adversarial attacks [10, 22]. In this work, we distill only the image encoder E_{img} into a lightweight student model $\mathcal{S}$, while keeping the teacher's text encoder E_{txt} fixed for both teacher and student. Our objective is to preserve $\mathcal{T}$'s strong classification performance in $\mathcal{S}$, while enhancing adversarial robustness.

Fig. 1 illustrates an overview of our proposed method. First, we exploit the teacher model's gradient signals as perceptual guidance to strengthen $\mathcal{S}$'s robustness. Next, a dual adaptive weighting mechanism balances sample- and class-level influences, ensuring more stable knowledge transfer. Finally, we introduce multi-level adversarial samples to progressively align feature distributions, leading to a more robust distillation process for medical VLMs.

2.2 Gradient Calibration for Perceptual Guidance

Unlike previous works [1, 5, 23] that require costly annotations or extra training to obtain perceptually aligned gradients, we propose a gradient calibration method that exploits the intrinsic gradient information provided by the teacher and does not require human intervention.

Specifically, for a clean input image x and a text prompt p, both the teacher model $\mathcal{T}$ and the student model $\mathcal{S}$ generate joint representations using their respective image encoders $E^{\mathcal{T}}_{\text{img}}$ and $E^{\mathcal{S}}_{\text{img}}$, along with their fixed shared text encoder E_{txt}. We introduce a gradient alignment loss, $\mathcal{L}_{\text{PAG}}$, which minimizes the discrepancy between the gradients of the student model's image encoder, $\nabla_{\theta_{\text{img}}}\mathcal{S}(x,p)$, and those of the teacher model's image encoder $\nabla_{\theta_{\text{img}}}\mathcal{T}(x,p)$:

$$\mathcal{L}_{\text{PAG}} = \mathbb{E}_{x,p}\left[\left\|\nabla_{\theta_{\text{img}}}\mathcal{S}(x,p) - \nabla_{\theta_{\text{img}}}\mathcal{T}(x,p)\right\|_2^2\right] \tag{1}$$

where $\nabla_{\theta_{\text{img}}}\mathcal{T}(x,p) = \frac{\partial \mathcal{L}_{\text{CE}}(\mathcal{T}(x,p),y)}{\partial \theta_{\text{img}}}$ denotes the gradient of the teacher model's image encoder with respect to its parameters, and $\nabla_{\theta_{\text{img}}}\mathcal{S}(x,p)$ is defined similarly for the student model. Here, $\mathcal{T}(x,p)$ and $\mathcal{S}(x,p)$ denote the predicted probabilities by the teacher and student models, respectively. y denotes the ground truth label, and $\mathcal{L}_{\text{CE}}$ is the cross-entropy loss.

2.3 Dynamic Weighted Gradient Training for Robustness

While the aforementioned gradient alignment loss effectively enhances model robustness, it treats all samples and categories equally, neglecting the distinct impact of sample and category differences on model robustness. Specifically, difficult samples and vulnerable classes play a critical role in determining the overall robustness of the model [19]. We address this limitation by introducing sample- and category-level weighting to modulate the $\mathcal{L}_{\text{PAG}}$.

To prioritize hard-to-learn samples, we define a dynamic weight $\alpha(x)$ for each input x based on the teacher model's prediction confidence, where higher weights are assigned to samples with low prediction confidence:

$$\alpha(x) = 1 - \max_{c\in\{1,\ldots,C\}} \mathcal{T}(x,p)_c \tag{2}$$

where $\mathcal{T}(x,p)_c$ denotes the softmax confidence of the teacher for class c given input x and text prompt p. At the categoriy level, inspired by [32], we address inter-class vulnerability by dynamically adjusting the category-specific weight τ_c during training. It adaptively assigns larger weights to more vulnerable categories, as their effective learning largely contributes to the model's overall robustness. The weight τ_c is updated based on the category's adversarial risk, which is evaluated using adversarial sample $\tilde{x}$ and computed as:

$$\tau_c^t = \tau_c^{t-1} - \gamma \cdot \frac{R(S(\tilde{x}_c,p)) - \bar{R}(S(\tilde{x},p))}{\max_{k\in\{1,\ldots,C\}} \left|R(S(\tilde{x}_k,p)) - \bar{R}(S(\tilde{x},p))\right|} \tag{3}$$

where $R(S(\tilde{x}_c, p))$ denote the average Cross-Entropy loss for all samples in category c, $\bar{R}(S(\tilde{x}, p)) = \frac{1}{C}\sum_{j=1}^{C} R(S(\tilde{x}_j, p))$ is the average error risk across all categories, and $\gamma \in (0, 1)$ is a scaling factor.

The Dual-Level Weighted Gradient Alignment Loss is computed as:

$$\mathcal{L}_{\text{DWGA}} = \frac{1}{C}\sum_{c=1}^{C} \mathbb{E}_{x\in\mathcal{X}c}\left[\Big(\alpha(x) + \phi \cdot \tau_c\Big)\mathcal{L}_{\text{PAG}}(x)\right] \tag{4}$$

where C is the number of categories, $\mathcal{X}c$ is the set of samples in category c, ϕ balances the influence of the category-level weight.

2.4 Hierarchical Perturbation Feature Alignment

Traditional adversarial training [17, 31] enhances model robustness by aligning features between clean and adversarial samples. However, existing methods rely on the strongest adversarial samples [6, 34], neglecting diverse information from intermediate perturbation levels. We propose Hierarchical Perturbation Feature Alignment (HPFA) by using adversarial examples at multiple perturbation levels, which capture diverse decision boundary insights and mitigate over-fitting.

Let $\mathcal{A} = \{\tilde{x}_s\}_{s=1}^{S}$ denote a set of N_m adversarial samples generated using a gradient-based attack, where s represents the perturbation strength, S is the maximum perturbation strength, and N_m is the number of intermediate adversarial samples. We align the features of these adversarial samples with those of clean inputs to promote robust representation learning. To balance the contributions of adversarial samples with varying strengths during training, we introduce a training time-varying weighting mechanism. The HPFA loss is defined as:

$$\mathcal{L}_{\text{HPFA}} = \sum_{s=1}^{S} w_{s,t} \left\| E_{img}^{\mathcal{S}}(x) - E_{img}^{\mathcal{S}}(\tilde{x}_s) \right\|_2^2, \quad (\tilde{x}_s \in \mathcal{A}) \tag{5}$$

where $E_{img}^{\mathcal{S}}$ denotes the student model's image encoder, and $w_{s,t}$ is the weight for adversarial sample $\tilde{x}_s$ at epoch t which progressively transitions from weaker to stronger perturbations. It initially assigns larger weights to mild adversarial examples for stable feature learning and gradually increases weights for stronger perturbations to enable precise boundary refinement, and is computed as:

$$w_{s,t} = \frac{\exp\left(\beta \cdot s \cdot \frac{t}{T}\right)}{\sum_{s'=1}^{S} \exp\left(\beta \cdot s' \cdot \frac{t}{T}\right)} \tag{6}$$

with T being the total number of training epochs and β a hyperparameter controlling the rate of weight transition.

2.5 Optimization Process

The optimization process for our robust knowledge distillation method consists of two key steps: inner maximization and outer minimization. In the inner maximization step, adversarial examples are generated by perturbing the input image

x to maximize the dissimilarity between its representation and the corresponding ground truth text representation. It is formulated as:

$$\tilde{x} = \arg\max_{\delta} \mathcal{L}_{\text{CE}}(\mathcal{S}(x+\delta, p), y) \quad \text{s.t. } \|\delta\|_\infty \leq \epsilon \tag{7}$$

where $\mathcal{S}(x+\delta, p)$ denotes the logits output of the student model for the adversarial example $\tilde{x} = x + \delta$. δ is a deliberate perturbation applied to the clean input, $\mathcal{L}_{CE}$ is the cross-entropy loss, and ϵ constrains the perturbation magnitude.

In the outer minimization step, the student model is optimized by transferring knowledge from the teacher model. In addition to our robust distillation losses $\mathcal{L}_{\text{DWGA}}$ and $\mathcal{L}_{\text{HPFA}}$, we also follow standard knowledge distillation methods to encourage consistent outputs between $\mathcal{S}$ and $\mathcal{T}$ by KL divergence and cross-entropy losses:

$$\mathcal{L}_{\text{KL}} = \mathbb{E}_{x,p}\left[\mathcal{T}(x,p)\log\frac{\mathcal{T}(x,p)}{\mathcal{S}(x,p)}\right] \tag{8}$$

$$\mathcal{L}_{all} = \mathcal{L}_{\text{CE}}(\mathcal{S}(x,p),y) + \mathcal{L}_{\text{CE}}(\mathcal{S}(\tilde{x},p),y) + \mathcal{L}_{\text{KL}} + \lambda_1\mathcal{L}_{\text{DWGA}} + \lambda_2\mathcal{L}_{\text{HPFA}} \tag{9}$$

where λ_1 and λ_2 are coefficients to balance the corresponding loss terms.

3 Experiments

3.1 Experimental Setup

Models and Datasets We evaluated the effectiveness of our robust distillation method on two medical image datasets: the pathology dataset **CRC100K** (nine classes, 100,000 training and 7,000 validation images) and the X-ray dataset **RSNA** (two classes, 26,684 training and 3,000 validation images). Following [14, 29], we split the dataset into train/valid/test sets with a ratio of 70%, 15%, and 15% for the classification task. For the teacher models, we adopt the pre-trained foundation model **Conch** [14] (ViT-B-16) for pathology and **CXR-CLIP** [29] (ViT-Tiny) for X-ray. Note that both VLMs did not see the downstream datasets in the pre-training stage. For the student model, we use the ResNet-18 architecture as the visual encoder.

Implementation details Our method was implemented in PyTorch on an NVIDIA 4090 GPU with 24GB of memory. The image size and intensity were normalized 224×224 and the range of [0,1], respectively. For distillation, we only trained the visual encoder of the student for 50 epochs using an SGD optimizer with an initial learning rate of 0.1, momentum of 0.9 and weight decay of 5×10^{-4}. The batch size was 128 and the hyper-parameters λ_1 and λ_2 were set to 1, γ and β were set to 0.1 and 2. For the inner maximization, we use a 10-step PGD attack

Table 1: Accuracy and Robustness comparison between different methods. The first section shows methods without distillation from VLMs, while the second section show methods using VLMs for distillation.

Method	Conch [14] → CRC100K					CXR-CLIP [29] → RSNA				
	ACC	FGSM	PGD	CW	AA	ACC	FGSM	PGD	CW	AA
Baseline	92.7	62.6	56.2	49.7	44.5	78.2	35.6	33.7	32.4	29.5
PGD-AT [27]	75.8	72.6	67.8	64.2	62.7	70.2	56.8	52.3	52.4	46.7
RobustWRN [12]	77.6	75.9	71.7	72.4	66.5	72.6	57.3	51.6	52.9	50.2
CTRW [15]	86.9	80.0	76.7	78.5	72.4	74.9	60.3	57.5	56.9	54.5
TRADES [30]	87.5	85.9	85.7	84.4	79.5	78.5	74.3	67.5	66.3	60.5
RSLAD [34]	90.5	87.1	87.3	86.6	83.7	81.3	78.5	68.3	68.6	63.6
AdaAD [11]	93.2	91.8	88.2	87.2	84.5	83.2	79.3	73.1	72.5	66.9
Ours	95.8	**93.7**	**91.5**	**89.5**	**86.3**	85.4	**80.7**	**75.2**	**74.4**	**68.4**
Teacher(zero-shot)	79.8	58.7	55.2	53.5	50.6	80.2	51.1	46.3	45.4	40.3
Teacher(finetuning)	**97.3**	67.5	64.7	64.4	63.6	**89.5**	62.7	59.2	60.3	54.5

with a random start size of 0.001, a step size of 2/255, and an L_∞ perturbation bound of $\epsilon = 8/255$.

We evaluated the model's performance by measuring its accuracy on the natural samples (referred to as clean accuracy) as well as its resilience to adversarial attacks on the adversarial examples (referred to as robust accuracy). The robust accuracy was measured by four widely used metrics, including: FGSM [7], PGD [16], CW_∞ [2], and AutoAttack (AA) [3]. To calculate these metrics, we set the maximum perturbation size to 8/255 and employed 20 steps for PGD and CW, each with a step size of 2/255.

3.2 Comparison with State-of-the-art Methods

Firstly, our method was compared with three teacher-free robust training methods: PGD-AT [27], RobustWRN [12], and CTRW [15], without distillation from pre-trained VLMs. They were also compared with the baseline method of standard training with cross entropy loss on the downstream dataset, without considering robustness. Then, we introduce VLMs as the teacher and compare our method with three state-of-the-art adversarial distillation methods, including TRADES [30], RSLAD [34], and AdaAD [11]. The results on the two datasets are summarized in Table 1. The baseline obtained clean accuracies of 92.7% on CRC100K and 78.2% on RSNA, respectively, but its robustness was only 44.5% and 29.5% under AutoAttack (AA), respectively. Among the distillation-free robust training methods, CTRW [15] achieved the best robustness, i.e., 79.5% on CRC100K and 54.5% on RSNA under AA, respectively. However, it leads to a decrease of clean accuracy compared with the baseline.

Benefiting from the superior capabilities of large pre-trained VLMs, distillation-based methods generally outperformed the distillation-free methods in Table 1.

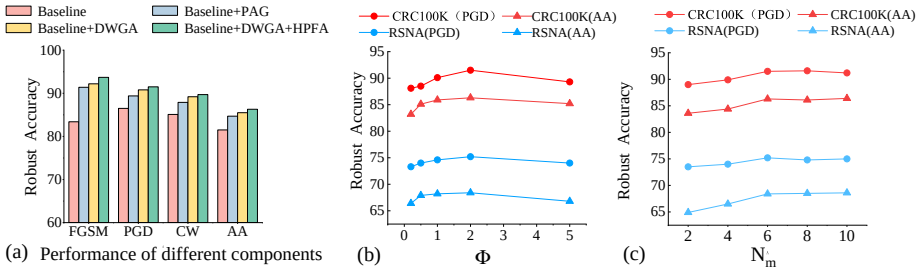

Fig. 2: Ablation study of our method.

Among the existing methods, AdaAD [11] achieved the best performance with a clean accuracy of 93.2% and a robustness of 84.5% under AA on the CRC100K dataset, and the corresponding values were 83.2% and 66.9% on the RSNA dataset, respectively. In contrast, our method outperformed all the existing methods in terms of all the metrics. For example, compared with AdaAD [11], it improves clean accuracy and robustness (AA) by 2.6 and 1.8 percentage points respectively on CRC100K. Additionally, Table 1 shows that our student model significantly outperformed using the pre-trained teacher for zero-shot inference in terms of both clean accuracy and robustness. Compared with fine-tuning the teacher (with all parameters updated) on the downstream dataset, our method has a close clean accuracy with much higher scores under all the robustness metrics. These results demonstrate the effectiveness of our method in ensuring model robustness, particularly in diverse adversarial attack scenarios.

3.3 Ablation Study

To analyze the impact of each component of our method on robustness, we adopted standard adversarial distillation as the baseline and incrementally incorporated three key modules: PAG, DWGA and HPFA. Ablation study results on the CRC100K dataset are presented in Fig. 2(a), which shows that the PAG significantly enhances the robustness of the model. Furthermore, by replacing PAG with DWGA and introducing HPFA, the robustness of the student model is further improved. Our method has two key hyper-parameters, the ratio ϕ and the number of intermediate samples N_m in Section 2.4. The results in Fig. 2(b) show that $\phi = 2$ obtains the best result, while a too small and too large value leads to inferior results. In addition, Fig. 2(c) shows that generally a larger N_m obtains a better result, and a plateau is obtained when $N_m \geq 6$.

4 Conclusion

We propose a robust knowledge distillation framework tailored for medical foundation models. We introduce a dynamic gradient calibration mechanism and

hierarchical adversarial feature alignment for balanced robust knowledge transfer. Thorough evaluation across two widely accessible medical VLMs and downstream datasets confirms the effectiveness of our method. Furthermore, this approach does not rely on robust teacher models and exhibits strong classification performance inherited from the foundation model. Our work has the potential to ensure the safe adoption of Med-VLMs before their deployment.

Acknowledgments. This work was supported by the National Key Research and Development Program of China (2022ZD0160705).

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

References

1. Aggarwal, G., Sinha, A., Kumari, N., Singh, M.: On the benefits of models with perceptually-aligned gradients. arXiv preprint arXiv:2005.01499 (2020)
2. 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)
3. 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)
4. Fang, G., Song, J., Shen, C., Wang, X., Chen, D., Song, M.: Data-free adversarial distillation. arXiv preprint arXiv:1912.11006 (2019)
5. Ganz, R., Kawar, B., Elad, M.: Do perceptually aligned gradients imply robustness? In: International Conference on Machine Learning. pp. 10628–10648. PMLR (2023)
6. Goldblum, M., Fowl, L., Feizi, S., Goldstein, T.: Adversarially robust distillation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 3996–4003 (2020)
7. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)
8. Gou, J., Yu, B., Maybank, S.J., Tao, D.: Knowledge distillation: A survey. International Journal of Computer Vision **129**(6), 1789–1819 (2021)
9. Hafner, M., Katsantoni, M., Köster, T., Marks, J., Mukherjee, J., Staiger, D., Ule, J., Zavolan, M.: Clip and complementary methods. Nature Reviews Methods Primers **1**(1), 1–23 (2021)
10. He, B., Jia, X., Liang, S., Lou, T., Liu, Y., Cao, X.: Sa-attack: Improving adversarial transferability of vision-language pre-training models via self-augmentation. arXiv preprint arXiv:2312.04913 (2023)
11. Huang, B., Chen, M., Wang, Y., Lu, J., Cheng, M., Wang, W.: Boosting accuracy and robustness of student models via adaptive adversarial distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24668–24677 (2023)
12. Huang, H., Wang, Y., Erfani, S., Gu, Q., Bailey, J., Ma, X.: Exploring architectural ingredients of adversarially robust deep neural networks. Advances in Neural Information Processing Systems **34**, 5545–5559 (2021)

13. Hussein, N., Shamshad, F., Naseer, M., Nandakumar, K.: Promptsmooth: Certifying robustness of medical vision-language models via prompt learning. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 698–708. Springer (2024)
14. Lu, M.Y., Chen, B., Williamson, D.F., Chen, R.J., Liang, I., Ding, T., Jaume, G., Odintsov, I., Le, L.P., Gerber, G., et al.: A visual-language foundation model for computational pathology. Nature Medicine **30**(3), 863–874 (2024)
15. Ma, Y., Dong, M., Xu, C.: Adversarial robustness through random weight sampling. Advances in Neural Information Processing Systems **36** (2024)
16. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083 (2017)
17. Maini, P., Wong, E., Kolter, Z.: Adversarial robustness against the union of multiple perturbation models. In: International Conference on Machine Learning. pp. 6640–6650. PMLR (2020)
18. Papernot, N., McDaniel, P., Wu, X., Jha, S., Swami, A.: Distillation as a defense to adversarial perturbations against deep neural networks. In: 2016 IEEE symposium on security and privacy (SP). pp. 582–597. IEEE (2016)
19. Park, H., Min, D.: Dynamic guidance adversarial distillation with enhanced teacher knowledge. In: European Conference on Computer Vision. pp. 204–219. Springer (2024)
20. Park, W., Kim, D., Lu, Y., Cho, M.: Relational knowledge distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3967–3976 (2019)
21. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763. PMLR (2021)
22. Shi, C., Rezai, R., Yang, J., Dou, Q., Li, X.: A survey on trustworthiness in foundation models for medical image analysis. arXiv preprint arXiv:2407.15851 (2024)
23. Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., Madry, A.: Robustness may be at odds with accuracy. arXiv preprint arXiv:1805.12152 (2018)
24. Wang, Z., Wu, Z., Agarwal, D., Sun, J.: Medclip: Contrastive learning from unpaired medical images and text. arXiv preprint arXiv:2210.10163 (2022)
25. Wu, B., Chen, J., Cai, D., He, X., Gu, Q.: Do wider neural networks really help adversarial robustness? Advances in Neural Information Processing Systems **34**, 7054–7067 (2021)
26. Wu, K., Peng, H., Zhou, Z., Xiao, B., Liu, M., Yuan, L., Xuan, H., Valenzuela, M., Chen, X.S., Wang, X., et al.: Tinyclip: Clip distillation via affinity mimicking and weight inheritance. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21970–21980 (2023)
27. Xie, C., Wang, J., Zhang, Z., Ren, Z., Yuille, A.: Mitigating adversarial effects through randomization. arXiv preprint arXiv:1711.01991 (2017)
28. Yin, S., Xiao, Z., Song, M., Long, J.: Adversarial distillation based on slack matching and attribution region alignment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24605–24614 (2024)
29. You, K., Gu, J., Ham, J., Park, B., Kim, J., Hong, E.K., Baek, W., Roh, B.: Cxr-clip: Toward large scale chest x-ray language-image pre-training. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 101–111. Springer (2023)

30. 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)
31. Zhang, J., Xu, X., Han, B., Niu, G., Cui, L., Sugiyama, M., Kankanhalli, M.: Attacks which do not kill training make adversarial learning stronger. In: International conference on machine learning. pp. 11278–11287. PMLR (2020)
32. Zhao, S., Wang, X., Wei, X.: Improving adversarial robust fairness via anti-bias soft label distillation. arXiv preprint arXiv:2312.05508 (2023)
33. Zhong, L., Liao, X., Zhang, S., Zhang, X., Wang, G.: Vlm-cpl: consensus pseudo labels from vision-language models for human annotation-free pathological image classification. arXiv preprint arXiv:2403.15836 (2024)
34. Zi, B., Zhao, S., Ma, X., Jiang, Y.G.: Revisiting adversarial robustness distillation: Robust soft labels make student better. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16443–16452 (2021)