Transferable Adversarial Attack for Both Vision Transformers and Convolutional Networks via Momentum Integrated Gradients

Wenshuo Ma[1], Yidong Li[2], Xiaofeng Jia[3], Wei Xu[1]
[1]IIIS, Tsinghua University, [2]Beijing Jiaotong University, [3]Beijing Big Data Centre
mwenshuo@gmail.com, ydli@bjtu.edu.cn, jiaxf@jxj.beijing.gov.cn, weixu@tsinghua.edu.cn

Abstract

Visual Transformers (ViTs) and Convolutional Neural Networks (CNNs) are the two primary backbone structures extensively used in various vision tasks. Generating transferable adversarial examples for ViTs is difficult due to ViTs' superior robustness, while transferring adversarial examples across ViTs and CNNs is even harder, since their structures and mechanisms for processing images are fundamentally distinct. In this work, we propose a novel attack method named Momentum Integrated Gradients (MIG), which not only attacks ViTs with high success rate, but also exhibits impressive transferability across ViTs and CNNs. Specifically, we use integrated gradients rather than gradients to steer the generation of adversarial perturbations, inspired by the observation that integrated gradients of images demonstrate higher similarity across models in comparison to regular gradients. Then we acquire the accumulated gradients by combining the integrated gradients from previous iterations with the current ones in a momentum manner and use their sign to modify the perturbations iteratively. We conduct extensive experiments to demonstrate that adversarial examples obtained using MIG show stronger transferability, resulting in significant improvements over state-of-the-art methods for both CNN and ViT models.

1. Introduction

Vision Transformers (ViTs) have become increasingly popular and are widely adopted [1, 4, 5, 11], following the success of Convolutional Neural Networks (CNNs) [16, 25, 33, 34], due to their impressive performance. ViTs process images as sequences of patches and use self-attention to model global dependencies among them, while CNNs exploit spatial structures of images using convolutional filters to capture local features like edges and textures. Regardless of these different image processing mechanisms used in CNNs and ViTs, their widespread applications in visual domains emphasize the importance of developing adversarial examples that can attack both types of models, which helps identify models' vulnerabilities before their deployment in real-world scenarios.

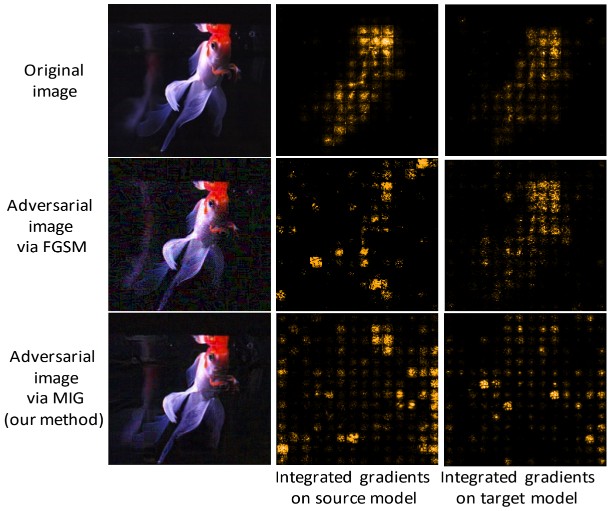
Original image
Adversarial image via FGSM
Adversarial image via MIG (our method)
Integrated gradients on source model
Integrated gradients on target model

Figure 1. Integrated gradients of original image (top row), adversarial images generated with FGSM [14] (middle row) and our Momentum Integrated Gradients (MIG) method (bottom row). Adversarial examples are crafted on source model DeiT-S [43] and used to attack target model ViT-B [11]. Compared to FGSM, adversarial image generated using MIG exhibits more consistent integrated gradients on both source and target models, indicating that MIG effectively interferes with both models, leading them to focus on unimportant background regions rather than the main objects.

In adversarial attacks, a malicious attacker perturbs an input image by applying a small, human-imperceptible perturbation to break the prediction of a machine learning model [6, 12, 14, 49]. Depending on whether the attacker has full access to the target victim model, adversarial attack methods can be classified into two types, *white-box attacks* and *black-box attacks*. Considering that the target models are usually not so readily accessible, generating a transferable adversarial example from a white-box surrogate model and then transferring it to the target black-box model is a common strategy for attacking black-box models. Therefore, the transferability of adversarial perturbations plays a critical role in adversarial attacks.

However, due to the differences between ViTs and CNNs mentioned earlier, previous attack methods designed for CNNs show very limited transferability when applied to ViTs. This leaves obstacles for attacking black-box ViTs using transfer-based attack schemes, especially because ViTs are shown to be more robust than CNNs [3, 29] and few adversarial attacks against ViTs have been studied.

In this work, we propose a novel attack method based on Integrated Gradients and Momentum iterative strategy, called Momentum Integrated Gradients (MIG) attack. MIG can not only successfully attack both ViTs and CNNs, but also show better transferability across models. That is, adversarial perturbations generated on the white-box model are still likely to cause the black-box target model make mistakes, regardless of its architecture.

Specifically, we first employ *Integrated Gradients (IG)* [39] method to compute the *saliency score* [19] of model prediction with respect to input. The sign of integrated gradients is then used to guide the updating of the perturbations. As an attribution method, integrated gradients reflect the sensitivity of model's outputs to inputs at different locations. Intuitively, there should be more pronounced perturbations at locations that have greater impacts on the model to have a sufficient impact on output. Besides, IG satisfies an excellent property *implementation invariance*, which further improves the transferability of adversarial examples.

Furthermore, we update perturbations with momentum-based iterative strategy to obtain better attack performance as in [9]. This strategy is similar to momentum gradient descent algorithm [31], where the gradient of the current iteration is accumulated by adding the velocity vector obtained from previous iterations. Applying momentum in MIG helps accumulate the impact of historical gradients, speeding up perturbation updating and making the loss function increase faster. Additionally, it helps to iteratively jump out of poor local optima caused by model-specific noise and find the globally optimal perturbations. As a result, these perturbations transfer better across different models.

As an example, we visualize integrated gradients for a clean image and adversarial images generated using FGSM [14] and MIG in Figure 1. Adversarial image using MIG exhibits similar integrated gradients on both the source model (DeiT-S [43]) and the target model (ViT-B [11]). This observation indicates that MIG effectively misleads both models to prioritize irrelevant regions instead of the main object.

We conduct extensive experiments on the ImageNet *val* dataset [35]. Experiment results demonstrate that MIG significantly enhances the transferability of adversarial examples across ViTs and CNNs. Compared to classical attacks such as FGSM [14], PGD [26], advanced attack MI [9], and even transfer-based attack methods specifically designed for ViTs [28], MIG achieves state-of-the-art attack success rate under a small perturbation budget.

We briefly summarize our main contributions as follows.

- We propose Momentum Integrated Gradients (MIG) attack method based on Integrated Gradients and Momentum updating strategy, which can attack ViTs with high transfer attack success rate.
- The transferability of adversarial perturbations generated using MIG outperforms SOTA attack methods when transferring across ViT and CNN models, which has been overlooked in prior research.
- We empirical demonstrate that attribution-based transfer attacks are valid for ViTs. This suggests an intrinsic connection between model interpretability and model robustness against attacks.

2. Related Work

2.1. Robustness of Vision Transformer

There exist some studies demonstrating the superior adversarial robustness of ViTs over CNNs [3, 29, 30]. Additionally, [36] shows that features learned by ViTs contain less high-frequency patterns and are less sensitive to high-frequency perturbations. [50] examines the role of self-attention in learning robust representations and verifies it as a contributor of the improved robustness.

As for transfer-based attacks towards ViTs, [27] finds that model ensemble can improve robustness without sacrificing clean accuracy against a black-box adversary. [45] and [27] show that existing transfer-based attacks struggle to effectively transfer adversarial examples from CNNs to ViTs, which can be attributed to the excellent robustness of ViTs. Our experiments also support these findings.

2.2. Adversarial Attack

Adversarial attack methods can generally be classified into white-box attacks [2, 13, 41] and black-box attacks [7, 18, 48], according to the amount of information available to the adversary about the target model. In white-box attacks, the malicious party can fully access to victim models and construct adversarial examples using loss and gradients of the victim models, like one-step fast gradient sign method FGSM [14] and iterative gradient-based methods [26, 41].

Compared to white-box attacks, black-box attacks are more challenging as they only have access to models' outputs through queries. Some black-box methods utilize feedback from these queries to guide the generation of adversarial examples, called query-based attacks [7, 8, 37]. Other approaches realize black-box attacks based on transferability of adversarial examples. For example, MI [9] enhances transferability by incorporating a momentum term

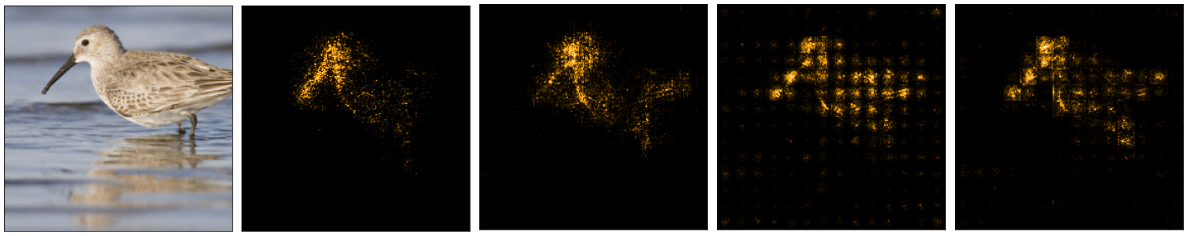

Figure 2. An original image and the corresponding integrated gradients from ResNet-50 [16], Inception-v4 [40], DeiT-T [43] and ViT-T [11], from left to right. Different models have similar integrated gradients distributions for the same image. Compared with CNN model, ViTs process images as sequences of patches, so the distribution of integrated gradients is also patch-like. To provide a clearer illustration, we smoothed the integrated gradients by adding random Gaussian noise to the original image five times and took their average.

and model logits ensemble, while DIM [47] applies random resizing to inputs to address overfitting on white-box models. TIM [10] adopts a set of images to calculate gradients, SIM [22] introduces the scale-invariant property to enhance the transferability, and [18] combines several existing methods and uses integrated gradients to generate perturbations. These methods focus exclusively on attacks against CNNs, and all of them perform poorly on ViTs, even using ViTs as source models. Recently, [28] utilizes self-ensemble and token refinement to improve attack transferability for ViTs. Different from these methods, our MIG method achieves excellent transferability for both CNNs and ViTs, without requiring any ensemble of models or inputs.

3. Methodology

In this section, we first introduce the motivation of using integrated gradients and momentum iterative method, and their preliminaries. Then we give a detailed description of Momentum Integrated Gradients (MIG). We further demonstrate that MIG can improve transferability across different architectures (ViTs and CNNs), simply by appropriately ensembling only two models.

3.1. Capture Model-agnostic Critical Regions via Integrated Gradients

Integrated gradients (IG) [39] is an axiomatic interpretability method that attributes the prediction of a neural network to its inputs. It calculates integrated gradients for each pixel by approximating the integral of gradients along the given path from a baseline image to the input image. Intuitively, these obtained gradients can be considered as the *importance* or *saliency scores* [18, 19] for all pixels.

We describe the process of calculating integrated gradients first. Let $f : \mathbb{R}^n \rightarrow [0, 1]$ represent a deep network which classifies an input image x into a certain class with probabilities. Suppose we have a baseline image b which contains no valid information for this network. In practice, the baseline image is usually a black image. Consider the straight-line path in $\mathbb{R}^n$ from baseline b to input x, and compute gradients at all points along the path. Integrated gradient along the i-th dimension for the input x and baseline b is obtained by cumulating these gradients together:

$$IG_i(f, x, b) = (x_i - b_i) \times \int_{\xi=0}^{1} \frac{\partial f(b + \xi \times (x - b))}{\partial x_i} d\xi, \tag{1}$$

where $\frac{\partial f(x)}{\partial x_i}$ is the gradient of $f(x)$ along the i-th dimension. For simplicity, we use $IG(f, x, b)$ to represent $[IG_0(f, x, b), IG_1(f, x, b), \ldots, IG_{n-1}(f, x, b)]$, where n is the total number of image pixels.

Compared to direct gradients and raw attentions, integrated gradients satisfy a good property: *Implementation Invariance*, which is the key factor to improve the transferability of adversarial perturbations between different models. Intuitively, if two networks produce identical outputs for all inputs, even though they have distinct implementations, they are functionally equivalent. Implementation invariance implies that integrated gradients will be consistent between such networks because they only depend on the input and output of the model and are not influenced by implementation details, as shown in Equation 1. Therefore, when IG is used to generate adversarial perturbations, two models that are functionally equivalent should have the same outputs. Then the adversarial perturbation that is valid on the source model should still be valid when transferred to the target model. In practice, it's difficult to obtain fully equivalent models, but this property can qualitatively help explain the enhancement of adversarial transferability brought by IG, and empirical ablation experiments in Section 4.6 further confirm its utility.

Figure 2 shows an example of integrated gradients of a clean image across different models. Various models have different integrated gradients for the same image, but all of them highlight the salient regions and downplay the trivial background areas. This observation inspires us to leverage IG as a guiding signal for generating adversarial examples.

3.2. Speed Up Perturbation Update via Momentum

Since visual models typically have a large number of parameters, and the calculation of integrated gradients is also intricate, we leverage momentum iterative strategy to accelerate perturbation updating and prevent the algorithm from being trapped in suboptimal regions due to model-specific noise, similar to momentum gradient descent algorithms [31, 32] and MI [9]. Traditional gradient-based attacks, such as [14, 21], generate adversarial examples using only the current gradient and not considering past steps. In contrast, the momentum method [31] accelerates the gradient descent algorithm by accumulating the velocity vector in the direction of the gradients from past iterations.

We apply this momentum strategy to our integrated-gradients-based attack to produce stable adversarial examples faster and with improved transferability across various models. Specifically, the accumulated integrated gradients g_t in the t-th iteration can be calculated as:

$$g_t = \mu * g_{t-1} + \frac{\Delta_t}{\|\Delta_t\|_1}, \tag{2}$$

where g_{t-1} is the accumulated integrated gradients of the previous iteration, Δ_t is the integrated gradients of the current iteration, and we normalize it by L_1 norm before addition. We use a momentum factor, i.e., the decay factor μ to regulate the decay rate of historical gradients. A lower μ discounts older accumulated gradients faster.

3.3. Momentum Integrated Gradients Method

As described above, by introducing integrated gradients, our method perturbs the most critical regions of the model's prediction in the model-agnostic attribution space. By incorporating momentum, previous gradients' energy is accumulated, and the speed can be maintained when entering flat regions, preventing stagnation in poor local optima and resulting in stable and optimal perturbations. We describe the complete procedure of our proposed Momentum Integrated Gradients (MIG) approach in Algorithm 1.

Note that although we only present the un-targeted attack, we can easily modify it to be a targeted version. Starting with a clean input image x and its corresponding ground-truth label y, the un-targeted adversarial example x_{adv} can be expressed as follows:

$$f(x_{adv}) \neq y, \quad \text{s.t.}\ \|x - x_{adv}\|_\infty \leq \epsilon, \tag{3}$$

where ϵ is the perturbation budget of the adversarial attack.

Concretely, for an input image x, we use T iterations to generate the adversarial example x_{adv} with a total perturbation budget ϵ. In the t-th iteration, we first calculate the integrated gradients of the model output with respect to the current input x_{t-1}. In practice, we approximate IG efficiently by summing up the gradients of the points along the straight-line path from the baseline image b to the input image x, which are spread sufficiently close to each other.

Algorithm 1: Momentum Integrated Gradients Attack.

```
Input: The source white-box model $f$, original clean
       image $x$ and its label $y$.
Parameter: The perturbation budget $\epsilon$, iteration
           number $T$, momentum factor $\mu$ and
           baseline image $b$.
Output: Adversarial image $x_{adv}$.
Initial accumulated integrated gradients $g_0 = 0$;
$x_0 = x$, $\alpha = \frac{\epsilon}{T}$;
for $t = 1$ to $T$, do
    // Calculate integrated gradients for $x_{t-1}$:
    $\Delta_t$ = IG($f$, $x_{t-1}$, $b$);
    // Update $g_t$ via momentum iterative method:
    $g_t$ = $\mu * g_{t-1}$ + $\frac{\Delta_t}{\|\Delta_t\|_1}$
    // Update $x_t$ according to the sign of $g_t$:
    $x_t = Clip_\epsilon\{x_{t-1} + \alpha \cdot sign(g_t)\}$
end
$x_{adv} = x_T$;
Return $x_{adv}$
```

Specifically, using the summation, we approximate IG as:

$$IG_i(f, x, b) \approx (x_i - b_i) \times \sum_{k=1}^{s} \frac{\partial f(b + \frac{k}{s} \times (x - b))}{\partial x_i} \times \frac{1}{s}, \tag{4}$$

where s is the order of approximation, i.e., the order of Taylor expansion. A larger s implies a more accurate approximation, while the computational cost increases. We set s to 20 to achieve a balance between accuracy and efficiency.

We then update the accumulated gradients g_t by gathering the gradients from previous iterations with the current integrated gradients in the momentum iterative manner. The adversarial image x_t then moves along the direction of the sign of g_t, with step size $\alpha = \epsilon/T$. After T iterations, the final adversarial image x_{adv} is generated.

3.4. MIG with Model Ensemble

Following Algorithm 1, MIG generates transferable perturbations on a white-box model, regardless of whether it is a CNN or a ViT. Although using a single model already shows good transfer attack ability, perturbations crafted solely on a source model may be limited in its specific architecture, given the different ways in which CNNs and ViTs process images. So we also support ensembling MIG with two separate source models to overcome this limitation.

We investigate three ensemble strategies: perturbation ensemble, logit ensemble, and integrated gradients (IG) ensemble. In perturbation ensemble, we generate perturbations separately using two source models and combine them to obtain the final perturbation. In logit ensemble, we directly fuse the logits of two models like [9]. In IG ensemble, we compute the integrated gradients of two models and fuse them to form the new IG for updating the perturbation. The ensemble process can be summarized as:

$$z(x) = \sum\nolimits_{m=1}^{M} w_m z_m(x), \tag{5}$$

where $z_m(x)$ can be the perturbation, logit or IG of the m-th model, w_m is the ensemble weight with $w_m \geq 0$ and $\sum_{m=1}^{M} w_m = 1$, $z(x)$ is the final ensembled term, and M is the number of ensemble models. Specifically, we set $M = 2$ since ensembles of two models are effective enough.

While expanding MIG to a two-model ensemble is straightforward, it can further improve the overall transferability beyond single-model MIG, especially in scenarios where different model architectures are used. Subsequent experiments in Section 4.5 empirically confirm the effectiveness and convenience of this ensemble approach.

4. Experiments

4.1. Datasets and Evaluation

We evaluate the performance of MIG mainly on ImageNet dataset [35], following [9, 18, 28]. Specifically, we randomly sample 5,000 images from ImageNet *val* split with 5 images per class, all of which can be correctly classified by all white-box source models. That is, on the source models used to craft adversarial examples, these 5K images can achieve 100% classification accuracy.

After generating adversarial examples on selected images using MIG, we measure the transferability of them in terms of *attack success rate* (also known as *fool rate*), i.e., we count the number of images that have successfully fooled the target black-box model, and compute their ratio to the overall images. We also use *mean attack success rate (MASR)* to measure the overall performance of an attack method across multiple target models.

4.2. Experiment Settings

We evaluate the transferability of adversarial examples generated by different methods using CNNs and ViTs as source (surrogate) models and target (victim) models, respectively, following [28]. For source models, we study three vision transformers from DeiT family [43] (DeiT-T, DeiT-S, and DeiT-B) due to their efficiency, as well as three CNNs: VGG-19$_{bn}$ (VGG19) [38], MNAS [42] and Inception-v4 (Incep-v4) [40]. For target models, we use three CNNs including DenseNet-201 (DN201) [17], ResNet50-BiT-M (BiT) [20], and state-of-the-art ConvNeXt (CNeXt) [25], as well as five ViT models: ViT-S, ViT-B, ViT-L [11], Transformer iN Transformer (TNT) [15] and advanced Swin-Transformer (Swin) [24].

We compare MIG with widely-used and SOTA attacks, including classic gradient-based attacks such as FGSM [14] and PGD [26], as well as ensemble-based attack methods including MI [9], PGD-RE [28] and MI-RE [28].

As for implementation details, we set $\epsilon = 16/255$ as the perturbation budget, to make a fair comparison with previous works [9, 28]. We by default set iteration numbers $T = 25$ and momentum factor $\mu = 1$. In Section 4.7, we also evaluate the performance of MIG with other hyperparameter settings. We use pre-trained models provided in Timm library [46]. All experiments are conducted on four RTX 3090 GPUs.

4.3. Results of MIG with ViTs as Source Models

We first take DeiT family as source models to generate adversarial examples, and evaluate the attack success rate of these examples on different target models, including both CNNs and ViTs. Table 1 reports experiment results and demonstrates the effectiveness of MIG. Compared to traditional non-ensemble methods such as FGSM and PGD, MIG significantly improves the mean attack success rate by more than 20%. For example, when using DeiT-B as the source model, MIG outperforms FGSM by 47.77% on CNNs, and 52.55% on ViTs, and outperforms PGD by 55.49% on CNNs and 40.58% on ViTs.

Our method also exceeds state-of-the-art ensemble-based methods such as MI, PGD-RE and MI-RE, which use ensembles of logits from different models or class tokens from different ViT blocks. For example, when using DeiT-B as the source model, MIG achieves 30.12% better MASR than MI and 10.93% than MI-RE when transferring to CNNs, 21.92% better MASR than MI and 6.76% than MI-RE when transferring to ViTs.

Moreover, MIG shows consistent improvements when transferring to both CNN models and ViT models, while previous attacks may perform well on only one class of models but poorly on another.

4.4. Results of MIG with CNNs as Source Models

Given that CNNs are still the mainstream models for various vision applications and easily accessible, the transferability of adversarial examples crafted on CNNs is also worth investigating. In this section, we use VGG-19$_{bn}$, MNAS and Inception-v4 as source models and report the transfer attack success rate in Table 2. Note that both PGD-RE and MI-RE methods boost adversarial transferability through token refinement, which is only meaningful for ViTs, so we only compare MIG with the other three methods: FGSM, PGD and MI.

Results have illustrated that our method has a consistent performance improvement compared to previous methods. For example, when using Inception-v4 as the source model to generate adversarial examples, MIG increases the MASR on CNNs from 34.56% (using FGSM) to 87.05%, brings about 50% performance enhancement. For ViT mod-

Table 1. Attack success rate (%) and mean attack success rate (MASR,%) of MIG and other attacks, using DeiTs [43] as source models.

Source	Attack	Target Model								MASR	MASR	MASR
Model	Method	DN201	BiT	CNeXt	ViT-S	ViT-B	ViT-L	TNT	Swin	(CNNs)	(ViTs)	
DeiT-T	FGSM [14]	47.44	42.97	28.34	67.33	58.62	42.95	50.30	35.38	39.58	50.92	46.67
	PGD [26]	28.75	25.77	13.74	61.42	44.79	34.81	40.27	22.01	22.75	40.66	33.95
	PGD-RE[28]	40.56	33.46	27.02	96.34	71.03	58.60	55.52	29.38	33.68	62.17	51.48
	MI [9]	59.09	53.62	34.69	87.43	70.49	65.52	73.02	41.92	49.13	67.68	60.72
	MI-RE [28]	66.58	56.48	47.11	**98.57**	82.58	79.02	78.14	47.97	56.72	77.26	69.56
	MIG (**ours**)	**82.28**	**75.90**	**51.40**	96.98	**84.34**	**79.35**	**88.62**	**57.46**	**69.86**	**81.35**	**77.04**
DeiT-S	FGSM [14]	42.47	37.30	27.86	56.03	46.95	38.59	42.63	30.82	35.88	43.00	40.33
	PGD [26]	32.92	26.15	19.87	70.92	54.07	34.09	46.18	20.93	26.31	45.24	38.14
	PGD-RE[28]	49.56	30.55	43.61	96.25	79.56	62.77	79.84	41.22	41.24	71.93	60.42
	MI [9]	60.94	51.66	44.78	87.99	78.45	60.87	81.28	49.75	52.46	71.67	64.47
	MI-RE [28]	73.49	62.75	60.88	**99.35**	90.55	84.06	92.75	64.02	65.70	86.15	78.73
	MIG (**ours**)	**84.39**	**80.63**	**68.17**	97.78	**92.52**	**85.99**	**96.39**	**73.69**	**77.73**	**89.28**	**84.95**
DeiT-B	FGSM [14]	41.06	34.69	24.66	48.51	39.97	35.56	36.35	26.26	33.47	37.33	35.89
	PGD [26]	28.97	27.03	21.24	72.33	57.82	37.55	55.27	23.54	25.75	49.30	40.47
	PGD-RE[28]	57.36	49.24	37.95	94.66	70.31	58.17	80.22	46.37	48.18	69.95	61.79
	MI [9]	57.83	49.00	46.54	85.24	69.28	60.63	74.55	50.09	51.12	67.96	61.65
	MI-RE [28]	78.27	68.58	64.08	94.63	80.63	81.39	92.41	66.55	70.31	83.12	78.32
	MIG (**ours**)	**86.33**	**82.67**	**74.73**	**96.20**	**91.12**	**87.53**	**95.07**	**79.47**	**81.24**	**89.88**	**86.64**

Table 2. Attack success rate (%) and mean attack success rate (MASR, %) of MIG and other attacks, using CNNs as source models.

Source	Attack	Target Model								MASR	MASR	MASR
Model	Method	DN201	BiT	CNeXt	ViT-S	ViT-B	ViT-L	TNT	Swin	(CNNs)	(ViTs)	
VGG19	FGSM [14]	44.08	38.00	23.09	26.80	19.40	14.23	23.54	22.62	35.06	21.32	26.47
	PGD [26]	25.65	24.25	14.64	11.65	8.43	7.38	13.05	10.84	21.51	10.27	14.49
	MI [9]	56.11	45.73	43.27	26.25	20.11	13.73	33.23	17.84	48.37	22.23	32.03
	MIG (**ours**)	**95.73**	**91.92**	**67.57**	**60.89**	**48.84**	**31.17**	**64.36**	**47.58**	**85.07**	**50.57**	**63.51**
MNAS	FGSM [14]	40.06	30.07	22.14	26.35	20.13	15.06	28.66	22.84	30.76	22.61	25.67
	PGD [26]	27.46	22.24	17.69	15.47	11.29	9.93	15.55	11.74	22.46	12.80	16.42
	MI [9]	62.94	52.05	32.58	32.78	27.09	22.88	38.70	26.83	49.19	29.66	36.98
	MIG (**ours**)	**94.08**	**86.65**	**60.83**	**70.63**	**53.56**	**37.10**	**71.59**	**52.89**	**80.52**	**57.15**	**65.92**
Incep -v4	FGSM [14]	46.29	35.09	22.29	21.99	18.32	13.81	26.55	17.87	34.56	19.71	25.28
	PGD [26]	27.96	24.40	15.26	18.20	10.04	7.08	13.86	12.22	22.54	12.28	16.13
	MI [9]	58.73	48.44	39.66	33.02	23.84	16.16	36.55	24.95	48.94	26.90	35.17
	MIG (**ours**)	**95.38**	**90.61**	**75.15**	**73.04**	**62.45**	**49.40**	**70.38**	**59.62**	**87.05**	**62.98**	**72.01**

els, MIG brings 43.27% improvement over FGSM. For ensemble-based methods such as MI, our method outperforms it by 38.11% on CNNs and 36.08% on ViTs.

Remarkably, MIG exceeds these attack methods by a large margin when transferring to ViT models, approaching comparable attack success rate as attacks to CNNs. In fact, other methods struggle to adapt to ViT models, highlighting the effectiveness and versatility of our approach. These results also provide evidence that ViT models are not that robust against attacks, as long as the regions that are critical for the prediction of ViTs can be effectively identified and perturbed. And MIG is effective in finding vulnerable regions that are susceptible to both CNNs and ViTs, leading to a more stable and robust attack against them.

4.5. Results of MIG with Ensemble

MIG performs well without requiring any model or input ensembles. Prior studies [23, 44] demonstrate that ensembling can be an effective strategy for further improving the transferability of attacks. In this section, we examine the impact of combining MIG with model and input ensembles.

We first employ three different model ensemble strategies introduced in Section 3.4: perturbation ensemble, logit ensemble, and integrated gradients (IG) ensemble. We evaluate the impact of these model ensembles on three groups of models: an ensemble of two CNNs, an ensemble of two ViTs, and an ensemble of one CNN and one ViT model.

The results are shown in Table 3. We can draw four conclusions from these experiments. First, using logit ensem-

Table 3. Attack success rate (%) and mean attack success rate (MASR, %) of MIG with and without model ensemble. Perturbation, logit, and IG denote using perturbation ensemble, logit ensemble, and integrated gradients ensemble, respectively.

Source Model	Ensemble Strategy	Target Model							MASR (CNNs)	MASR (ViTs)	MASR
		DN201	BiT	CNeXt	ViT-S	ViT-B	ViT-L	TNT			
VGG19	No Ensemble	95.73	91.92	67.57	60.89	48.84	31.17	64.36	85.07	51.32	65.78
MNAS		94.08	86.65	60.83	70.63	53.56	37.10	71.59	80.52	58.22	67.78
DeiT-T		82.28	75.90	51.40	96.98	84.34	79.35	88.62	69.86	87.32	79.84
DeiT-S		84.39	80.63	68.17	97.78	92.52	85.99	96.39	77.73	93.17	86.55
VGG19	Perturbation	90.56	82.73	54.12	53.92	39.96	25.25	59.24	75.80	44.59	57.97
+	Logit	97.64	95.98	79.42	78.31	63.36	45.63	82.13	91.01	67.36	77.50
MNAS	IG	97.84	96.34	80.57	78.87	66.21	45.73	81.93	**91.58**	**68.19**	**78.21**
DeiT-T	Perturbation	76.36	71.64	53.46	92.17	82.33	70.43	90.16	67.15	83.77	76.65
+	Logit	90.26	87.70	74.00	98.54	94.98	89.76	97.69	83.99	95.24	90.42
DeiT-S	IG	90.36	87.85	73.84	98.49	95.38	89.82	98.04	**84.02**	**95.43**	**90.55**
VGG19	Perturbation	80.87	77.21	63.40	94.48	86.90	76.66	92.47	73.83	87.63	81.71
+	Logit	95.78	92.47	82.18	96.34	91.82	84.64	96.08	**90.14**	92.22	91.33
DeiT-S	IG	95.61	93.09	81.38	96.64	92.27	84.94	95.93	90.03	**92.45**	**91.41**

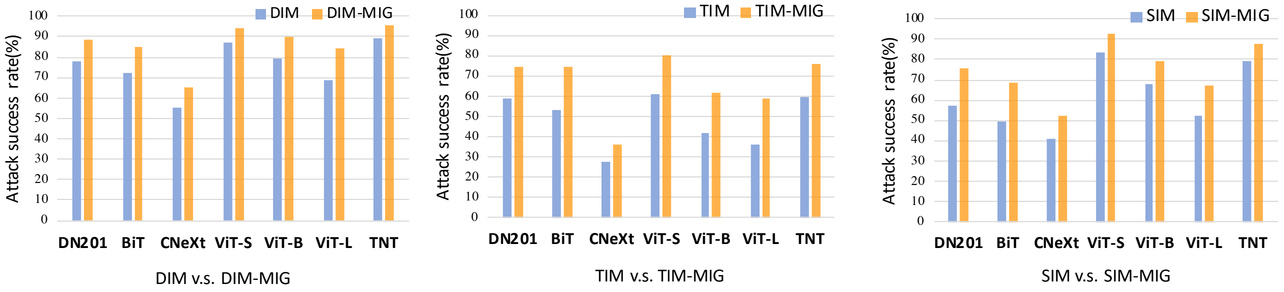

Figure 3. Attack success rate (%) of DIM [47], TIM [10], SIM [22] and their MIG-enhanced versions, using DeiT-S [43] as source model.

ble or IG ensemble can significantly improve the adversarial transferability and outperform two source models used for ensemble by 3.87% $\sim$ 25.63%, as indicated by MASR. As a comparison, direct ensembling perturbations leads to poor performance. This is possibly because directly integrating perturbations ignores the feedback of adversarial examples at each iteration. Second, ensembles of single-class models, especially ensembles of two CNN models, have more limited performance improvement when using ViTs as target models, but still leads to around 10% MASR improvement than single CNN setting, confirming that ViT models are indeed more robust than CNNs.

Third, the best performance is achieved when using one ViT and one CNN as source models, which may be due to the elimination of inherent limitations caused by differences in model architectures. In addition, we find that the effects of ensembling IG and logit are very similar. This further indicates that integrated gradients are strongly correlated with the model's output (logits) and are less related to implementation details, as described in Section 3.1.

As a conclusion, in practical scenarios with unknown target models, we only need to use an IG or logit ensemble of one CNN and one ViT to achieve good attack success rate.

To see the effects of using input ensembles, we integrate our method into previous input transformation attacks such as DIM [47], TIM [10] and SIM [22], which are introduced in Section 2.2. We maintain the default parameter settings for these methods, and use DeiT-S as source model, seven other models including both CNNs and ViTs as target models. Figure 3 shows the results, which demonstrates that MIG improves the attack success rates by about 5% $\sim$ 23% for these input-ensemble-based attack methods on various target models, indicating the effectiveness of MIG.

Additional results on various model and input ensemble settings can be found in the supplementary material.

4.6. Ablation Study

In this section, we study the contributions of all components used in MIG: a) Employ Integrated Gradients (IG) to replace direct gradients; b) Iteratively update the adversarial perturbations; and c) Accumulate historical gradients in a momentum manner. Regard these techniques as key components of MIG, we conduct ablation experiments. Specifically, we evaluate the attack success rate of adversarial examples by using ViT-L as target model and DeiT-S, DeiT-B, MNAS, and Inception-v4 as source models. To enhance

Table 4. Ablation study. Regard Integrated Gradients (IG), Iterative Update and Momentum Iterative Update as key components of MIG, we study the effectiveness of all these components respectively, using ViT-L [11] as target model.

IG	Iterative	Momentum Update	Source Model			
			DeiT-S	DeiT-B	MNAS	Incep-v4
			40.23	35.63	16.77	15.75
	✓		52.45	45.90	18.22	17.84
✓			19.87	13.86	12.57	9.55
✓	✓		62.19	65.91	28.83	27.36
	✓	✓	61.50	60.17	23.05	16.79
✓	✓	✓	**86.33**$^{(+46.10)}$	**87.85**$^{(+52.22)}$	**37.45**$^{(+20.68)}$	**50.25**$^{(+34.50)}$

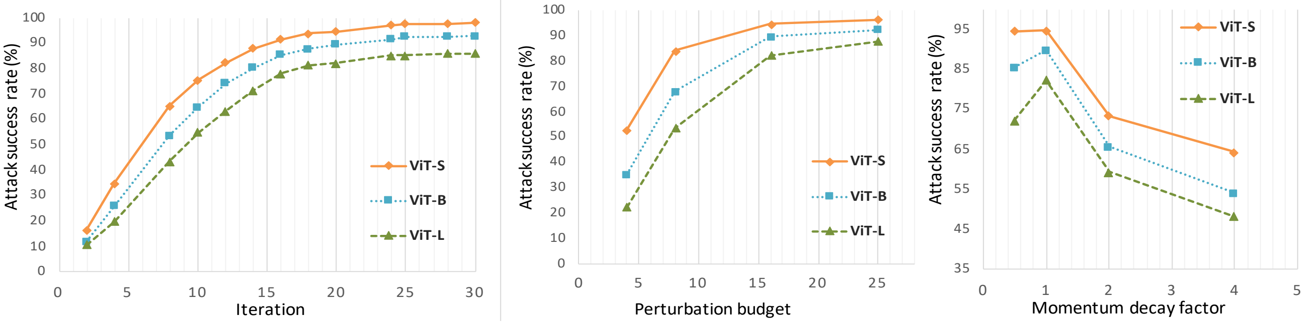

Figure 4. Attack success rate (%) of MIG attack under different hyper-parameter settings, using DeiT-S [43] as source model.

computational efficiency, we conduct experiments on a subset of the dataset in Section 4.1, consisting of 1,000 images.

The results are shown in Table 4. Note that even using ViTs as source models, attack success rates are still low without MIG. Concretely, using IG and iterative strategy to guide the generation of adversarial perturbations can bring around 12% $\sim$ 30% performance improvement when using different source models, indicating that the direct gradients may be sub-optimal. The momentum updating strategy then provides improvement by 8.62% $\sim$ 24.14%, suggesting that the accumulation of historical gradients can assist the algorithm to obtain perturbations that transfer better.

Note that using IG solely and crafting perturbations in a single-step leads to poor performance. This is probably because IG only considers the prediction of the classification model, and a single step update with IG's guidance cannot generate a significant enough impact on the loss function.

4.7. Hyper-parameters Selection

We investigate the impact of different hyper-parameters on the performance of MIG. Specifically, we vary the decay factor μ, the number of iterations T, and the perturbation budget ϵ, and record the transfer performance under different settings. For each setting, we use DeiT-S as the source model and evaluate the transferability of the adversarial examples on ViT-S, ViT-B, and ViT-L, respectively.

Figure 4 shows the attack success rate of MIG under different hyper-parameter settings. We observe that the performance increases as the number of iterations and the perturbation budget increase, and reaches stability at approximately iteration number $T = 25$, perturbation budget $\epsilon = 16/255$. As for the momentum decay factor μ, the optimal performance is obtained when $\mu = 1$, after which the performance decreases as the decay factor increases.

Note that when we increase the scale of the target model (varying from ViT-S to ViT-L), the trend of attack success rate varies consistently with the hyper-parameters. That is, the optimal performance can be achieved with the same hyper-parameter setting for all these target models.

5. Conclusion

Vision Transformers (ViTs) have demonstrated greater resilience in image classification than standard Convolutional Neural Networks (CNNs), and it has been a question whether we can produce an attack that works on different ViT models, or even on both ViTs and CNNs. Intuition suggests that ViTs are more robust in that they identify the global dependencies and semantically relevant regions in images better, while integrated gradients (IG) can effectively indicate these regions across different models. Therefore, we explicitly target this advantage by perturbing these regions with the guidance of IG, and apply the momentum iterative method to further enhance attack quality. Our approach not only achieves state-of-the-art transfer attack success rates on ViTs, ViT-CNN transfers, and model ensembles, but also provides an intuitive suggestion of ViTs' possible vulnerabilities in general.

References

[1] Alaaeldin Ali, Hugo Touvron, Mathilde Caron, Piotr Bojanowski, Matthijs Douze, Armand Joulin, Ivan Laptev, Natalia Neverova, Gabriel Synnaeve, Jakob Verbeek, et al. Xcit: Cross-covariance image transformers. *Advances in neural information processing systems*, 34:20014–20027, 2021.

[2] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. In *International conference on machine learning*, pages 284–293. PMLR, 2018.

[3] 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 international conference on computer vision*, pages 10231–10241, 2021.

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 213–229. Springer, 2020.

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[6] Jianbo Chen, Michael I Jordan, and Martin J Wainwright. Hopskipjumpattack: A query-efficient decision-based attack. In *2020 ieee symposium on security and privacy (sp)*, pages 1277–1294. IEEE, 2020.

[7] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM workshop on artificial intelligence and security*, pages 15–26, 2017.

[8] Minhao Cheng, Thong Le, Pin-Yu Chen, Jinfeng Yi, Huan Zhang, and Cho-Jui Hsieh. Query-efficient hard-label black-box attack: An optimization-based approach. *arXiv preprint arXiv:1807.04457*, 2018.

[9] 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*, pages 9185–9193, 2018.

[10] Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4312–4321, 2019.

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[12] Gamaleldin Elsayed, Shreya Shankar, Brian Cheung, Nicolas Papernot, Alexey Kurakin, Ian Goodfellow, and Jascha Sohl-Dickstein. Adversarial examples that fool both computer vision and time-limited humans. *Advances in neural information processing systems*, 31, 2018.

[13] Yonggan Fu, Shunyao Zhang, Shang Wu, Cheng Wan, and Yingyan Lin. Patch-fool: Are vision transformers always robust against adversarial perturbations? *arXiv preprint arXiv:2203.08392*, 2022.

[14] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[15] Kai Han, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing Xu, and Yunhe Wang. Transformer in transformer. *Advances in Neural Information Processing Systems*, 34:15908–15919, 2021.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[17] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[18] Yi Huang and Adams Wai-Kin Kong. Transferable adversarial attack based on integrated gradients. *arXiv preprint arXiv:2205.13152*, 2022.

[19] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, 1998.

[20] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 491–507. Springer, 2020.

[21] Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, pages 99–112. Chapman and Hall/CRC, 2018.

[22] Jiadong Lin, Chuanbiao Song, Kun He, Liwei Wang, and John E Hopcroft. Nesterov accelerated gradient and scale invariance for adversarial attacks. *arXiv preprint arXiv:1908.06281*, 2019.

[23] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. *arXiv preprint arXiv:1611.02770*, 2016.

[24] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[25] 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*, pages 11976–11986, 2022.

[26] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[27] 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.

[28] Muzammal Naseer, Kanchana Ranasinghe, Salman Khan, Fahad Shahbaz Khan, and Fatih Porikli. On improving adversarial transferability of vision transformers. *arXiv preprint arXiv:2106.04169*, 2021.

[29] Muhammad Muzammal Naseer, Kanchana Ranasinghe, Salman H Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers. *Advances in Neural Information Processing Systems*, 34:23296–23308, 2021.

[30] Sayak Paul and Pin-Yu Chen. Vision transformers are robust learners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2071–2081, 2022.

[31] Boris T Polyak. Some methods of speeding up the convergence of iteration methods. *Ussr computational mathematics and mathematical physics*, 4(5):1–17, 1964.

[32] Ning Qian. On the momentum term in gradient descent learning algorithms. *Neural networks*, 12(1):145–151, 1999.

[33] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[34] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[35] 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*, 115:211–252, 2015.

[36] Rulin Shao, Zhouxing Shi, Jinfeng Yi, Pin-Yu Chen, and Cho-Jui Hsieh. On the adversarial robustness of visual transformers. *arXiv preprint arXiv:2103.15670*, 2(7), 2021.

[37] Yucheng Shi and Yahong Han. Decision-based black-box attack against vision transformers via patch-wise adversarial removal. *arXiv preprint arXiv:2112.03492*, 2021.

[38] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[39] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.

[40] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

[41] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[42] Mingxing Tan, Bo Chen, Ruoming Pang, Vijay Vasudevan, Mark Sandler, Andrew Howard, and Quoc V Le. Mnasnet: Platform-aware neural architecture search for mobile. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2820–2828, 2019.

[43] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021.

[44] Xiaosen Wang and Kun He. Enhancing the transferability of adversarial attacks through variance tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1924–1933, 2021.

[45] 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*, volume 36, pages 2668–2676, 2022.

[46] Ross Wightman. Pytorch image models. `https://github.com/rwightman/pytorch-image-models`, 2019.

[47] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2730–2739, 2019.

[48] Zheng Yuan, Jie Zhang, and Shiguang Shan. Adaptive image transformations for transfer-based adversarial attack. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part V*, pages 1–17. Springer, 2022.

[49] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pages 7472–7482. PMLR, 2019.

[50] Daquan Zhou, Zhiding Yu, Enze Xie, Chaowei Xiao, Animashree Anandkumar, Jiashi Feng, and Jose M Alvarez. Understanding the robustness in vision transformers. In *International Conference on Machine Learning*, pages 27378–27394. PMLR, 2022.