Gradient Correlation Whitening and Wavelet Packet Decomposition Based Transferable Adversarial Attacks on Vision Transformers

Shunhui Ji[1], Xu Zhang[1], Yunhe Li[1], Hai Dong[2], Yingying Shen[1], Pengcheng Zhang[1*]
[1]College of Computer Science and Software Engineering, Hohai University, Nanjing, China
[2]School of Computing Technologies, RMIT University, Melbourne, Australia

***Abstract*—Vision Transformers (ViTs) have demonstrated remarkable performance across various vision tasks, yet they remain vulnerable to adversarial attacks. While transfer-based attacks offer a practical black-box threat, existing methods often suffer from limited cross-architecture transferability due to token-level gradient overfitting and inadequate handling of frequency-sensitive features across models. In this work, we propose a novel method that integrates Gradient Correlation Whitening (GCW) and Wavelet Packet Decomposition (WPD) to address these limitations. GCW mitigates gradient overfitting by applying a global whitening transformation to decorrelate token-wise gradients, thereby reducing structural dependency on the source ViT. Meanwhile, WPD replaces the conventional Discrete Cosine Transform with a multi-level wavelet packet decomposition, enabling finer-grained frequency analysis and more adaptive perturbation enhancement across diverse spectral bands. Experiments on five ViTs and four CNNs demonstrate that our method surpasses the previous state-of-the-art in transferable adversarial attacks, achieving an average attack success rate improvement of 2.1% across models. Code is available at https://github.com/lingyu700/GCW-WPD.**

***Index Terms*—Vision Transformers, Transferability, Black-box attack, Adversarial perturbation**

I. Introduction

Vision Transformers (ViTs) have emerged as a dominant architecture in computer vision, achieving state-of-the-art performance across various image recognition tasks, including image classification [1], object detection [2], and medical imaging [3]. However, like convolutional neural networks (CNNs), ViTs remain susceptible to adversarial attacks—carefully crafted perturbations that can deceive models while remaining nearly imperceptible to humans [4], [5]. In practical scenarios, attackers often operate under black-box conditions, where they cannot access the target model's internal parameters or gradients [6]. Transfer-based attacks offer a viable strategy in such settings: an adversary generates adversarial examples on a surrogate model and transfers them to the target model [7], [8].

Despite their practical relevance, transfer-based attacks on ViTs often exhibit limited cross-architecture transferability, largely because the generated adversarial perturbations become overly dependent on the specific attention patterns of the source model [9], [10]. Unlike CNNs, which rely on localized feature extraction, ViTs employ global self-attention mechanisms that introduce unique vulnerabilities. One key manifestation of this issue is token-level gradient overfitting: during backpropagation, gradients corresponding to different image patches show high structural similarity, as visualized in Fig. 2. This elevated cosine similarity between token gradients causes the attack to overfit to the surrogate model's attention structure, thereby hindering generalization to other architectures. Although recent gradient regularization methods, such as Pay No Attention (PNA) [11] and Token Gradient Regularization (TGR) [12], attempt to alleviate this by discarding or scaling certain gradients, they typically operate from a local perspective, which can discard meaningful gradient information and restrict overall attack effectiveness.

In parallel, frequency-domain attack strategies have shown promise in enhancing transferability by exploiting spectral sensitivities across models [9]. Existing methods, such as those based on the Discrete Cosine Transform (DCT), perform global frequency transformations but lack the ability to capture localized frequency structures [13], [14]. This limits their adaptability to models that respond differently to high-frequency details, edges, and textures.

To address these challenges, we introduce a novel method: Gradient Correlation Whitening and Wavelet Packet Decomposition (GCW-WPD). Our approach integrates Gradient Correlation Whitening (GCW), a global gradient decorrelation strategy that reduces token-wise gradient similarity through a whitening transformation, breaking the structural dependency on the surrogate ViT. Complementarily, we propose Wavelet Packet Decomposition (WPD) to replace conventional DCT, enabling multi-level, sub-band-specific frequency analysis and perturbation enhancement. This allows the attack to adaptively target the most influential frequency components across diverse models. By synergistically combining GCW and WPD, our method not only alleviates gradient overfitting at its source but also enriches input diversity in the frequency domain, leading to adversarial examples with markedly improved cross-architecture transferability.

In sum, we make the following main contributions:

- **Gradient Correlation Whitening.** We propose GCW, a gradient regularization method that applies a whitening

The work is supported by the National Natural Science Foundation of China (62272145 and U21B2016) and the Henan Provincial Key Research and Development Program (251111210500).

* Corresponding Author: pchzhang@hhu.edu.cn

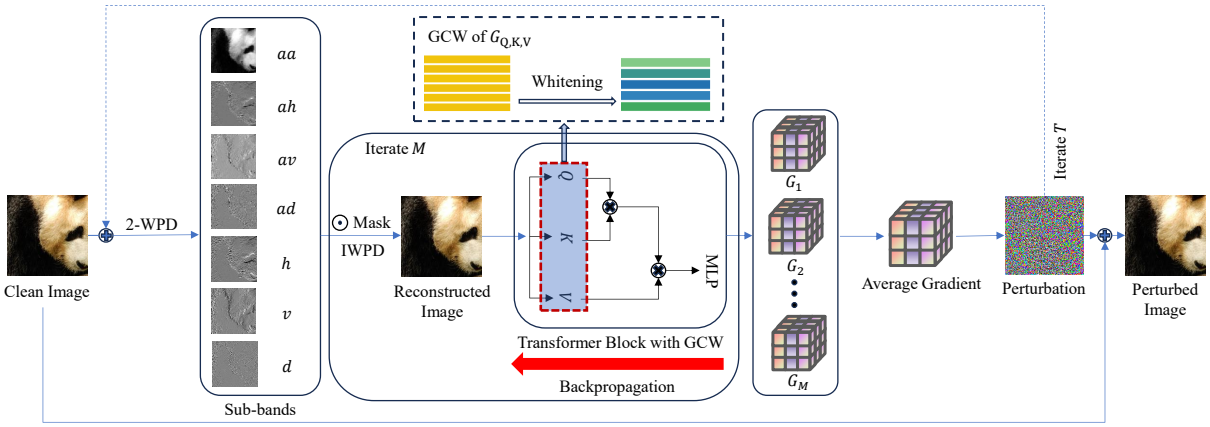


Fig. 1: Overview of the proposed GCW-WPD method.

transformation to decorrelate token gradients globally, reducing overfitting with source model.

- **Wavelet Packet Decomposition.** We introduce WPD for frequency-aware attacks, which provides finer-grained, adaptive control over perturbation enhancement across multiple frequency sub-bands.
- **Experimental Evaluation.** Experiments on five ViTs and four CNNs demonstrate that our method surpasses state-of-the-art methods. It achieves an average attack success rate improvement of 2.1% on different models, confirming its superior cross-architecture transferability.

II. Related Work

A. Adversarial Attacks on CNNs

Extensive research has been conducted on transferable adversarial attacks in the context of CNNs. Under the white-box setting, attacks such as Projected Gradient Descent (PGD) [15] generate perturbations by iteratively maximizing the model's loss. For black-box transfer attacks, the Momentum Iterative Fast Gradient Sign Method (MI-FGSM) [16] stabilizes the update direction with a momentum term. The Skip Gradient Method (SGM) [17] further explores the architectural vulnerability by leveraging gradients from skip connections. These approaches lay the foundation for attacks targeting ViTs.

B. Adversarial Attacks on Vision Transformers

The unique architecture of ViTs, particularly their global self-attention mechanism, introduces distinct adversarial vulnerabilities. For instance, PNA [11] discards gradients within the attention module. Token Gradient Regularization (TGR) [12] suppresses extreme gradient values to reduce variance, and Virtual Dense Connection (VDC) [18] smooths gradient propagation via virtual skip connections. More recently, TGD-MFA [14] combines gradient regularization with frequency enhancement using DCT. However, these methods often operate from a local perspective in gradient space and employ global frequency transformations, which may not fully capture fine-grained frequency structures. Consequently, they fall short in simultaneously addressing token-level gradient overfitting and adapting to model-specific frequency sensitivities, limiting cross-model transferability. To address these limitations, we propose the GCW-WPD method, which integrates global gradient whitening with multi-resolution wavelet packet decomposition for more transferable adversarial attacks.

III. Method

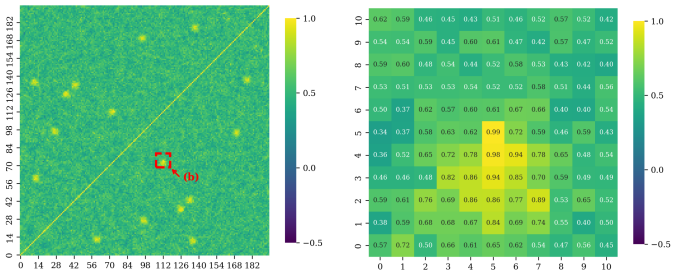
A. Problem Formulation and Overall Framework

Given a clean image $x \in \mathbb{R}^{H\times W\times C}$ with height H, width W, and channels C, and its ground-truth label y, the goal of a transferable adversarial attack is to craft an adversarial example $x^{adv} = x+\delta$ that can mislead not only the white-box surrogate model f_s but also a black-box target model f_t. The perturbation δ is constrained within an ℓ_∞-norm ball: $\|\delta\|_\infty \leq \epsilon$, where ϵ is the perturbation budget ensuring imperceptibility. The objective can be formulated as:

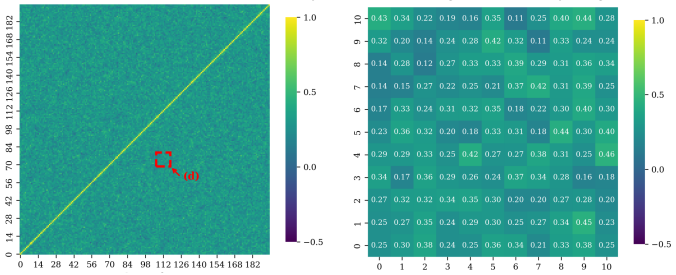
$$\max_{\delta} \ \mathcal{L}(f_t(x+\delta), y) \quad \text{s.t.} \quad \|\delta\|_\infty \leq \epsilon, \tag{1}$$

where $\mathcal{L}$ is the loss function (e.g., cross-entropy). Since f_t is inaccessible in a black-box setting, the attack is performed by maximizing the loss of the surrogate model f_s, relying on the transferability of adversarial perturbations.

As illustrated in Fig. 1, our proposed GCW-WPD method operates in two complementary dimensions to enhance transferability: the gradient domain and the frequency domain. During the iterative attack process, each iteration involves two core stages. First, in the frequency enhancement stage, the current adversarial image is decomposed in the frequency domain via WPD. Instead of using a global DCT, WPD provides a multi-resolution analysis, generating several frequency sub-bands. A mask is computed for each sub-band based on its gradient sensitivity, and these masks guide a stochastic enhancement process that simulates diverse model frequency responses. The enhanced frequency components are then reconstructed back to the spatial domain. Second, in the gradient modification stage, the backpropagated token gradients from the ViT are intercepted and processed by our GCW module.



(a) Token-level cosine similarity (b) High-similarity region



(c) After GCW (d) Same region after GCW

Fig. 2: Gradient similarity visualization before and after whitening. Token gradients exhibit high structural similarity in the surrogate ViT (a, b), which is effectively reduced by Gradient Correlation Whitening (c, d).

GCW applies a global whitening transformation to decorrelate the token-wise gradients, thereby alleviating the token-level gradient overfitting specific to the source ViT's attention patterns. The modified gradients are then used to update the adversarial perturbation. This dual-stream process—gradient decorrelation via GCW and frequency-aware diversification via WPD—forms a synergistic loop that progressively guides the perturbation towards more generalizable directions.

B. Gradient Correlation Whitening

A key obstacle to transferability in ViTs is the high structural similarity among gradients of different tokens, leading to overfitting on the source model's attention landscape. Let $G \in \mathbb{R}^{N\times D}$ denote the matrix of token gradients for a layer, where N is the number of tokens (excluding the class token) and D is the gradient dimension. The empirical covariance matrix Σ of the centered gradients $\widetilde{G} = G - \mathbb{E}[G]$ reveals these correlations:

$$\Sigma = \frac{1}{N-1}\widetilde{G}^T\widetilde{G}. \tag{2}$$

High off-diagonal values in Σ indicate strong linear dependencies between token gradients (as visualized in Fig. 2a and 2b), which confines the adversarial update to a narrow, model-specific subspace.

To break this structural dependency, we propose Gradient Correlation Whitening (GCW). The core operation is a whitening transformation that maps the correlated gradients $\widetilde{G}$ to a new set G_{white} whose covariance is the identity matrix:

$$G_{\text{white}} = \widetilde{G}\, W, \quad \text{where } W = \Sigma^{-\frac{1}{2}}. \tag{3}$$

Here, W is the whitening matrix, computed via the inverse square root of the covariance matrix. This transformation orthogonalizes the gradient directions, effectively dispersing the update across a broader set of token features and reducing over-reliance on any specific subset.

Direct computation of $\Sigma^{-\frac{1}{2}}$ via full eigen-decomposition is expensive. Instead, we adopt an efficient adaptive strategy using Newton's iteration to compute the inverse square root of the covariance matrix. This ensures computational efficiency for typical ViT configurations (e.g., $N = 196, D = 768$). As visualized in Fig. 2c and 2d, after our GCW operation, the token gradients achieve global decorrelation during back-propagation. To maintain stability and retain useful gradient information, we blend the whitened gradients with the original gradients using a controllable mixing coefficient:

$$G_{\text{modified}} = (1-\sigma)G + \sigma \cdot G_{\text{white}}, \tag{4}$$

where $\sigma \in [0, 1]$ controls the overall strength of the whitening effect. This soft modification ensures a balance between decorrelation and exploitation of the original attack direction.

C. Wavelet Packet Decomposition for Frequency-Aware Attack

Existing frequency-domain attacks rely on the DCT, which provides only a global, fixed spectral representation. This limits their ability to adapt to the varied sensitivity that different models exhibit towards localized frequency components (e.g., high-frequency edges or textures). To overcome this, we replace DCT with Wavelet Packet Decomposition (WPD) [19], which offers a multi-resolution, localized frequency analysis.

Given the adversarial image x_t^{adv} at iteration t, we apply a two-level WPD using a selected wavelet basis (e.g., Daubechies [20]). In the first level, the image is decomposed into four sub-bands: a (approximation, capturing low-frequency coarse structures), h (horizontal details, containing high-frequency variations along the horizontal direction), v (vertical details, containing high-frequency variations along the vertical direction), and d (diagonal details, capturing high-frequency variations along diagonal orientations). The approximation sub-band a, which retains the most global structural information, is then further decomposed in the second level into its own set of four sub-bands: aa, ah, av, and ad. This yields a total of seven sub-bands for analysis: $C = aa, ah, av, ad, h, v, d$. This hierarchical representation captures a rich spectrum of frequency information, ranging from broad structural outlines to localized texture details, enabling fine-grained and adaptive perturbation in the frequency domain.

For each sub-band $C_i \in C$, we compute its importance for the current attack by evaluating the gradient of the loss with respect to that sub-band's coefficients. Let x' be the image reconstructed solely from C_i (setting other sub-bands to zero). The gradient mask M_i is computed as:

$$g_i = \nabla_{C_i}\mathcal{L}(f_s(x'), y), \quad M_i = \frac{|g_i|}{\max(|g_i|) + \xi}, \tag{5}$$

where ξ is a small constant for numerical stability. This yields a set of normalized masks $\{M_i\}_{i=1}^{7}$ that highlight the

frequency regions most influential to the surrogate model's decision.

In each inner iteration m of the attack, we create an enhanced version of the image to simulate diverse model responses. First, spatial Gaussian noise $\eta_s \sim \mathcal{N}(0,1)$ is added to x_t^{adv}. The noisy image is then decomposed via WPD to obtain sub-bands $\{C_i^{(m)}\}$. Each sub-band is perturbed by multiplicative frequency noise and scaled by its corresponding mask:

$$\widehat{C}_i^{(m)} = \left(C_i^{(m)} \odot \mathcal{N}(1, \lambda_f)\right) \odot M_i, \tag{6}$$

where $\odot$ denotes element-wise multiplication, and $\mathcal{N}(1, \lambda_f)$ is Gaussian noise with mean 1 and variance λ_f^2, simulating varying frequency attention across models. The enhanced sub-bands $\{\widehat{C}_i^{(m)}\}$ are then reconstructed via the inverse WPD to form an augmented image $\widehat{x}_m$.

For M such augmentations, we compute the gradients of the loss with respect to each $\widehat{x}_m$ using the GCW-enhanced surrogate model f_{GCW}. The final gradient for the outer iteration is the average:

$$g = \frac{1}{M} \sum_{m=0}^{M-1} \nabla_{\widehat{x}_m} \mathcal{L}(f_{\text{GCW}}(\widehat{x}_m), y). \tag{7}$$

This aggregated gradient is then used in a standard iterative update to obtain x_{t+1}^{adv}, followed by projection onto the ℓ_∞-ball of radius ϵ. The synergy between GCW and WPD is crucial: GCW provides a robust, decorrelated gradient foundation, while WPD ensures the perturbations are effectively targeted at the most vulnerable, multi-resolution frequency components across potential target models. Our algorithm steps are shown in Algorithm 1.

Algorithm 1 GCW-WPD Method

Require: Clean image x, label y, perturbation budget ϵ, outer steps T, inner steps M, surrogate model f_s, model f_{GCW}, loss $\mathcal{L}$, parameter λ_f for Gaussian noise $\mathcal{N}(1, \lambda_f)$
Ensure: Adversarial example x^{adv}

$x_0^{adv} \leftarrow x$; $\eta \leftarrow \epsilon / T$
for $t = 0$ to $T-1$ **do**
  $\mathcal{C} \leftarrow \text{WPD}(x_t^{adv})$ {7 sub-bands: $\mathcal{C} = \{C_i\}_{i=1}^7$}
  for $i = 1$ to 7 **do**
    $g_i \leftarrow \nabla_{C_i} \mathcal{L}(f_s(x'), y)$ by Eq.(5)
    $M_i \leftarrow |g_i| / (\max(|g_i|) + \xi)$ by Eq.(5)
  end for
  for $m = 0$ to $M-1$ **do**
    $C_i^{(m)} \leftarrow \text{WPD}(x_t^{adv} + \mathcal{N}(0,1))$
    for $i = 1$ to 7 **do**
      $\widehat{C}_i^{(m)} \leftarrow (C_i^{(m)} \odot \mathcal{N}(1, \lambda_f)) \odot M_i$ by Eq.(6)
    end for
    $\hat{x}_m \leftarrow \text{InverseWPD}(\{\widehat{C}_i^{(m)}\}_{i=1}^7)$
    $g_m \leftarrow \nabla_{\hat{x}_m} \mathcal{L}(f_{\text{GCW}}(\hat{x}_m), y)$ by Eq.(7)
  end for
  $x_{t+1}^{adv} \leftarrow \text{Clip}_{x,\epsilon}\{x_t^{adv} + \eta \cdot \text{sign}(\frac{1}{M} \sum_{m=0}^{M-1} g_m)\}$
end for
return x^{adv}

IV. Experiments

A. Experimental Setup

Dataset. We evaluate our method on the ImageNet dataset [21], which is widely used for evaluating adversarial attacks. Following standard protocols in transferable adversarial attack research, we randomly select 1,000 correctly classified images from the validation set, covering 1,000 different categories, to serve as our test set.

Baseline Methods. To assess the effectiveness of our proposed GCW-WPD method, we compare it against several state-of-the-art transferable attack methods. These include: MI-FGSM [16], PNA [11], TGR [12], VDC [18], TGD-MFA [14]. For all baselines, we use the official implementations and follow the recommended hyper-parameter settings as described in the original papers.

Models. To thoroughly evaluate the transferability, we employ a diverse set of both surrogate models and target models. For surrogate models, we employ three ViTs: ViT-Base (ViT-B), CaiT-Small (CaiT-S), and PiT-Tiny (PiT-T). For target models, we consider nine models, including five ViTs and four CNNs: ViT-B, CaiT-S, PiT-Base (PiT-B), Visformer-Small (Vis-S), Swin-Base (Swin-B), ResNet-50 (RN50), VGG-16, MobileNetV2 (MN-v2), and Inception-v3 (Inc-v3).

Evaluation Metrics. We use the Attack Success Rate (ASR) [14] as the primary evaluation metric, which is defined as the percentage of adversarial examples that successfully cause the target model to make an incorrect prediction. We assess statistical significance using Wilcoxon signed-rank tests with Bonferroni correction (significance level $p = 0.01$), with significant results marked by asterisks (*) in the tables.

Implementation Details. For a fair comparison, all attacks are conducted under the same setting [18]: the maximum ℓ_∞ perturbation budget is set to $\epsilon = 16/255$, the number of iterations is $T = 10$, and the step size is $\alpha = \epsilon/T$. For our method, the whitening strength σ in GCW is set to 0.12. In the WPD module, we use the Daubechies-4 wavelet and perform a two-level decomposition, resulting in seven sub-bands as described in Section III-C. The number of inner frequency augmentations M is set to 20.

B. Performance Comparison

To evaluate the effectiveness of our proposed GCW-WPD framework, we conduct experiments on three surrogate models and nine target models. As shown in Table I, GCW-WPD consistently achieves the highest ASR across all settings, demonstrating superior transferability and robustness.

When using ViT-B as the surrogate, GCW-WPD attains an average black-box ASR of 82.4%, outperforming the strongest baseline TGD-MFA by 2.1%. Notably, our method exhibits remarkable cross-architecture generalization: it surpasses TGD-MFA by 5.3% on RN50 and by 2.5% on Inc-v3. This indicates its capability to learn model-agnostic adversarial patterns rather than overfitting to ViT-specific attention mechanisms. With CaiT-S as the surrogate, GCW-

TABLE I: Attack success rates (%) for transfer attacks on various ViTs and CNNs. The average (Avg.) column excludes the white-box results (grey background). An asterisk (*) indicates GCW-WPD's statistically significant superiority over the corresponding baseline under Wilcoxon signed-rank tests ($p = 0.01$, Bonferroni-corrected). Best results are in bold.

Model	Method	ViT-B	CaiT-S	PiT-B	Vis-S	Swin-B	RN50	VGG-16	MN-v2	Inc-v3	Avg.
ViT-B	MI-FGSM	97.7	51.4	42.1	44.9	51.8	39.5	56.9	53.3	45.7	48.2
	PNA	93.2	60.1	52.5	54.7	56.6	41.7	61.8	54.3	43.7	53.2
	TGR	99.7	82.0	61.7	69.8	75.2	59.1	80.3	76.8	62.8	71.0
	VDC	99.8	84.7	64.9	65.5	68.2	51.6	71.4	65.5	54.7	65.8
	TGD-MFA	99.9	95.2	78.9	81.0	79.5	69.4	83.1	79.9	75.0	80.3
	GCW-WPD	99.8	**95.8***	**81.2***	**81.8***	**80.4***	**74.7***	**84.6***	**82.8***	**77.5***	**82.4**
CaiT-S	MI-FGSM	70.5	98.3	53.9	49.0	58.7	46.3	66.5	58.1	49.3	56.5
	PNA	72.0	91.7	58.9	62.2	68.6	55.1	68.3	62.7	57.6	63.2
	TGR	87.6	99.9	69.0	76.2	85.1	67.9	84.4	79.9	60.4	76.3
	VDC	88.2	99.7	74.1	79.9	88.7	73.4	82.2	74.6	59.7	77.6
	TGD-MFA	95.9	99.8	88.4	92.3	90.8	77.0	88.1	83.5	86.1	87.7
	GCW-WPD	**96.5***	99.8	**90.1***	**93.1***	**93.9***	**79.3***	**88.7***	**84.8***	**86.6***	**89.2**
PiT-T	MI-FGSM	30.9	37.5	38.7	44.7	50.1	39.9	65.4	67.9	49.7	47.2
	PNA	40.3	51.0	49.9	56.4	58.1	48.3	73.2	80.5	62.6	57.8
	TGR	39.0	49.4	47.6	55.9	58.2	52.7	81.6	85.3	65.7	59.5
	VDC	42.5	53.2	51.0	57.8	59.8	56.2	82.9	88.4	69.3	62.3
	TGD-MFA	46.5	53.7	54.5	60.9	62.4	60.0	82.1	89.3	71.8	64.6
	GCW-WPD	**47.5***	**55.9***	**54.9***	**62.6***	**63.8***	**61.7***	**84.5***	**90.1***	**73.0***	**66.0**

WPD maintains its leading performance, achieving an average ASR of 89.2% compared to TGD-MFA's 87.7%. The gains are especially pronounced on challenging ViT targets such as Swin-B, with an improvement of 3.1%, and PiT-B, with an improvement of 1.7%. This shows that our gradient whitening strategy effectively mitigates architectural biases that hinder transferability within the ViT family. In the most challenging scenario, pooling-based PiT-T serves as the surrogate, GCW-WPD still consistently surpasses all baselines, achieving an average ASR of 66.0% compared to 64.6% for TGD-MFA. Compared to gradient regularization methods such as PNA, TGR, and VDC, GCW-WPD demonstrates substantial and consistent gains, particularly in cross-architecture transfers.

Overall, GCW-WPD establishes a new state-of-the-art in transferable adversarial attacks, showing remarkable consistency by achieving the highest average ASR for every surrogate-target combination. The synergistic integration of gradient correlation whitening and wavelet packet decomposition proves to be a generalizable solution for enhancing adversarial transferability across both ViT and CNN architectures.

C. Ablation Studies

To validate the contribution of each component in our GCW-WPD framework, we conduct ablation studies across both ViT and CNN targets. Results are summarized in Table II and Table III, comparing four configurations: attack without GCW or WPD, GCW only, WPD only, and the full GCW-WPD.

The results in Tables II and III consistently demonstrate the superiority of the full GCW-WPD over its individual components across both ViT and CNN architectures. When using ViT-B as the surrogate, the full method achieves an average ASR of 86.3%, significantly outperforming the GCW-only (79.8%) and WPD-only (66.7%) variants, indicating clear synergistic effects. This trend holds for other surrogates like CaiT-S and PiT-T, where the combined approach yields the best performance. On CNN targets, GCW-WPD also shows substantial gains, especially in cross-architecture transfer—for instance, with ViT-B as the surrogate, the full method attains an average ASR of 79.9% on CNNs, compared to 68.9% for GCW-only and 59.2% for WPD-only. The results validate our design: GCW mitigates gradient-level overfitting, while WPD provides complementary, frequency-aware input diversification. Their integration is particularly beneficial when the surrogate model has limited capacity, as WPD's multi-resolution analysis becomes more critical.

TABLE II: Ablation studies on the components of GCW-WPD across different ViT-based models.

Model	Method	ViT-B	CaiT-S	PiT-B	Vis-S	Avg.
ViT-B	None	99.9	70.1	39.1	40.6	49.9
	GCW-Only	99.9	88.2	72.8	78.5	79.8
	WPD-Only	99.5	76.9	58.9	64.3	66.7
	GCW-WPD	99.8	**95.8**	**81.2**	**81.8**	**86.3**
CaiT-S	None	77.1	99.9	50.3	57.5	61.6
	GCW-Only	92.4	99.8	78.0	84.1	84.8
	WPD-Only	83.5	99.4	73.2	76.7	77.8
	GCW-WPD	**96.5**	99.8	**90.1**	**93.1**	**93.2**
PiT-T	None	33.5	38.7	36.9	37.4	36.6
	GCW-Only	43.9	52.9	47.3	54.7	49.7
	WPD-Only	41.2	51.0	45.2	49.9	46.8
	GCW-WPD	**47.5**	**55.9**	**54.9**	**62.6**	**55.2**

TABLE III: Ablation studies on the components of GCW-WPD across different CNN-based models.

Model	Method	RN50	VGG-16	MN-v2	Inc-v3	Avg.
ViT-B	None	36.4	43.1	39.5	38.0	39.3
	GCW-Only	67.2	70.6	66.8	70.9	68.9
	WPD-Only	55.8	59.8	58.5	62.6	59.2
	GCW-WPD	**74.7**	**84.6**	**82.8**	**77.5**	**79.9**
CaiT-S	None	46.2	48.9	45.1	46.0	46.6
	GCW-Only	70.5	73.6	69.2	73.7	71.8
	WPD-Only	58.7	60.4	59.9	66.5	61.4
	GCW-WPD	**79.3**	**88.7**	**84.8**	**86.6**	**84.9**
PiT-T	None	37.5	43.7	46.5	35.8	40.9
	GCW-Only	55.9	64.9	71.1	58.8	62.7
	WPD-Only	50.2	55.0	60.7	52.7	54.7
	GCW-WPD	**61.7**	**84.5**	**90.1**	**73.0**	**77.3**

D. Parameter Sensitivity Analysis

To further investigate the influence of the whitening parameters in GCW, we conduct a sensitivity analysis on the mixing coefficient σ.

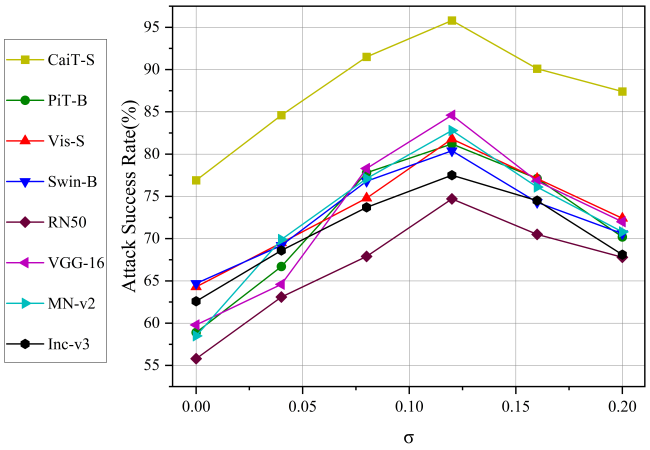


Fig. 3: Impact of parameter σ on attack success rates (%) across different models, using ViT-B as the surrogate.

As shown in Fig. 3, when $\sigma = 0$, the method reduces to the baseline without whitening, yielding the lowest ASR. Performance peaks at $\sigma = 0.12$, indicating that moderate decorrelation enhances transferability. However, larger values degrade performance, as excessive whitening disrupts gradient alignment or introduces noise, harming optimization. This confirms the necessity of an appropriate σ in the proposed whitening mechanism. In summary, the selected value $\sigma = 0.12$ is demonstrated to be within the optimal value for generating highly transferable adversarial examples.

V. Conclusion

In this paper, we introduced GCW-WPD, a novel method for enhancing the cross-architecture transferability of adversarial attacks. GCW reduces structural dependency on the source model by decorrelating token-wise gradients via a global whitening transformation. Complementarily, WPD enables fine-grained frequency analysis, allowing adaptive perturbation enhancement across diverse spectral bands. Experiments on a wide range of ViT and CNN models demonstrate that GCW-WPD consistently outperforms state-of-the-art transfer-based attack methods, achieving significant improvements in adversarial transferability.

References

[1] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie, "A convnet for the 2020s," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11976–11986.

[2] Ziyu Wang, Wenhao Li, and Ji Wu, "Dq3d: Depth-guided query for transformer-based 3d object detection in traffic scenarios," in *2025 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2025, pp. 1–8.

[3] Weili Jiang, Yihao Li, Zaiyi Liu, Lin An, Gwenolé Quellec, and Chubin Ou, "Saradapter: Prioritizing attention on semantic-aware representative tokens for enhanced medical image segmentation," *IEEE Transactions on Medical Imaging*, 2025.

[4] 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*, 2021, pp. 7838–7847.

[5] Xu Guo, Peng Chen, Zhihui Lu, Hongfeng Chai, Xin Du, and Xudong Wu, "Towards transferable adversarial attacks on vision transformers for image classification," *Journal of Systems Architecture*, vol. 152, pp. 103155, 2024.

[6] Meixi Zheng, Xuanchen Yan, Zihao Zhu, Hongrui Chen, and Baoyuan Wu, "Blackboxbench: A comprehensive benchmark of black-box adversarial attacks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[7] Sensen Guo, Baocang Wang, Xiaoyu Li, Peican Zhu, Lianwei Wu, Wenning Wu, and Lin Jia, "Alda: Enhancing the transferability of adversarial attacks with attention-guided look-ahead and data augmentation," *Pattern Recognition*, p. 112686, 2025.

[8] Xiyao Liu, Junxing Ma, Xinda Wang, Qianyu Lin, Jian Zhang, Gerald Schaefer, Cagatay Turkay, and Hui Fang, "Recoverable facial identity protection via adaptive makeup transfer adversarial attacks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025, vol. 39, pp. 514–522.

[9] Zhiyu Zhu, Xinyi Wang, Zhibo Jin, Jiayu Zhang, and Huaming Chen, "Enhancing transferable adversarial attacks on vision transformers through gradient normalization scaling and high-frequency adaptation," in *The Twelfth International Conference on Learning Representations*, 2024.

[10] Yajie Wang, Chuan Zhang, Huipeng Zhou, Zuobin Ying, Zehui Xiong, Wanlei Zhou, and Liehuang Zhu, "Boosting adversarial transferability of vision transformers," *IEEE Transactions on Dependable and Secure Computing*, 2025.

[11] 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 AAAI Conference on Artificial Intelligence*, 2022, vol. 36, pp. 2668–2676.

[12] Jianping Zhang, Yizhan Huang, Weibin Wu, and Michael R Lyu, "Transferable adversarial attacks on vision transformers with token gradient regularization," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 16415–16424.

[13] Yajie Wang, Yi Wu, Shangbo Wu, Ximeng Liu, Wanlei Zhou, Liehuang Zhu, and Chuan Zhang, "Boosting the transferability of adversarial attacks with frequency-aware perturbation," *IEEE Transactions on Information Forensics and Security*, vol. 19, pp. 6293–6304, 2024.

[14] Cong Hu, Zhi-chao He, Yuan-bo Li, and Xiao-Jun Wu, "Robust token gradient and frequency-aware transferable adversarial attacks on vision transformers," *IEEE Transactions on Information Forensics and Security*, 2025.

[15] 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.

[16] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li, "Boosting adversarial attacks with momentum," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 9185–9193.

[17] Dongxian Wu, Yisen Wang, Shu-Tao Xia, James Bailey, and Xingjun Ma, "Skip connections matter: On the transferability of adversarial examples generated with resnets," in *International Conference on Learning Representations*, 2020.

[18] Jianping Zhang, Yizhan Huang, Zhuoer Xu, Weibin Wu, and Michael R Lyu, "Improving the adversarial transferability of vision transformers with virtual dense connection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, vol. 38, pp. 7133–7141.

[19] Zhengyao Song, Yongqiang Li, Danni Yuan, Li Liu, Shaokui Wei, and Baoyuan Wu, "Wpda: frequency-based backdoor attack with wavelet packet decomposition," *Neural Networks*, vol. 194, pp. 108074, 2026.

[20] Cédric Vonesch, Thierry Blu, and Michael Unser, "Generalized daubechies wavelet families," *IEEE transactions on signal processing*, vol. 55, no. 9, pp. 4415–4429, 2007.

[21] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al., "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.