Generalization Analysis for Adversarial Vision Transformers

Ziwen Jiang[1] , **Chang Cao**[1] , **Han Li**[1] , **Hong Chen**[1] , **Rushi Lan**[2,*]
[1]Engineering Research Center of Intelligent Technology for Agriculture, Ministry of Education, and College of Informatics, Huazhong Agricultural University, Wuhan 430070, China
[2]Guangxi Key Laboratory of Image and Graphic Intelligent Processing, Guilin University of Electronic Technology, Guilin 541004, China
jziwen4@gmail.com

Abstract

Vision Transformers (ViTs) exhibit notable susceptibility to adversarial attacks, presenting a significant challenge for their deployment in security-sensitive applications. Despite their considerable empirical successes, a rigorous theoretical foundation for ViT's adversarial generalization behavior has not been adequately established. To address this limitation, we leverage empirical Rademacher complexity to analyze the mechanism of perturbation accumulation through deep ViTs layers. We establish a high-probability generalization bound for ViTs in classification tasks under adversarial settings. Our theoretical framework elucidates the roles of several factors in mitigating perturbation effects, norm regularization of weight matrices (in both MLP and attention modules) and depth-wise propagation constraints on layer-wise norms. Extensive experiments on benchmark datasets corroborate our theoretical insights, bridging the gap between ViTs architecture design and adversarial robustness.

1 Introduction

The ViTs has demonstrated superior performance compared to state-of-the-art convolutional neural networks (CNNs) on standard image recognition tasks [Dosovitskiy, 2020]. Unlike CNNs that rely on local convolution operations, ViTs process images as sequences of patches and utilize self-attention mechanisms to effectively capture long-range dependencies and global context. This novel architecture has not only excelled in image classification but has also become a foundational model for a wide range of other vision tasks, including object detection [Li *et al.*, 2022; Fang *et al.*, 2023] and semantic segmentation [Thisanke *et al.*, 2023].

However, recent studies reveal that ViTs remain vulnerable to adversarial attacks, where imperceptible perturbations to input images can cause misclassifications [Bhojanapalli *et al.*, 2021; Shao *et al.*, 2022; Deng *et al.*, 2025]. This vulnerability has attracted increasing interest in enhancing model robustness through adversarial training. While existing works predominantly focuses on empirical validation [Gu *et al.*, 2022; Wei *et al.*, 2022; Wu *et al.*, 2022], the underlying theoretical foundations remain under-explored.

Recent years have witnessed significant theoretical progress in understanding adversarial generalization problems from statistical learning perspective. This includes unified convergence analyses grounded in hypothesis space capacity measures, such as VC dimension [Attias *et al.*, 2022], Rademacher complexity [Yin *et al.*, 2019; Awasthi *et al.*, 2020], and covering numbers [Mustafa *et al.*, 2022; Wen *et al.*, 2023]. Moreover, related studies have also been conducted based on algorithmic stability [Wu and Cheng, 2021; Xiao *et al.*, 2022]. While these frameworks provide a solid foundation, their direct application to ViTs is challenging due to the architecture's unique properties. Compared to CNN, ViTs lack explicit inductive biases such as spatial connectivity and translation equivariance [Edelman *et al.*, 2022; Sultana *et al.*, 2022]. This implies that the complexity of the ViTs hypothesis space scales precipitously with the input dimension or the sequence length. Consequently, deriving generalization bounds that are independent or only weakly dependent of the dimension d is crucial for establishing the adversarial generalization of ViTs in the finite-data regime. Additionally, the non-Lipschitz nature of dot-product attention on unbounded inputs induces gradient instability [Kim *et al.*, 2021], while the propagation of adversarial perturbations via global self-attention interactions makes the attention weight distribution unpredictable [Shao *et al.*, 2022; Mao *et al.*, 2022]. Unlike classical approaches that operate on single inputs, in Transformer networks which process sequences of tokens, the joint effect of different perturbations on the image patches tokens makes the adversarial loss highly non-convex and intractable, rendering the classical estimation methods ineffective [Yin *et al.*, 2019; Attias *et al.*, 2022].

To address the aforementioned challenges, we characterize the generalization properties of ViTs by analyzing a surrogate objective derived from the supremum of the adversarial loss, which yields a tighter generalization guarantee. The main contributions of this paper are summarized as follows:

- To circumvent the non-convex optimization challenges inherent in the original adversarial loss, we derive an explicit surrogate upper bound $\widehat{\ell}(\cdot)$, which decomposes the adversarial risk into a standard classification error term

*Corresponding author.

Reference	Adversarial Setting	Technique	Bound
[Edelman *et al.*, 2022]	No	Covering Number	$\mathcal{O}(C^{\mathcal{O}(L)}\sqrt{\log(dnm)/n})$
[Trauger and Tewari, 2024]	No	Rademacher Complexity	$\mathcal{O}(C^{\mathcal{O}(L)}\sqrt{\log(d)/n})$
[Li *et al.*, 2023c]	No	Algorithms Stability	$\mathcal{O}(C\log n/\sqrt{nT})$
[Bai *et al.*, 2023]	No	Covering Number	$\mathcal{O}(\sqrt{D/n}+d/m)$
Ours	**Yes**	Rademacher Complexity	$\mathcal{O}(C^{\mathcal{O}(L)}\epsilon\sqrt{\log(d)/n})$

Table 1: Comparison of theoretical analysis for Transformer architectures. (m: the sequence length, n: the number of samples, d: the input dimension, D: the model scale, ϵ: the perturbation budget, T: the number of task, C: a constant bigger than 1.)

and a worst-case perturbation margin. This formulation transforms the intractable adversarial optimization problem into a tractable regularization objective by imposing layer-wise constraints on perturbation propagation throughout the ViT architecture.

- We establish a high-probability generalization bound for multi-layer ViTs in adversarial classification by utilizing adversarial Rademacher complexity. Crucially, our derived bound scales with $\mathcal{O}(\epsilon\sqrt{\log(d)/n})$, demonstrating that the generalization error depends logarithmically on the token dimension d. The derived bound quantifies the interplay between adversarial perturbations and generalization error, providing theoretical insights into how sample size n, sequence length m and weight matrix regularization collectively mitigate adversarial degradation.
- Extensive experimental results on benchmark datasets demonstrate the effectiveness of our theoretical findings in reducing generalization error and achieving good generalization performance.

2 Related Work

2.1 Adversarial Defense on ViTs

Recent research has shown that carefully crafted perturbations can induce ViTs models to make incorrect predictions with high confidence [Bhojanapalli *et al.*, 2021; Shao *et al.*, 2022]. To counter such attacks, a broadly range of defense mechanisms have been developed. Several studies have primarily focused on improving attention mechanisms, utilizing methods such as re-weighting and multi-scale fusion to enhance robustness [Mahmood *et al.*, 2021; Mao *et al.*, 2022]. Other efforts aim to modify the core Transformer components or design specialized token mechanisms to reduce adversarial vulnerability [Guo *et al.*, 2023; Gong *et al.*, 2024]. Furthermore, training strategies such as adversarial pretraining [Xu *et al.*, 2023; Yao *et al.*, 2024] and robust fine-tuning [Wang *et al.*, 2024; Gopal *et al.*, 2025] have proven effective in improving model robustness.

Beyond these targeted architectural and strategic enhancements, adversarial training remains a more fundamental and extensively validated paradigm [Shao *et al.*, 2022; Herrmann *et al.*, 2022; Mo *et al.*, 2022; Jain and Dutta, 2024], recognized as one of the most effective approaches for securing ViTs. Despite its empirical success across various domains [Manzari *et al.*, 2023; Gu *et al.*, 2024; Xiao *et al.*, 2025], the underlying generalization properties of ViTs under adversarial conditions remain poorly understood.

2.2 Theoretical Analysis of Transformers

Recent theoretical advances have provided comprehensive insights into the properties of Transform architectures. Yun et al. [2019] established that Transformers are universal approximators of continuous sequence-to-sequence functions, revealing that self-attention layers play a critical role by computing contextual mappings of input sequences. Deora et al. [2023] then conducted a theoretical analysis of the multi-head attention mechanism, demonstrating that under over-parameterized conditions, it can achieve efficient convergence and robust generalization when trained via gradient descent. Edelman et al.[2022] derived bounds for norm-constrained Transformers using covering number techniques, demonstrating a logarithmic dependency on context length. To address the role of sequence length more broadly, Trauger and Tewari [2024] established sequence-length-independent guarantees by introducing innovative linear covering number measures. More recently, the learning paradigm has been extended to In-Context Learning (ICL), Li et al. [2023c] related the excess risk of ICL to the model's algorithmic stability, providing a unified explanation for its consistent performance across multi-task distributions. In parallel, Bai et al. [2023] established a rigorous statistical theory for the pretraining phase, deriving end-to-end guarantees for sample complexity and proving the architecture's capacity for adaptive in-context algorithm selection. Table 1 summarizes the key comparisons with related theoretical works.

Despite these extensive theoretical foundations for standard Transformers, the specific interplay between ViT architectural nuances and their generalization behavior under adversarial perturbations remains a burgeoning area of inquiry.

3 Problem Formulation

Notations. Let $[L] = \{1, 2, \ldots, L\}$ denote the set of indices up to L. Throughout this paper, vectors are represented by bold lowercase letters (e.g., $\mathbf{w}$), with their individual components denoted by italicized subscripts (e.g., $w_1, \ldots, w_d$ for $\mathbf{w} \in \mathbb{R}^d$). Matrices are designated by bold uppercase letters (e.g., $\mathbf{W}$). For a matrix $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_{d_2}] \in \mathbb{R}^{d_1 \times d_2}$, its (p, q)-norm is defined as: $\left\| \left(\|\mathbf{w}_1\|_p, \ldots, \|\mathbf{w}_{d_2}\|_p \right) \right\|_q$, where

$\|\cdot\|_p$ denotes the standard ℓ_p-norm. Additionally, we denotes the i column of $\mathbf{W}$ as $\mathbf{W}_{*i}$.

Preliminaries. Given n training samples $\{(\mathbf{X}^i, y^i)\}_{i=1}^n$ generated from an unknown distribution $\mathcal{D}$ and a fair initial model, the goal is to find an improved model that maps $\mathbf{X}$ to y for any $(\mathbf{X}, y) \sim \mathcal{D}$. Here, each data point contains m tokens $\mathbf{x}_1^i, \mathbf{x}_2^i, \ldots, \mathbf{x}_m^i$, i.e., $\mathbf{X}^i = \{\mathbf{x}_1^i, \mathbf{x}_2^i, \ldots, \mathbf{x}_m^i\} \in \mathbb{R}^{m\times d}$ and each token is d-dimensional. Following conventional ViTs architectures [Dosovitskiy, 2020], these tokens typically correspond to non-overlapping image patches.

We consider a K-category classification problem within a hypothesis class $\mathcal{F}$. A model $f \in \mathcal{F}$ maps an input $\mathbf{X}$ to a vector of prediction scores $f(\mathbf{X}) \in \mathbb{R}^K$. The predicted class label is then given by:

$$\arg\max_{y'\in[K]}[f(\mathbf{X}^i)]_{y'}.$$

To guide the learning process, the quality of prediction is typically measured by a loss function, we adopt the ramp loss $\ell_\gamma(\cdot,\cdot) : \mathbb{R} \to [0, 1]$ in this work, defined by

$$\ell_\gamma(v, y) = \begin{cases} 1 & \text{if } M(v, y) \le 0 \\ 1 - M(v, y)/\gamma & \text{if } 0 < M(v, y) < \gamma \\ 0 & \text{if } M(v, y) \ge \gamma, \end{cases}$$

where $M(v, y) := v_y - \max_{j\neq y} v_j$ denotes the margin operator. It is worth noting that $\ell_\gamma(v, y)$ is $\|\cdot\|_\infty$-lipschitz with constant $1/\gamma$ and is an upper bound on the zero-one loss. Consequently, the optimal model is obtained by minimizing the empirical training error:

$$\frac{1}{n}\sum_{i=1}^n \ell_\gamma(f(\mathbf{X}^i), y^i).$$

However, in the presence of adversaries, imperceptible perturbations on image patches can deceive the model to make wrong predictions [Shao *et al.*, 2022]. Following the previous work [Chen *et al.*, 2022; Gu *et al.*, 2022; Wei *et al.*, 2022], we postulate that the adversarial image patches originate from a perturbed neighborhood, formally defined as:

$$\mathcal{B}_\infty^\epsilon(\mathbf{X}) = \Big\{\widetilde{\mathbf{X}} = \big[\widetilde{\mathbf{x}}_1, \widetilde{\mathbf{x}}_2, \ldots, \widetilde{\mathbf{x}}_m\big] : \|\widetilde{\mathbf{x}}_i - \mathbf{x}_i\|_\infty \le \epsilon, \forall i \in [m]\Big\},$$

where ϵ denotes the maximum perturbation budget. Given an input image $\mathbf{X}$, its ground-truth label y and the perturbation budget $\epsilon > 0$, the adversary seeks the optimal adversarial patch sequence $\widetilde{\mathbf{X}}_* = [\widetilde{\mathbf{x}}_{*1}, \ldots, \widetilde{\mathbf{x}}_{*m}]$ via:

$$\widetilde{\mathbf{X}}_* = \arg\max_{\widetilde{\mathbf{X}}\in\mathcal{B}_\infty^\epsilon(\mathbf{X})} \ell_\gamma(f(\widetilde{\mathbf{X}}), y).$$

Therefore, the adversarial loss of f is defined as:

$$\widetilde{\ell}(f(\mathbf{X}), y) := \max_{\widetilde{\mathbf{X}}\in\mathcal{B}_\infty^\epsilon(\mathbf{X})} \ell_\gamma(f(\widetilde{\mathbf{X}}), y). \tag{1}$$

Among defense strategies against adversarial perturbations, adversarial training (AT) is widely adopted [Tramèr *et al.*, 2018; Bai *et al.*, 2021], which minimizes the adversarial training error:

$$\widetilde{R}_n(f) := \frac{1}{n}\sum_{i=1}^n \widetilde{\ell}(f(\mathbf{X}^i), y^i),$$

which measures the worst-case performance of the predictor under adversarial perturbations. We are interested in the generalization behavior measured by the adversarial expected error:

$$\widetilde{R}(f) := \mathbb{E}_{(\mathbf{X},y)\sim\mathcal{D}}\Big[\widetilde{\ell}(f(\mathbf{X}), y)\Big].$$

To bridge the gap between empirical performance and theoretical analysis, we define the generalization gap as:

$$\mathcal{E}(f) = \Big|\widetilde{R}(f) - \widetilde{R}_n(f)\Big|.$$

A common step in establishing generalization bounds is to quantify the capacity of the hypothesis class $\mathcal{F}$. The Empirical Rademacher Complexity (ERC) serves as a widely used measure for this purpose.

Definition 3.1 (Empirical Rademacher Complexity)**.** *For any function class $\mathcal{H} \subseteq \mathbb{R}^{\mathcal{Z}}$ and a fixed sample $S := \{z_1, \ldots, z_n\}$, the empirical rademacher complexity is defined as:*

$$\mathfrak{R}_S(\mathcal{H}) := \mathbb{E}_\sigma\left[\sup_{h\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^n \sigma_i h(z_i)\right],$$

where $\{\sigma_i\}_{i=1}^n$ are independent random variables such that $\mathbb{P}(\sigma_i = +1) = \mathbb{P}(\sigma_i = -1) = 0.5$.

We introduce the following classic result in [Yin *et al.*, 2019] to adversarial settings, which shows that the generalization gap can be controlled by adversarial ERC.

Lemma 3.2 ([Yin *et al.*, 2019])**.** *Suppose that the range of the loss function is $[0, B]$. For any $\delta \in (0, 1)$, with probability at least $1 - \delta$, the following holds for all $f \in \mathcal{F}$:*

$$\mathcal{E}(f) \le 2B\mathfrak{R}_S(\widetilde{\ell}\circ\mathcal{F}) + 3B\sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{2n}}.$$

To evaluate the generalization performance of adversarially trained ViTs, this paper focuses on explicitly characterizing the upper bound of the robust risk $\mathfrak{R}_S(\widetilde{\ell}\circ\mathcal{F})$. Unlike the inherent locality of CNNs, the self-attention mechanism in Transformer blocks allows a single perturbed token to globally influence all other tokens. This global modeling of long-range interactions renders the optimization problem for identifying worst-case perturbations highly non-convex and analytically intractable. To circumvent this challenge, we derive a surrogate upper bound on the adversarial loss by analyzing the propagation of inputs through the ViTs architecture. Subsequently, we establish a novel risk bound based on the Rademacher complexity of this surrogate function.

4 Main Results

4.1 Adversarial Generalization Analysis for Single-Layer ViTs

We follow the definition of self-attention and Vision Transformers set forth by [Li *et al.*, 2023b]. For a single-layer

single-head ViTs containing a feed-forward neural network $\mathbf{W}_C \in \mathbb{R}^{d_1 \times d}$, with $\mathbf{W}_V \in \mathbb{R}^{d \times d_1}$, and $\mathbf{W}_Q, \mathbf{W}_K \in \mathbb{R}^{d \times m}$ as trainable weight matrices. Let σ be an L_σ-lipschitz activation function that is applied element-wise and has the property $\sigma(0) = 0$. Finally, a Vision Transformer block is defined as:

$$\sigma\Big(\text{softmax}\big(\mathbf{X}\mathbf{W}_Q\mathbf{W}_K^T\mathbf{X}^T\big)\mathbf{X}\mathbf{W}_V\Big)\mathbf{W}_C.$$

Since $\mathbf{W}_Q$ and $\mathbf{W}_K$ are only ever multiplied together, we combine them into a single matrix $\mathbf{W}_{QK} \in \mathbb{R}^{d \times d}$ for convenience of analysis. Finally, a learnable token $\mathbf{x}_{[\mathbf{cls}]} \in \mathbb{R}^d$ is prepended to the input token sequence $\mathbf{X}$, resulting in an augmented sequence of dimension $\mathbb{R}^{(m+1)\times d}$. We take the output feature corresponding to the [cls] token and denote it as $Z_{[\mathbf{cls}]} \in \mathbb{R}^d$. We also apply a trainable prediction vector $\mathbf{w} \in \mathbb{R}^{d \times K}$, and multiply it with $Z_{[\mathbf{cls}]}$ to get our output $\mathbf{w}^T Z_{[\mathbf{cls}]}$. In K-category classification problems, we define the label $y \in [K]$ and consider the hypothesis class $\mathcal{F} : \mathbb{R}^{(m+1)\times d} \to \mathbb{R}^K$.

Let $\Theta = \{\mathbf{w}, \mathbf{W}_{QK}, \mathbf{W}_V, \mathbf{W}_C\}$ denote the parameter set, then the function class $\mathcal{F}$ is defined as:

$$\mathcal{F} := \Big\{ f(\Theta, \mathbf{X}) = \sigma(\text{softmax}(\mathbf{X}\mathbf{W}_{QK}\mathbf{x}_{[\mathbf{cls}]})\mathbf{X}\mathbf{W}_V)\mathbf{W}_C\mathbf{w} : \big\|\mathbf{W}_C^T\big\|_2 \le \mathbf{B}_C, \big\|\mathbf{W}_V^T\big\|_2 \le \mathbf{B}_V \Big\}. \tag{2}$$

The primary difficulty in optimizing the adversarial loss for function class $\mathcal{F}$ stems from the presence of the inner maximization over the perturbation set $\mathcal{B}_\infty^\epsilon$. In prior work, Yin et al. [2019] adopted a surrogate loss based on semidefinite programming (SDP) relaxation to address the outer maximization problem associated with adversarial loss in multi-class classification tasks. However, this SDP relaxation-based surrogate loss method is inherently restricted to shallow architectures with a single hidden layer. Inspired by the work of [Wen *et al.*, 2025], we adopted a paired boundary analysis tailored for adversarial perturbations. This approach can be applied to both single-layer and multi-layer ViTs, deriving a tighter upper bound for the original adversarial loss.

Lemma 4.1. *Let the robust surrogate loss be defined by*

$$\widehat{\ell}(f(\Theta, \mathbf{X}^i), y) = \ell_\gamma\left(M(f(\Theta, \mathbf{X}^i), y) - \Lambda(f(\Theta, \mathbf{X}^i))\right),$$

where the worst-case error is:

$$\Lambda(f(\Theta, \mathbf{X}^i)) = 2L_\sigma\epsilon \max_{k\in[K]} \big\|\mathbf{w}_{*k}^T\big\|_1 \big\|\mathbf{W}_C^T\big\|_{1,\infty} \big\|\mathbf{W}_V^T\big\|_{1,\infty} \times \left(\frac{\big\|\mathbf{W}_{QK}\big\|_1 \big\|\mathbf{X}\big\|_{1,\infty} \big\|\mathbf{x}_{[\boldsymbol{cls}]}\big\|_\infty}{2} + 1\right).$$

Then, we have

$$\begin{aligned}&\max_{\widetilde{\mathbf{X}}\in\mathcal{B}_\infty^\epsilon(\mathbf{X})} \mathbb{1}\Big\{y^i \neq \arg\max_{y'\in[K]}[f(\Theta, \widetilde{\mathbf{X}}^i)]_{y'}\Big\} \\ &\le \widetilde{\ell}\big(f\big(\Theta, \mathbf{X}^i\big), y^i\big) \le \widehat{\ell}\big(f\big(\Theta, \mathbf{X}^i\big), y^i\big) \\ &\le \mathbb{1}\Big\{M\big(f\big(\Theta, \mathbf{X}^i\big), y^i\big) - \Lambda\big(f\big(\Theta, \mathbf{X}^i\big)\big) \le \gamma\Big\}.\end{aligned}$$

Remark 4.2. *The proof is provided in Appendix B. The surrogate loss consists of two components: $M\left(f\left(\mathbf{X}\right), y\right)$ maintains the optimization objective of the original classification task, while $\Lambda\left(f\left(\mathbf{X}\right)\right)$ quantifies the worst-case prediction deviation caused by adversarial perturbations. In the training phase, the magnitude of ϵ should be strictly controlled to prevent the model from excessively sacrificing its classification performance on standard data in the pursuit of robustness.*

Remark 4.3. *The inequality chain provides the theoretical justification for utilizing $\widehat{\ell}$ as a surrogate optimization objective. The first inequality establishes that $\widehat{\ell}$ serves as a rigorous upper bound on the intractable adversarial* 0-1 *risk, ensuring that minimizing this surrogate effectively suppresses the true worst-case error. Furthermore, the model is guaranteed to be robust against perturbations within $\mathcal{B}_\infty^\epsilon$ whenever the standard error $M(\cdot)$ is separated from the worst-case deviation $\Lambda(\cdot)$ by a margin γ.*

With the contraction inequality [Ledoux and Talagrand, 2013] and Lemma 4.1, we obtain the following structural result:

$$\mathfrak{R}_S(\widetilde{\ell} \circ \mathcal{F}) \le \mathfrak{R}_S(\widehat{\ell} \circ \mathcal{F}) \le \frac{1}{\gamma}\Big(\mathfrak{R}_S(M \circ \mathcal{F}) + \mathfrak{R}_S(\Lambda \circ \mathcal{F})\Big).$$

In the following theorem, we present the generalization bound of ViTs for multi-classification tasks by applying Lemma 3.2. The proof is provided in Appendix B.

Theorem 4.4. *Let $\mathcal{F} : \mathbb{R}^{m\times d} \to \mathbb{R}^K$ be the class of single-layer ViTs as defined in (2). Consider the robust surrogate loss defined in Lemma 4.1. For any fixed $\gamma > 0$, with probability at least $1 - \delta$,*

$$\begin{aligned}&\mathbb{P}_{(\mathbf{X},y)\sim\mathcal{D}}\Big\{\exists\widetilde{\mathbf{X}} \in \mathcal{B}_\infty^\epsilon\left(\mathbf{X}\right) \text{ s.t. } y \neq \arg\max_{y'\in[K]}[f(\Theta, \widetilde{\mathbf{X}})]_{y'}\Big\} \\ &\le \frac{1}{n}\sum_{i=1}^n \mathbb{1}\Big\{\left[f\left(\Theta, \mathbf{X}^i\right)\right]_{y^i} \le \gamma + \max_{y'\neq y^i}[f\left(\Theta, \mathbf{X}^i\right)]_{y'} \\ &\quad + \Lambda\left(f\left(\Theta, \mathbf{X}^i\right)\right)\Big\} + 2\mathfrak{R}_S(\widehat{\ell} \circ \mathcal{F}) + 3\sqrt{\frac{\log\frac{2}{\delta}}{2n}},\end{aligned}$$

where

$$\begin{aligned}\mathfrak{R}_S(\widehat{\ell} \circ \mathcal{F}) \le &\big(2 + K\epsilon\sqrt{2\log(2d)}\mathbf{B}_X^2 + C_1 m^{\frac{3}{2}}\big\|\mathbf{w}^T\big\|_2\mathbf{B}_X^3\big) \\ &\times L_\sigma\mathbf{B}_C\mathbf{B}_V\big\|\mathbf{W}_{QK}\big\|_1\gamma^{-1}n^{-1/2},\end{aligned}$$

with $C_1 = \sqrt{2}[\frac{\Gamma(3/2)}{\sqrt{\pi}}]^{1/2}$ and $\mathbf{B}_X = \max_i \|\mathbf{x}_i\|_2$.

Remark 4.5. *The theorem decomposes the upper bound of the true adversarial risk into three governing components: i) the empirical robust surrogate error, which measures the model's performance on training data under worst-case perturbations and margin constraints; ii) the Rademacher complexity term, which quantifies the capacity of the ViT hypothesis class; and iii) a confidence term that vanishes as sample size n increases. Crucially, the explicit bound on $\mathfrak{R}_S(\widehat{\ell} \circ \mathcal{F})$ reveals that the generalization gap scales with $\mathcal{O}(\mathbf{B}_C\mathbf{B}_V)$. This observation also implies a challenge for deep architectures: as the depth increases to L layers, the complexity*

bound tends to grow exponentially with the weight norms eg $\mathcal{O}(\mathbf{B}_C\mathbf{B}_V)^L$. *Consequently, without strict constraints, the multiplicative accumulation of weight norms can render the bound vacuous. This theoretical insight is empirically corroborated in Section 5, where applying* L_2 *regularization specifically to the* $\mathbf{W}_C$ *weights is shown to reduce the adversarial generalization error.*

4.2 Adversarial Generalization Analysis for Multi-Layer ViTs

The preceding analysis established the generalization bound for a single-layer ViTs. However, a single Transformer block has limited representational power and lacks the capacity to capture complex hierarchical features inherent in natural images. To achieve better performance, multiple Transformer blocks are typically stacked to construct deep architectures in practical applications [Dosovitskiy, 2020]. The standard multi-layer ViTs constructs a deep, expressive network by inductively stacking L core blocks. We inductively define an L-layer Vision Transformer block with layer normalization, denoting the weights of layer i by:

$$\mathbf{W}^{(i)} := \{\mathbf{W}_{QK}^{(i)}, \mathbf{W}_V^{(i)}, \mathbf{W}_C^{(i)}\}.$$

And we denote the set of weights up to layer i by:

$$\mathbf{W}^{1:i} = \{\mathbf{W}^{(1)}, \ldots, \mathbf{W}^{(i-1)}\}.$$

Define the layer function:

$$f\left(\mathbf{X}; \mathbf{W}^{(i)}\right) = \sigma\left(\text{softmax}\left(\mathbf{X}\mathbf{W}_{QK}^{(i)}\mathbf{X}^T\right)\mathbf{X}\mathbf{W}_V^{(i)}\right).$$

Initialize the base case:

$$g_{\text{block}}^1\left(\mathbf{X}; \mathbf{W}^{1:1}\right) = \mathbf{X}.$$

The output of layer $i+1$ is:

$$g_{\text{block}}^{(i+1)}\left(\mathbf{X}; \mathbf{W}^{1:i+1}\right) =$$

$$\Pi_{\text{norm}}\left(\sigma\left(\Pi_{\text{norm}}\left(f\left(g_{\text{block}}^{(i)}\left(\mathbf{X}; \mathbf{W}^{1:i}\right); \mathbf{W}^{(i)}\right)\right)\right)\mathbf{W}_C^{(i)}\right),$$

where Π_{norm} projects each row onto the unit ball. Unlike the unconstrained propagation in shallow models, the operator Π_{norm} ensuring that feature magnitudes remain bounded as depth increases. Note that the norm of the entire input can still have a dependence on m.

Let the class of L-layer transformer blocks be defined as:

$$\mathcal{G}_{\text{block}}^{(L)} := \Big\{\mathbf{X} \mapsto g_{\text{block}}^{(L+1)}\left(\mathbf{X}; \mathbf{W}^{1:L+1}\right) : \forall i \in [L],$$

$$\left\|\mathbf{W}_V^{(i)}\right\|_2 \le \mathbf{B}_V^{(i)}, \left\|\mathbf{W}_C^{(i)}\right\|_2 \le \mathbf{B}_C^{(i)}, \|\mathbf{W}_{QK}^{(i)}\|_2 \le \mathbf{B}_{QK}^{(i)}\Big\}.$$

Then we define the adversarial scalar output function class be:

$$\mathcal{G}_{\text{scalar}}^{(L)} := \left\{\mathbf{X} \to \mathbf{w}^T g(\mathbf{X}) : g \in \mathcal{G}_{\text{block}}^{(L)}, \mathbf{w} \in \mathbb{R}^{d\times K}\right\}. \quad (3)$$

Analogous to the single-layer analysis, explicitly characterizing the adversarial perturbation for multi-layer ViTs is challenging since the intractability of adversarial Loss. Consequently, we extend the surrogate loss framework to accommodate deep architectures. This robust surrogate objective provides a tractable upper bound for the empirical adversarial loss, which is essential for establishing the adversarial generalization bounds.

Lemma 4.6. *Let the robust surrogate loss be defined by*

$$\widehat{\ell}\left(\mathcal{G}\left(\mathbf{X}^i; \mathbf{W}^{1:L}, \mathbf{w}\right), y^i\right) =$$
$$\ell_\gamma\left(M\left(\mathcal{G}\left(\mathbf{X}^i; \mathbf{W}^{1:L}, \mathbf{w}\right), y^i\right) - \Lambda\left(\mathcal{G}\left(\mathbf{X}^i; \mathbf{W}^{1:L}, \mathbf{w}\right)\right)\right),$$

where the worst-case error is:

$$\Lambda\big(\mathcal{G}\big(\mathbf{X}; \mathbf{W}^{1:L}, \mathbf{w}\big)\big) = L_\sigma^L \max_{k\in[K]} \|\mathbf{w}_{*k}^T\|_1$$
$$\times \prod_{l=1}^{L}\left[\mathbf{B}_C^{(l)}\mathbf{B}_V^{(l)}\left(1 + \mathbf{B}_{QK}^{(l)}\Psi_{l+1}(\Psi_{l+1} + \Psi'_{l+1})\right)\right].$$
$$Here \quad \Psi_{l+1} = C_2\left\|\mathbf{W}_V^1\right\|_2\sqrt{m}\mathbf{B}_X,$$
$$and \quad \Psi'_{l+1} = C_2\left\|\mathbf{W}_V^1\right\|_1\left(\sqrt{d}\mathbf{B}_X + \epsilon\right),$$

the constant $C_2 = L_\sigma^L\|\mathbf{W}_C^1\|_2\prod_{l=2}^{L}\mathbf{B}_C^{(l)}\mathbf{B}_V^{(l)}$ *is composed of the parameter-norm bounds across layers.*

Remark 4.7. *The robust surrogate loss* $\widehat{\ell}(\cdot)$ *fundamentally captures the tension between clean sample performance and adversarial vulnerability. The construction of* Ψ_{l+1} *and* Ψ'_{l+1} *emerges from our layer-wise analysis, incorporating cumulative weight dependencies absent in Lemma 4.1, where* Ψ_{l+1} *bounds the norm amplification of clean inputs through successive transformer blocks and* Ψ'_{l+1} *extends this to adversarial examples by incorporating the perturbation* ϵ.

Theorem 4.8. *Let* $\mathcal{G} : \mathbb{R}^{m\times d} \to \mathbb{R}^K$ *be the class of L-layers ViTs as defined in (3). Consider the robust surrogate loss defined in Lemma 4.6. For any fixed* $\gamma > 0$, *with probability at least* $1 - \delta$,

$$\mathbb{P}_{(\mathbf{X},y)\sim\mathcal{D}}\left\{\exists\widetilde{\mathbf{X}} \in \mathcal{B}_\infty^\epsilon(\mathbf{X}) \text{ s.t. } y \neq \arg\max_{y'\in[K]}\mathcal{G}(\widetilde{\mathbf{X}})_{y'}\right\}$$
$$\le \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\Big\{\mathcal{G}(\widetilde{\mathbf{X}}^i)_{y^i} \le \gamma + \Lambda\left(\mathcal{G}\left(\mathbf{X}^i; \mathbf{W}^{1:L}, \mathbf{w}\right)\right)$$
$$+ \max_{y'\neq y^i}\mathcal{G}\left(\mathbf{X}^i\right)_{y'}\Big\} + 2\mathfrak{R}_S(\widehat{\ell}\circ\mathcal{G}_{scalar}^{(L)}) + 3\sqrt{\frac{\log\frac{2}{\delta}}{2n}},$$

where:

$$\mathfrak{R}_S(\widehat{\ell}\circ\mathcal{G}_{scalar}^{(L)}) \le$$
$$\frac{L_\sigma^L\mathcal{O}(\mathbf{B}_C\mathbf{B}_V)^L}{\gamma\sqrt{n}}\big(2K\sqrt{2\log 2d}\left\|\mathbf{W}_V^T\right\|_1\mathbf{\Omega} + \sqrt{m}\|\mathbf{w}\|_2\big)\mathbf{B}_X,$$

with $\mathbf{\Omega} = \prod_{l\in[L]}\left(1 + \mathbf{B}_{QK}^{(l)}\Psi_{l+1}(\Psi_{l+1} + \Psi'_{l+1})\right)$.

Remark 4.9. *The detailed proof is provided in Appendix C. This theorem establishes a precise connection between the adversarial generalization performance of deep ViTs and their architectural properties. The bound we derived explicitly captures the exponential dependence of the depth L on*

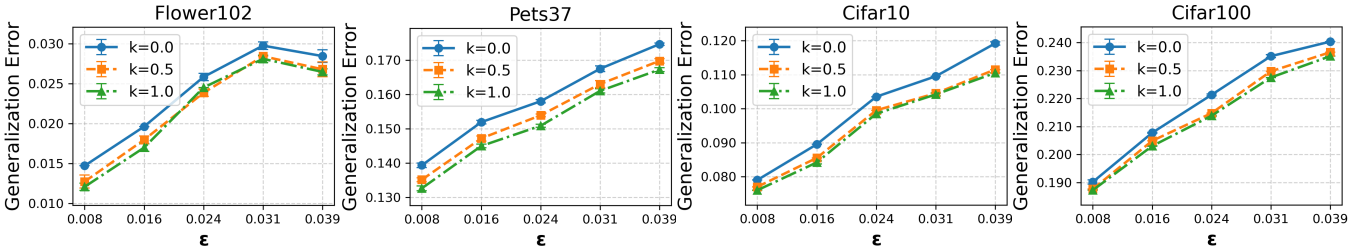

Figure 1: Empirical evaluation of generalization performance (mean ± std) for models trained with L_2 regularization under varying parameter k. ϵ denotes the maximum allowable adversarial perturbation.

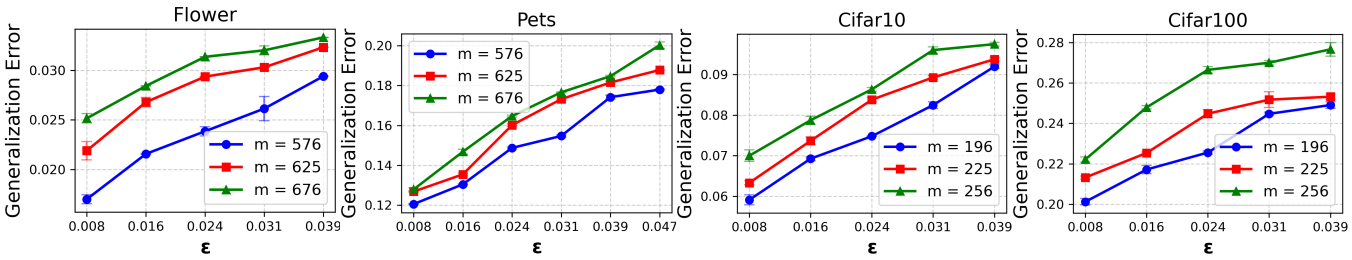

Figure 2: Empirical evaluation of generalization performance (mean ± std) for models trained with different sequence length. ϵ denotes the maximum allowable adversarial perturbation.

the Lipschitz constant of activation function and the norm of the weight matrix. To maintain adversarial robustness while leveraging the powerful expressive capacity of deep networks, it is necessary to strictly constrain the norms of weight matrices and employ activation functions with Lipschitz constants not exceeding 1 (e.g., sigmoid or relu), thereby controlling the propagation of perturbations through the network. In addition, the width-dependent term $\sqrt{\log(d)/n}$ provides a theoretical justification for scaling network width d proportionally with depth in robust ViTs architectures, suggesting that wider architectures can effectively counterbalance the complexity induced by increased depth.

Remark 4.10. *Our analysis reveals an explicit dependence of the generalization bound on the cumulative product of layer-wise norms. Given that the global self-attention mechanism can maximize the propagation of worst-case perturbations (captured by Ψ'_{l+1}), introducing structural sparsity into the attention matrix serves as an effective mechanism to reduce the block's effective Lipschitz constant. The Drop Attention mechanism evaluated in Section 5, as a theoretical safeguard that works in tandem with weight regularization to enhance the adversarial robustness of deep ViTs.*

Dataset	**Size** (n)	**Dimension** (d)
Oxford-IIIT Pets	7,349	224×224×3
Flowers-102	8,189	224×224×3
CIFAR-10	60,000	32×32×3
CIFAR-100	60,000	32×32×3

Table 2: The adopted datasets details.

5 Experiments

In this section, we conduct several experiments to validate our theoretical results.

Datasets

Our experiments are validated on multi-domain image classification datasets (Oxford-IIIT Pets, Flowers-102, CIFAR-10 and CIFAR-100). During fine-tuning, CIFAR-10/100 are adjusted to a fixed resolution 224×224 to adapt to the model, while Pets and Flowers-102 are resized to 384×384. Statistics of datasets are provided in Table 2.

Model Settings

We employ the ViTs-Base model pre-trained by [Dosovitskiy, 2020] (available on GitHub with ImageNet-21k pretrained weights) as the backbone. For fine-tuning, we strictly adhere to the original configuration, optimizing the model using Stochastic Gradient Descent (SGD) with a momentum of 0.9, a learning rate of 0.001, and training for 100 epochs.

We formulate the robust training objective as:

$$\min_{f\in\mathcal{F}} \sum_{i=1}^{n} \max_{\widetilde{\mathbf{X}}\in\mathcal{B}^{\epsilon}_{\infty}(\mathbf{X})} \ell(f(\widetilde{\mathbf{X}}), y) + k\|\mathbf{W}_C\|_2, \tag{4}$$

where $\ell(\cdot)$ is cross-entropy loss. Guided by our theoretical findings in Theorem 4.4 and 4.8 , we introduce L_2 regularization term for the weight matrix $\mathbf{W}_C$ in the first MLP layer, with the coefficient $k \in \{0, 0.5, 1.0\}$. This design aims to constrain the perturbation amplification without suppressing the representational power of deeper layers. For the inner maximization problem, we employ Projected Gradient Descent (PGD) [Madry *et al.*, 2018] to minimize the objective (4). The perturbation budget ϵ is set to $\{2, 4, 6, 8, 10\}/255$, with step size $\alpha = \varepsilon/4$ and 5 steps . During the evaluation,

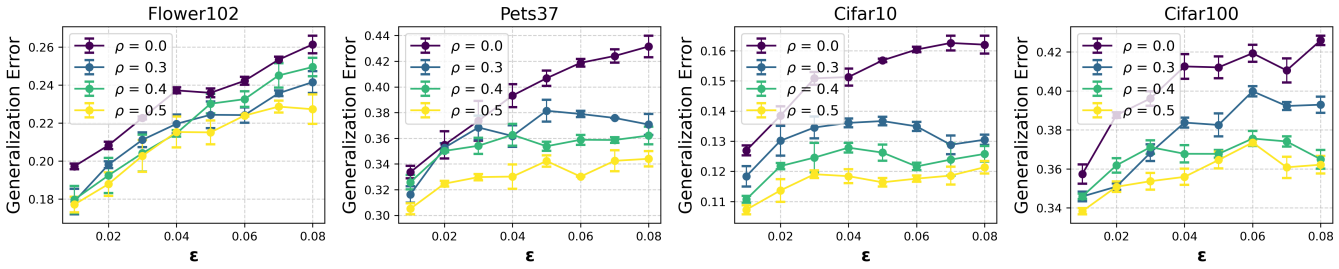

Figure 3: Impact of the Drop-Key mechanism on adversarial generalization error. The figures illustrate the generalization error under different drop ratios. ϵ denotes the maximum allowable adversarial perturbation. ρ denotes the drop-key ratio.

we test the model's robustness using the same PGD settings to ensure consistency between training and inference.

Generalization Analysis

Each experiment is independently repeated 10 times. Following the line of [Yin *et al.*, 2019; Xiao *et al.*, 2022], we quantify generalization performance of ViT model by calculating the generalization error:

$$|\text{Adversarial training accuracy} - \text{Adversarial test accuracy}|.$$

This metric serves as an empirical proxy for the theoretical generalization gap $\mathcal{E}(f)$ defined in Section 3.

5.1 Numerical Discussion

Regularization

Figure 1 illustrates the adversarial generalization error under PGD attacks with varying regularization coefficients k. We observe that the empirical generalization error for regularized models ($k \in 0.5, 1.0$) is consistently lower than that of the unregularized baseline ($k = 0$). These findings provide empirical corroboration for Theorem 4.8, confirming that explicitly constraining the norm of the weight matrix $\mathbf{W}_C$ effectively restricts the hypothesis space complexity and narrows the adversarial generalization gap.

Sequence Length

To investigate the impact of sequence length on generalization, we conduct experiments with varying token counts m. Since the sequence length is determined by the input resolution ($H \times W$) and a fixed patch size ($P = 16$) via $m = HW/P^2$, we adjusted the input resolutions to obtain different m. For CIFAR-10/100, we varied resolutions from 224×224 to 256×256, resulting in $m \in \{196, 225, 256\}$. For high-resolution datasets (Flowers-102 and Pets), we adopted resolutions of 384×384, 400×400, and 416×416, corresponding to $m \in \{576, 625, 676\}$. As illustrated in Figure 2, the empirical generalization error exhibits a monotonically decreasing trend as dimensionality decreases, suggesting the beneficial role of low-dimensional feature projection in generalization error reduction.

Sparsity Implementation

To investigate the impact of attention sparsity, we utilize the ViT-Tiny architecture (patch size 16, 224×224 input resolution) as the backbone for ablation studies. Motivated by the theoretical observation that constraining the attention matrix norm yields a tighter generalization bound, we implement structural sparsity via a Drop-Key mechanism. Specifically, following [Li *et al.*, 2023a], we incorporate a Drop-Key ratio $\rho \in [0, 0.5]$ and generate a binary mask $\mathbf{M} \sim \text{Bernoulli}(1-\rho)$ during training. This mask is applied to the attention scores $\mathbf{QK}^T$ prior to the softmax operation, effectively setting a random subset of attention weights to $-\infty$ (i.e., zero probability after softmax). This forces the model to rely on a sparse set of features.

As illustrated in Figure 3, the Drop-Key mechanism significantly narrows the generalization gap, and its regularization effect is amplified as the attack strength ϵ grows. While the baseline model's performance degrades sharply under large ϵ, models trained with higher Drop-Key ratios maintain a much tighter alignment between training and testing performance. This empirical evidence supports our theoretical assertion that imposing sparsity on the attention matrix tightens the bound on the worst-case error, thereby enhancing the model's intrinsic robustness.

6 Conclusions

This paper analyzes the generalization performance of ViTs under adversarial perturbations through the lens of Rademacher complexity. Our derived bounds characterize the intricate interplay between adversarial robustness and fundamental architectural parameters, including perturbation magnitude ϵ, input dimensionality d, and layer-wise weight norms. A crucial insight from our analysis is that the multiplicative accumulation of complexity with depth can be mitigated by constraining weight norms and attention mechanisms. Extensive experiments on benchmark datasets validate these findings, demonstrating that strategies such as ℓ_2 regularization on MLP weights and attention sparsity can reduce the adversarial generalization gap. These results effectively bridge the disconnect between theoretical guarantees and practical architectural design for robust ViTs.

Acknowledgments

This work is supported in part by Guangxi Key Technologies R&D Program No. AB25069496.

References

[Attias *et al.*, 2022] Idan Attias, Aryeh Kontorovich, and Yishay Mansour. Improved generalization bounds for ad-

versarially robust learning. *Journal of Machine Learning Research*, 23:1–31, 2022.

[Awasthi *et al.*, 2020] Pranjal Awasthi, Natalie Frank, and Mehryar Mohri. Adversarial learning guarantees for linear hypotheses and neural networks. In *Proceedings of the International Conference on Machine Learning*, pages 431–441, 2020.

[Bai *et al.*, 2021] Tao Bai, Jinqi Luo, Jun Zhao, Bihan Wen, and Qian Wang. Recent advances in adversarial training for adversarial robustness. *arXiv preprint arXiv:2102.01356*, 2021.

[Bai *et al.*, 2023] Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. Transformers as statisticians: Provable in-context learning with in-context algorithm selection. *Advances in Neural Information Processing Systems*, 36:57125–57211, 2023.

[Bhojanapalli *et al.*, 2021] Srinadh Bhojanapalli, Ayan Chakrabarti, Daniel Glasner, Daliang Li, Thomas Unterthiner, and Andreas Veit. Understanding robustness of transformers for image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10231–10241, 2021.

[Chen *et al.*, 2022] Zhaoyu Chen, Bo Li, Jianghe Xu, Shuang Wu, Shouhong Ding, and Wenqiang Zhang. Towards practical certifiable patch defense with vision transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15148–15158, 2022.

[Deng *et al.*, 2025] Haoyu Deng, Yanmei Fang, and Fangjun Huang. Patch attacks on vision transformer via skip attention gradients. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 554–567, Singapore, 2025.

[Deora *et al.*, 2023] Puneesh Deora, Rouzbeh Ghaderi, Hossein Taheri, and Christos Thrampoulidis. On the optimization and generalization of multi-head attention. *arXiv preprint arXiv:2310.12680*, 2023.

[Dosovitskiy, 2020] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[Edelman *et al.*, 2022] Benjamin L Edelman, Surbhi Goel, Sham Kakade, and Cyril Zhang. Inductive biases and variable creation in self-attention mechanisms. In *Proceedings of the International Conference on Machine Learning*, pages 5793–5831, 2022.

[Fang *et al.*, 2023] Yuxin Fang, Shusheng Yang, Shijie Wang, Yixiao Ge, Ying Shan, and Xinggang Wang. Unleashing vanilla vision transformer with masked image modeling for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6244–6253, 2023.

[Gong *et al.*, 2024] Huihui Gong, Minjing Dong, Siqi Ma, Seyit Camtepe, Surya Nepal, and Chang Xu. Random entangled tokens for adversarially robust vision transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24554–24563, 2024.

[Gopal *et al.*, 2025] Bhavna Gopal, Huanrui Yang, Mark Horton, and Yiran Chen. Safer: Sharpness aware layer-selective finetuning for enhanced robustness in vision transformers. *arXiv preprint arXiv:2501.01529*, 2025.

[Gu *et al.*, 2022] Jindong Gu, Volker Tresp, and Yao Qin. Are vision transformers robust to patch perturbations? In *European Conference on Computer Vision*, pages 404–421, 2022.

[Gu *et al.*, 2024] Junyi Gu, Mauro Bellone, Tomáš Pivoňka, and Raivo Sell. Clft: Camera-lidar fusion transformer for semantic segmentation in autonomous driving. *IEEE Transactions on Intelligent Vehicles*, page 1–12, 2024.

[Guo *et al.*, 2023] Yong Guo, David Stutz, and Bernt Schiele. Robustifying token attention for vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17557–17568, 2023.

[Herrmann *et al.*, 2022] Charles Herrmann, Kyle Sargent, Lu Jiang, Ramin Zabih, Huiwen Chang, Ce Liu, Dilip Krishnan, and Deqing Sun. Pyramid adversarial training improves vit performance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13419–13429, 2022.

[Jain and Dutta, 2024] Samyak Jain and Tanima Dutta. Towards understanding and improving adversarial robustness of vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24736–24745, 2024.

[Kim *et al.*, 2021] Hyunjik Kim, George Papamakarios, and Andriy Mnih. The lipschitz constant of self-attention. In *Proceedings of the International Conference on Machine Learning*, pages 5562–5571, 2021.

[Ledoux and Talagrand, 2013] Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media, 2013.

[Li *et al.*, 2022] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *European conference on computer vision*, pages 280–296, 2022.

[Li *et al.*, 2023a] Bonan Li, Yinhan Hu, Xuecheng Nie, Congying Han, Xiangjian Jiang, Tiande Guo, and Luoqi Liu. Dropkey for vision transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22700–22709, 2023.

[Li *et al.*, 2023b] Hongkang Li, Meng Wang, Sijia Liu, and Pin-Yu Chen. A theoretical understanding of shallow vision transformers: Learning, generalization, and sample complexity. In *International Conference on Learning Representations*, 2023.

[Li *et al.*, 2023c] Yingcong Li, Muhammed Emrullah Ildiz, Dimitris Papailiopoulos, and Samet Oymak. Transformers as algorithms: Generalization and stability in in-context learning. In *Proceedings of the International Conference on Machine Learning*, pages 19565–19594, 2023.

[Madry *et al.*, 2018] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

[Mahmood *et al.*, 2021] Kaleel Mahmood, Rigel Mahmood, and Marten Van Dijk. On the robustness of vision transformers to adversarial examples. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7838–7847, 2021.

[Manzari *et al.*, 2023] Omid Nejati Manzari, Hamid Ahmadabadi, Hossein Kashiani, Shahriar B. Shokouhi, and Ahmad Ayatollahi. Medvit: A robust vision transformer for generalized medical image classification. *Computers in Biology and Medicine*, 157:106791, 2023.

[Mao *et al.*, 2022] Xiaofeng Mao, Gege Qi, Yuefeng Chen, Xiaodan Li, Ranjie Duan, Shaokai Ye, Yuan He, and Hui Xue. Towards robust vision transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12042–12051, 2022.

[Mo *et al.*, 2022] Yichuan Mo, Dongxian Wu, Yifei Wang, Yiwen Guo, and Yisen Wang. When adversarial training meets vision transformers: Recipes from training to architecture. In *Advances in Neural Information Processing Systems*, volume 35, pages 18599–18611, 2022.

[Mustafa *et al.*, 2022] Waleed Mustafa, Yunwen Lei, and Marius Kloft. On the generalization analysis of adversarial learning. In *Proceedings of the International Conference on Machine Learning*, pages 16174–16196, 2022.

[Shao *et al.*, 2022] Rulin Shao, Zhouxing Shi, Jinfeng Yi, Pin-Yu Chen, and Cho-jui Hsieh. On the adversarial robustness of vision transformers. In *Advances in Neural Information Processing Systems*, 2022.

[Sultana *et al.*, 2022] Maryam Sultana, Muzammal Naseer, Muhammad Haris Khan, Salman Khan, and Fahad Shahbaz Khan. Self-distilled vision transformer for domain generalization. In *Proceedings of the Asian conference on computer vision*, pages 3068–3085, 2022.

[Thisanke *et al.*, 2023] Hans Thisanke, Chamli Deshan, Kavindu Chamith, Sachith Seneviratne, Rajith Vidanaarachchi, and Damayanthi Herath. Semantic segmentation using vision transformers: A survey. *Engineering Applications of Artificial Intelligence*, 126:106669, 2023.

[Tramèr *et al.*, 2018] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. In *International Conference on Learning Representations*, 2018.

[Trauger and Tewari, 2024] Jacob Trauger and Ambuj Tewari. Sequence length independent norm-based generalization bounds for transformers. In *Proceedings of the International Conference on Artificial Intelligence*, pages 1405–1413, 2024.

[Wang *et al.*, 2024] Sibo Wang, Jie Zhang, Zheng Yuan, and Shiguang Shan. Pre-trained model guided fine-tuning for zero-shot adversarial robustness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24502–24511, 2024.

[Wei *et al.*, 2022] Zhipeng Wei, Jingjing Chen, Micah Goldblum, Zuxuan Wu, Tom Goldstein, and Yu-Gang Jiang. Towards transferable adversarial attacks on vision transformers. In *Proceedings of the International Conference on Artificial Intelligence*, volume 36, pages 2668–2676, 2022.

[Wen *et al.*, 2023] Wen Wen, Han Li, Hong Chen, Rui Wu, Lingjuan Wu, and Liangxuan Zhu. Generalization bounds for adversarial metric learning. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 4397–4405, 2023.

[Wen *et al.*, 2025] Wen Wen, Han Li, Tieliang Gong, and Hong Chen. Towards generalization bounds of GCNs for adversarially robust node classification. In *International Conference on Learning Representations*, 2025.

[Wu and Cheng, 2021] xinxing Wu and Qiang Cheng. Algorithmic stability and generalization of an unsupervised feature selection algorithm. In *Advances in Neural Information Processing Systems*, volume 34, pages 19860–19875, 2021.

[Wu *et al.*, 2022] Boxi Wu, Jindong Gu, Zhifeng Li, Deng Cai, Xiaofei He, and Wei Liu. Towards efficient adversarial training on vision transformers. In *European Conference on Computer Vision*, pages 307–325, 2022.

[Xiao *et al.*, 2022] Jiancong Xiao, Yanbo Fan, Ruoyu Sun, Jue Wang, and Zhi-Quan Luo. Stability analysis and generalization bounds of adversarial training. *Advances in Neural Information Processing Systems*, 35:15446–15459, 2022.

[Xiao *et al.*, 2025] Chaoen Xiao, Sirui Peng, Lei Zhang, Jianxin Wang, Ding Ding, and Jianyi Zhang. A transformer-based adversarial network framework for steganography. *Expert Systems with Applications*, 269:126391, 2025.

[Xu *et al.*, 2023] Xiaoyun Xu, Shujian Yu, Zhuoran Liu, and Stjepan Picek. Mimir: Masked image modeling for mutual information-based adversarial robustness. *arXiv preprint arXiv:2312.04960*, 2023.

[Yao *et al.*, 2024] Yuchong Yao, Nandakishor Desai, and Marimuthu Palaniswami. Adversarial masked autoencoders are robust vision learners. *IEEE Transactions on Artificial Intelligence*, 2024.

[Yin *et al.*, 2019] Dong Yin, Ramchandran Kannan, and Peter Bartlett. Rademacher complexity for adversarially robust generalization. In *Proceedings of the International Conference on Machine Learning*, pages 7085–7094, 2019.

[Yun *et al.*, 2019] Chulhee Yun, Srinadh Bhojanapalli, Ankit Singh Rawat, Sashank Reddi, and Sanjiv Kumar. Are transformers universal approximators of sequence-to-sequence functions? In *International Conference on Learning Representations*, 2019.