Ensemble Diversity Facilitates Adversarial Transferability

Bowen Tang[1], Zheng Wang[12]*, Yi Bin[1], Qi Dou[3], Yang Yang[1], Heng Tao Shen[1]
[1]University of Electronic Science and Technology of China
[2]Institute of Electronic and Information Engineering of UESTC, Guangdong
[3]The Chinese University of Hong Kong
{tangbowenbw, dlyyang}@gmail.com, {zh_wang, yi.bin, shenhengtao}@hotmail.com, qidou@cuhk.edu.hk

Abstract

With the advent of ensemble-based attacks, the transferability of generated adversarial examples is elevated by a noticeable margin despite many methods only employing superficial integration yet ignoring the diversity between ensemble models. However, most of them compromise the latent value of the diversity between generated perturbation from distinct models which we argue is also able to increase the adversarial transferability, especially heterogeneous attacks. To address the issues, we propose a novel method of Stochastic Mini-batch black-box attack with Ensemble Reweighing using reinforcement learning (SMER) to produce highly transferable adversarial examples. We emphasize the diversity between surrogate models achieving individual perturbation iteratively. In order to customize the individual effect between surrogates, ensemble reweighing is introduced to refine ensemble weights by maximizing attack loss based on reinforcement learning which functions on the ultimate transferability elevation. Extensive experiments demonstrate our superiority to recent ensemble attacks with a significant margin across different black-box attack scenarios, especially on heterogeneous conditions. https://github.com/tangbwb/SMER

1. Introduction

It is acknowledged that CNNs are susceptible to adversarial attacks involving imperceptible perturbations [1, 11]. A plethora of studies [4, 37, 38, 49] suggests that the generated adversarial examples exhibit cross-model transferability, thus enhancing black-box attacks. Furthermore, these adversarial examples play a pivotal role in bolstering the robustness of CNNs as evidenced by improvements in error-tolerant rates [5, 31] and aiding other applications [39–41]. Additionally, the rise in popularity of ViTs [3] has spurred corresponding investigations into adversarial robustness [6, 26–28, 42]. A diverse array of technical approaches has been developed to achieve black-box adversarial attacks. Early methods primarily rely on gradient-based optimization techniques [11, 16]. Subsequent advancements [8–10, 20] elevate the potency of the attacks to a higher level.

Ensemble-based strategies [21, 22, 36] are pragmatic to integrate different models, albeit simply average the fusion. The advantages of ensemble-based methods over typical

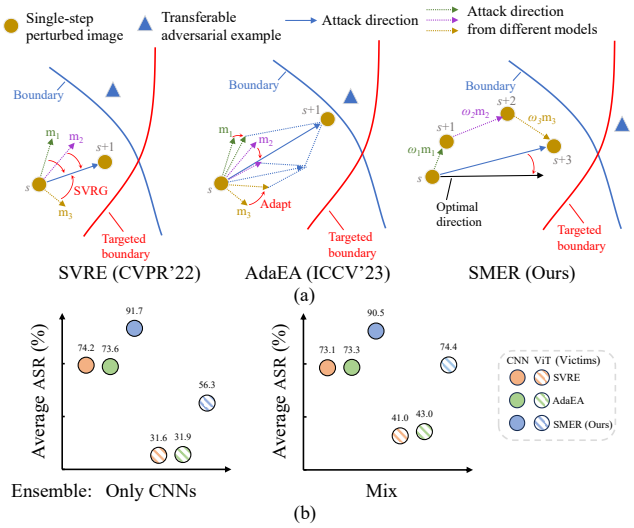

Figure 1. (a) The optimal direction search of SVRE, AdaEA and SMER. Existing approaches primarily focus on reducing ensemble diversity, whereas diversity plays a crucial role in enhancing transferability in our proposal. It is worth noting that for better clarification, our SMER incorporates three iterative perturbation steps, ranging from $s+1$ to $s+3$, whereas the others employ only a single iterative step. (b) The average attack performance, based on TI-DIM, of existing ensemble-based strategies sharply declines under heterogeneous conditions.

*Corresponding author. This work was supported part by the National Natural Science Foundation of China (U20B2063, 62306065, and 62220106008), part by Sichuan Science and Technology Program, China (2023YFG0289), and also part by the Guangdong Basic and Applied Basic Research Foundation (2022A1515110576, 2023A1515140104).

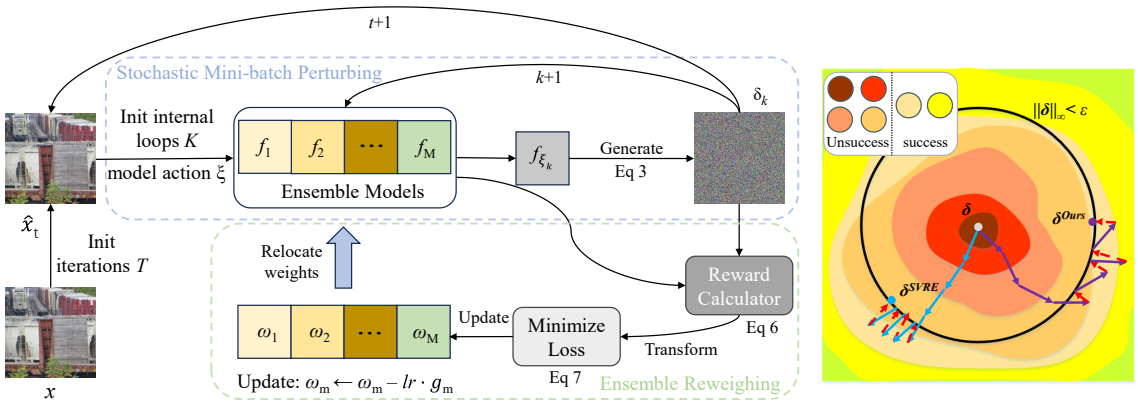

Figure 2. The diagram illustrates the procedure of our method attacking a benign image x. (**Left**) There is an ensemble set $\{f\}$ and its corresponding weights set $\{\omega\}$. In each internal loop, the perturbation δ_k is maximized under the currently stochastically selected model f_{ξ_k}. To leverage the contributions between models, we bring the reward of reinforcement learning to update the weights ω by optimizer with transformed reward. (**Right**) The graph illustrates the internal perturbation search process. δ^{SVRE} and δ^{Ours} denote the perturbation of SVRE and our algorithm respectively, while the redder color indicates the lower loss. The optimization is accelerated by the variant gradient along the bound in limited searches.

gradient-based approaches lie in two key factors: (1) They can be combined with advanced gradient-based methods, $e.g.$ translation-invariant method (TIM) [11], (2) They harness the strengths from diverse surrogate models. Typically, ensemble-based methods focus on leveraging the latter advantage, exploring how to integrate surrogate models to produce adversarial examples with high transferability. As depicted in Figure 1 (a), Stochastic Variance Reduced Ensemble Adversarial Attack (SVRE) [45] incorporates the optimization of Stochastic Variance Reduced Gradient (SVRG) [18] to reduce the diversity of generated perturbations from distinct models. Similarly, Adaptive Model Ensemble Adversarial Attack (AdaEA) [6] introduces a discrepancy ratio and utilizes a disparity-reduced filter to fuse ensemble outputs with reduced diversity. SVRE and AdaEA aim to update iterative perturbations using SVRG and adaptation techniques, respectively, to optimize ensemble average gradients. Consequently, optimization in each step across the entire ensemble can be viewed as a batch operation.

It is noteworthy that attacks can be categorized into three types: 1) **homogeneous attacks based on homogeneous ensemble**, $e.g.$ from only CNNs ensemble to CNN victims, 2) **heterogeneous attacks based on homogeneous ensemble**, $e.g.$ from only CNNs ensemble to ViT victims, 3) **heterogeneous ensemble**, $i.e.$ Mix ensemble. Existing methods manually decrease the discrepancy of generated perturbation between surrogates on the iterative batch-like optimization and obtain a satisfactory ASR in homogeneous attacks. However, ASR drops significantly in both heterogeneous attacks and ensemble scenarios. Thus, strategies for mitigating diversity may compromise individual superiority in either heterogeneous attacks or ensemble settings. It is worth considering allowing individual models more freedom, which may further enhance the versatility of black-box attacks in both heterogeneous attacks and ensembles.

In contrast to existing methods [6, 45], we refrain from constraining the diversity of generated perturbations across different surrogates, a challenging task given the variability in architectures. Instead, we advocate for producing adversarial examples from individual models independently, akin to mini-batch optimization for iterative perturbations as illustrated in Figure 1 (a). Additionally, we note that each individual surrogate impacts black-box results differently. Therefore, to further tailor the individuality between surrogates, we propose reweighing ensemble weights using reinforcement learning, aiming to maximize attack loss and determine the optimal direction. Consequently, the iteratively generated adversarial perturbation is designed to maximize the attack loss for the current surrogate.

The summary of our contributions is as follows:

- We introduce a novel stochastic mini-batch black-box attack with ensemble reweighing using reinforcement learning. Unlike conventional approaches, our method embraces the unconstrained diversity of generated perturbations, as illustrated on the left side of Figure 2.
- To enhance the influence of diversity, we propose stochastic mini-batch perturbation to generate adversarial examples from each individual surrogate in the current iteration. Furthermore, to differentiate the diversity within the ensemble, we introduce ensemble reweighing to adjust ensemble weights using reinforcement learning.
- SMER serves as a plug-and-play method, offering convenience in integrating with multiple ensemble surrogates across various basic attacks. Extensive experiments validate the high effectiveness and superior transferability of our generated adversarial examples in black-box settings.

2. Related Work

Gradient-based Attacks. Gradient descent, the primary optimization method for CNNs, can have an inverse effect when employing gradient ascent. Consequently, numerous methods have emerged to integrate single surrogates for transfer-based black-box attacks. Goodfellow *et al.* [11] introduced the fast gradient sign method (FGSM) for adversarial attacks in one step. Iterative strategies [20, 25] extend the Attack Success Rate (ASR) of one-step attacks. Attacks with momentum [8] enhance transferability and effectiveness in black-box scenarios. Techniques like random resizing, padding [44], translation-invariant methods [9], and scale-invariant methods [21] reshape inputs to improve transferable attack performance.

The performance of transfer-based black-box attacks with a single surrogate falls below expectations due to limited transferability.

Ensemble-based Attacks. The ensemble-based method has shown promise in enhancing the performance of transfer-based attacks [22]. Three common approaches to ensemble include ensemble on predictions, ensemble on loss, and ensemble on logits, with the latter proving more effective [8]. Consequently, recent studies predominantly focus on leveraging ensemble surrogates for improved efficacy and efficiency. Zheng *et al.* [47] explore generating adversarial examples using meta-learning with separate ensemble surrogates. Addressing potential issues with vanishing gradients during attacks, SVRE [45] reduces gradient variance using SVRG [18]. To mitigate discrepancies between CNNs and ViTs [3], AdaEA [6] fuses ensemble outputs by monitoring discrepancy ratios and updates direction using a disparity-reduced filter, thereby reducing perturbation discrepancies among surrogates.

Despite the significant results presented by the aforementioned pioneering ensemble-based approaches, their attack performance requires further improvement, whether the victim is homogeneous or heterogeneous.

Defensive Methods. Due to the susceptibility of Deep Neural Networks (DNNs) to adversarial attacks, several studies have emphasized the need for robustness enhancement. Adversarial training has proven effective in bolstering classifier robustness [25, 27, 36]. Additionally, other approaches directly operate on input adversarial examples, such as reversing adversarial features [29, 46], utilizing compression techniques [12, 23], and employing randomized smoothing [17]. Furthermore, the emergence of recent ViT [3] has demonstrated superior performance compared to conventional CNNs, albeit with a similar vulnerability to adversarial attacks. Consequently, defensive approaches have been proposed, focusing on structural modifications for state-of-the-art recognition models. For instance, the Robust Vision Transformer [28] enhances ViT blocks with robust transformations, while Mao *et al.* [26] improve the generalization and robustness of ViT through discrete representation, resulting in Robust ViTs equipped with innovative architectures and resilience against adversaries.

In general, the versatility of adversarial attacks against both CNNs and ViTs can be more effectively gauged through advanced defensive methods, especially considering that commonly trained models have been extensively scrutinized and exhibit superior ASR [45].

Algorithm 1 SMER with MI-FGSM attack algorithm

Input: A benign image x and its corresponding label y, ensemble surrogates $\{f_1, f_2, \ldots, f_M\}$, cross-entropy loss J, ensemble weights $\{\omega_1, \omega_2, \ldots, \omega_M\}$, reinforcement learning reward R, weights update scheme Sh, perturbation norm L_∞ and bound ε, number of iterations T, number of internal loops K, step size α, internal step size β, decay factor μ, internal decay factor $\tilde{\mu}$.

Output: Adversarial example $\hat{x}$ with $\|\hat{x} - x\|_\infty \leq \varepsilon$.

$\alpha = \varepsilon / T$;
$g_0 = 0$; $\hat{x}_0 = x$;
Initialize ensemble surrogates $\omega f(x)$;
Initialize Sh;
for $t = 0$ to $T - 1$ **do**
  $\tilde{g}_0 = 0$; $\tilde{x}_0 = \hat{x}_t$;
  Initialize model action ξ;
  for $k = 0$ to $K - 1$ **do**
    Calculate logits $\mathbf{l}(\tilde{x}_k) = \omega_{\xi_k} \cdot f_{\xi_k}(\tilde{x}_k)$;
    Update internal gradient $\tilde{g}_{k+1} = \tilde{\mu} \cdot \tilde{g}_k + \frac{\nabla J(\mathbf{l}(\tilde{x}_k), y)}{\|\nabla J(\mathbf{l}(\tilde{x}_k), y)\|_1}$;
    Calculate reward $R = J(\sum_{i=1}^{M} \omega_i f_i(\tilde{x}_k), y)$;
    Update w by $Sh(R)$;
    $\tilde{x}_{k+1} = Clip_x^{\varepsilon}(\tilde{x}_k + \beta \cdot sign(\tilde{g}_{k+1}))$;
  end for
  Update gradient $g_{t+1} = \mu \cdot g_t + \frac{\tilde{g}_K}{\|\tilde{g}_K\|_1}$;
  $\hat{x}_{t+1} = Clip_x^{\varepsilon}(\hat{x}_t + \alpha \cdot sign(g_{t+1}))$;
end for
return $\hat{x} = \hat{x}_T$.

3. Methodology

3.1. Preliminaries

The ensemble-based attacks [8] commonly employ the additive perturbation δ [7, 11] that is generated from the gradient backward to perturb a benign image x and produce adversarial example $\hat{x}$, *i.e.* $\hat{x} = x + \delta$. It is usual to utilize L_p norm, which can be denoted as $\|\delta\|_p \leq \varepsilon$, to constrain δ for the imperceptibility. To align with previous works, we employ L_∞ for the following comparisons. The perturbed image $\hat{x}$ is expected to fool the image recognition victim f, *i.e.* $f(\hat{x}) \neq y$, where y denotes the ground-truth label of the image x. Hence, the typical process of the iteratively

gradient-based attack with the single surrogate model can be described as:

$$\hat{x}_{t+1} = \hat{x}_t + \alpha \cdot sign(\nabla_x J(f(\hat{x}_t), y)), \tag{1}$$

where J is the loss function and $\hat{x}_t$ denotes the iteratively generated adversarial example. The gradient-based attack can be extended into an ensemble-based method when incorporating multiple models. As demonstrated in the results of MI-FGSM [8], it exhibits superior performance when applied to ensembles based on logits, a technique widely adopted in related works [45]. The ensemble operation can be represented as $\mathbf{l}(x) = \sum_{i=1}^{M} \omega_i f_i(x)$, where $\mathbf{l}(x)$ denotes the ensemble logits derived from M surrogates. Thus, incorporating cross-entropy loss, the ensemble-based attack can be expressed as:

$$\hat{x}_{t+1} = \hat{x}_t + \alpha \cdot sign(\nabla_x(-\mathbb{1}_y \cdot \log(softmax(\mathbf{l}(x))))), \tag{2}$$

where $\mathbb{1}_y$ represents the one-hot encoding of ground-truth label y. SVRE [45] is to mitigate the difference of generated perturbation between models. Their main contribution, the internal update, $i.e.$ $g_m = \nabla_x J(\hat{x}_m, y) - \nabla_x J(\hat{x}_t, y) + g_t^{ens}$, where the unbiased estimate g_m is to reduce the gradient variance, $i.e.$ generated perturbation diversity between surrogates and $\hat{x}_m$ denotes the generated adversarial example within internal loops. On the one hand, SVRE mitigates generated gradient diversity to satisfy $\mathbb{E}(\nabla_x J(\hat{x}_t, y)) = g_t^{ens}$ that makes generated perturbation from each surrogate converge. On the other hand, the batch-like optimization limits the performance when it encounters heterogeneous victims due to constraining the gradient diversity as shown in Figure 1 (b).

Therefore, we propose the method utilizing stochastic mini-batch to maximize the perturbation from individual models with ensemble reweighing using reinforcement learning for refining model diversity as depicted in the left of Figure 2.

3.2. Stochastic Mini-batch Perturbing

Considering the diversity of back-propagated gradients [6] and the inefficiency of methods mitigating diversity, particularly in cross-architecture attacks as depicted in Figure 1 (b), we propose a straightforward approach to exploit diversity by treating each surrogate as an individual. Thus, the iterative perturbation optimization resembles mini-batch optimization, wherein the iterative mini-batch perturbation is solely derived from the current surrogate. For ensemble-based attacks, M ensemble surrogates can be denoted as $\{f_1, f_2, \ldots, f_M\}$. The initial perturbation can be assumed as $\{\mathbf{0}\}^{C\times H\times W}$ which can be denoted in the right of Figure 2. In most gradient-based iterative attack methods, perturbation will be constantly optimized using gradient ascent after every iteration.

Therefore, we employ internal iterative perturbing to search the optimal direction for producing the outer perturbation. The ultimate desire is to produce the optimal perturbation, $i.e.$ δ^*, and perturb the benign image, $i.e.$ $\hat{x} = x + \delta^*$, for fooling the entire ensemble surrogates as much as possible. Hence, the perturbation iterative generation $G(x)$ resembles the perturbation optimization, and the process can be formulated as:

$$\delta^* = \inf_x G(x), \tag{3}$$

where the expectation is to satisfy $\mathbb{E}[\|G(x) - \delta^*\|] \leq \epsilon$. In the k inner loop, a single surrogate model is randomly selected to perform a one-step perturbation. It is essential to note that each surrogate model holds equal importance, requiring us to traverse all surrogate models after M loops. The model selected for generating stochastic mini-batches is determined by ξ, a stochastic sequence of model indices following the aforementioned selection protocol. Each model assesses the input image and generates scores, and the perturbation can achieve its optimal value for the current model following gradient calculation in the current loop. Consequently, the update for each inner loop can be described as:

$$\tilde{x}_{k+1} = \tilde{x}_k + \alpha \cdot sign(\nabla_x J(f_{\xi_k}(\tilde{x}_k), y)). \tag{4}$$

The current adversarial example $\tilde{x}_{k+1}$ is derived from the former one $\tilde{x}_k$. $G(x_k) = \alpha \cdot sign(\nabla_x J(f_{\xi_k}(\tilde{x}_k), y))$ denotes the perturbation generation, and it is bounded as $\|G(x)\|_\infty \leq \varepsilon$. Consequently, the search process can be demonstrated as:

$$\hat{x} = \arg\max_x J(f_\xi(x), y). \tag{5}$$

3.3. Ensemble Reweighing

To enhance the impact of diverse perturbations, we propose differentiating diversity further by individual models. Given the unknown specific expected adversarial examples, we utilize unsupervised ensemble reweighing with reinforcement learning to optimize the process, as outlined in Algorithm 1. In each loop, the weights $\{\omega_1, \omega_2, \ldots, \omega_M\}$ serve as the agents to be optimized through reinforcement learning, with updates based on the reward R. The ultimate goal is to generate perturbations that threaten all surrogate models, maximizing the total attack rewards, $i.e.$ $R_{total} = \max \sum_{j=1}^{T\times K} R_j$. It is important to note that each update in the reinforcement process relies solely on the immediately preceding result $\tilde{x}_k$, and the subsequent results are unpredictable. Empirical evidence suggests that maximizing the attack loss J leads to optimal attack performance, while the ensemble surrogates are accessible. Therefore, it is practical to maximize the internal reward R_j and conduct model-based reinforcement learning for the update. We simplify

Base	Attack	Adversarial Training				Transformation-based Method							Robust Vision Transformer			
		Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens}$	Average	R&P	NIPS-r3	Bit-R	JPEG	FD	ComDefend	Average	RVT-S*	Drvit	Vit+DAT	Average
I-FGSM	Ens [8]	27.1	24.5	15.7	22.4	15.2	18.9	26.0	41.8	37.1	56.0	32.5	31.8	33.9	20.1	28.6
	SVRE [45]	40.1	37.3	24.8	34.0	25.0	34.1	31.0	62.1	50.4	67.0	44.9	40.6	39.0	24.0	34.5
	AdaEA [6]	35.4	31.8	18.6	28.6	17.9	28.0	46.5	66.8	53.9	82.8	49.3	42.4	43.9	25.3	37.2
	SMER	**57.9**	**51.0**	**38.8**	**49.2**	**38.1**	**50.0**	**47.1**	**75.7**	**57.2**	**93.9**	**60.3**	**49.6**	**44.4**	**29.3**	**41.1**
MI-FGSM	Ens [8]	50.5	49.3	32.3	44.0	33.0	44.6	39.7	75.9	62.8	77.5	55.6	60.0	59.2	39.5	52.9
	SVRE [45]	64.5	59.0	39.1	54.2	40.7	59.5	43.4	89.1	73.3	86.6	65.4	70.6	66.6	44.0	60.4
	AdaEA [6]	45.5	41.4	24.2	37.0	26.1	41.9	69.8	81.4	73.6	91.1	64.0	59.1	59.5	37.5	52.0
	SMER	**75.2**	**66.0**	**45.2**	**62.1**	**53.7**	**70.6**	**83.2**	**94.2**	**88.9**	**98.6**	**81.6**	**75.2**	**69.9**	**48.6**	**64.6**
TIM	Ens [8]	73.5	68.1	59.7	67.1	60.5	67.2	49.3	82.6	74.8	85.1	69.9	52.8	69.1	40.7	54.2
	SVRE [45]	87.9	85.6	79.7	84.4	80.2	83.8	62.3	92.0	84.0	92.2	82.4	65.9	77.1	47.0	63.3
	AdaEA [6]	82.1	79.7	74.0	78.6	72.6	76.9	87.1	85.0	87.7	94.1	83.9	60.9	75.3	42.3	59.5
	SMER	**92.5**	**89.6**	**86.1**	**89.4**	**85.2**	**88.7**	**93.9**	**95.4**	**95.3**	**98.3**	**92.8**	**77.9**	**82.1**	**58.5**	**72.8**
TI-DIM	Ens [8]	87.4	84.3	77.6	83.1	81.2	85.7	63.0	91.7	84.3	91.9	83.0	58.2	72.9	42.7	57.9
	SVRE [45]	95.3	93.7	90.1	93.0	91.9	93.2	72.9	96.5	90.8	96.0	90.2	75.7	84.5	54.7	71.6
	AdaEA [6]	84.8	83.6	78.1	82.2	76.9	81.6	89.8	88.6	90.5	90.9	86.4	63.4	77.8	44.7	62.0
	SMER	**98.2**	**96.3**	**94.7**	**96.4**	**95.2**	**96.3**	**98.3**	**98.8**	**98.7**	**99.1**	**97.7**	**88.4**	**89.9**	**71.5**	**83.3**

Table 1. The ASR (%) on adversarial training, transformation-based methods and robust ViTs between Ens, SVRE, AdaEA and SMER. The adversarial examples are generated on four ensemble surrogates: Inc-v3, IncRes-v2, Res-101 and Inc-v4, with four attack baselines.

R_j to R, which can be described as:

$$R = \max_{\tilde{x}_k} J(\sum_{i=1}^{M} \omega_i f_i(\tilde{x}_k), y). \tag{6}$$

However, to conduct the update scheme in practice, we transform the reward function to a minimized loss. From our observation, loss $J_R = -\ln(R)$ contributes positive results to optimizing weights w, and stochastic gradient descent optimizer (SGD) is efficient in this task of limited iteration for parameterized weights. The update scheme for the weights w is denoted as Sh in Algorithm 1. The objective function can be written as:

$$w = \arg\min_{\tilde{x}_k} J_R(\sum_{i=1}^{M} \omega_i f_i(\tilde{x}_k), y). \tag{7}$$

The objective function of the SMER can be eventually summarized as:

$$\hat{x} = \arg\max_{x} J(\omega_\xi f_\xi(x), y) - J_R(\sum_{i=1}^{M} \omega_i f_i(x), y). \tag{8}$$

4. Experiments

In this section, we validate the performance of SMER under settings of both homogeneous and heterogeneous attacks, as well as against defensive methods. Additionally, we investigate feature attention alteration, iterative discrepancy, average attack loss performance, and ensemble types.

4.1. Experimental Setup

Datasets. Experiments are conducted on ImageNet-compatible datasets [30] which is deliberately built and utilized in recently related attack works [9, 45].

Models. We select abundant CNNs including Inception v3 (Inc-v3) [33], Inception ResNet v2 (IncRes-v2) [13], ResNet v2-101 (Res-101) [34], ResNet v2-152 (Res-152) [34], Inception v4 (Inc-v4) [34], ResNet-18 (Res-18) [14], ResNet-50 (Res-50) [14], WideResNet-50 (WRN50) [48], WideResNet-101 (WRN101) [48], BiT-M-R50×1 (BiT-50) [19] and BiT-M-R101 (BiT-101) [19]. In addition, three adversarial-trained models, Inc-v3$_{ens3}$ [36], Inc-v3$_{ens4}$ [36] and IncRes-v2$_{ens}$ [36] are chosen for evaluation. Moreover, a group of ViTs including ViT-Base (ViT-B) [3], ViT-Tiny (ViT-T) [3], ViT-Small (ViT-S) [3], DeiT-Base (DeiT-B) [35], DeiT-Tiny (DeiT-T) [35], DeiT-Small (DeiT-S) [35], Swin-Base (Swin-B) [24], Swin-Tiny (Swin-T) [24], Swin-Small (Swin-S) [24] are selected in our experiments. The specific configurations are shown below the tables.

Defenses. To evaluate the versatility of SMER, we choose defensive methods including R&P [43], NIPS-r3 [2], Bit-R [46], JPEG [12], FD [23], ComDefend [15]. Besides, another three robust vision transformers, RVT-S* [28], Drvit [26] and Vit+Dat [27] are employed for evaluation.

Baselines and comparisons. Following the latest works [6, 45], we also integrate our SMER method into four attack baselines, including $i.e.$ I-FGSM [11], MI-FGSM [8], TIM [9] and TI-DIM [9]. At the same time, we also share identical implementations with ensemble attack (Ens) [8], SVRE [45], and AdaEA [6].

Parameters. We implement experiments on the same structural sets. The perturbation bound $\varepsilon = 16/255$, the iterative number T is set to 10, and the step size $\alpha = \varepsilon/10$. For methods with momentum, the decay factor is 1.0. For methods equipping with translation kernel, the size is 5×5. The diverse and scale-invariant operation refer to [21]. The internal loops K are four times of model number. The internal

Base	Attack	CNN						ViT						
		Res-50	WRN101	Inc-v4	BiT-50	BiT-101	Average	ViT-B	ViT-S	DeiT-B	DeiT-S	Swin-B	Swin-S	Average
I-FGSM	Ens [8]	57.5	66.0	64.7	65.4	54.7	61.7	46.1	70.2	64.4	84.3	27.8	42.2	55.8
	SVRE [45]	63.9	72.1	71.2	69.8	58.6	67.1	38.9	58.2	55.1	71.3	27.6	44.7	49.3
	AdaEA [6]	74.0	80.3	74.7	78.3	69.4	75.3	68.0	90.8	85.1	95.1	46.9	**66.6**	75.4
	SMER	**77.8**	**83.1**	**79.0**	**81.7**	**73.5**	**79.0**	**69.1**	**91.0**	**85.8**	**97.3**	**47.0**	63.5	**75.6**
MI-FGSM	Ens [8]	80.6	84.1	82.4	83.8	77.9	81.8	73.0	90.2	88.2	96.0	53.7	69.9	78.5
	SVRE [45]	85.3	87.8	87.2	87.2	79.5	85.4	67.5	85.6	81.8	91.1	52.1	70.7	74.8
	AdaEA [6]	83.5	84.4	81.4	85.3	77.5	82.4	77.5	94.5	89.5	97.9	60.5	74.4	82.4
	SMER	**91.6**	**92.3**	**91.3**	**91.7**	**87.0**	**90.8**	**88.0**	**97.8**	**95.6**	**99.6**	**66.7**	**82.5**	**88.4**
TIM	Ens [8]	51.4	67.6	71.4	75.5	63.5	65.9	40.5	62.8	52.9	73.6	17.0	29.5	46.1
	SVRE [45]	62.3	78.3	80.6	79.7	68.5	73.9	42.0	66.4	57.5	75.5	21.8	36.3	49.9
	AdaEA [6]	60.9	75.0	77.7	80.5	71.4	73.1	53.8	81.3	69.6	88.0	24.5	41.8	59.8
	SMER	**79.5**	**87.8**	**87.4**	**89.7**	**82.2**	**85.3**	**72.9**	**91.3**	**84.6**	**96.2**	**37.1**	**59.2**	**73.6**
TI-DIM	Ens [8]	60.2	74.7	77.5	80.8	71.9	73.0	39.2	63.0	50.2	71.5	18.1	32.1	45.7
	SVRE [45]	69.5	83.2	90.0	87.7	79.8	82.0	48.0	73.8	62.6	81.3	24.7	43.7	55.7
	AdaEA [6]	63.5	77.4	79.2	83.4	73.5	75.4	48.1	74.8	63.2	82.5	24.2	40.9	55.6
	SMER	**90.8**	**95.5**	**96.9**	**97.8**	**95.7**	**95.3**	**86.5**	**97.0**	**94.3**	**98.5**	**53.4**	**77.6**	**84.6**

Table 2. The ASR (%) of Ens, SVRE, AdaEA and SMER attacking CNNs and ViTs. The adversarial examples are generated on four ensemble surrogates: Res-18, Inc-v3, ViT-T and DeiT-T, with four attack baselines respectively.

Base	Attack	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens}$	Average
TIM	Ens [8]	1.6	1.7	1.8	1.7
	SVRE [45]	4.4	4.7	3.0	4.0
	AdaEA [6]	1.3	0.8	0.8	1.0
	SMER	**19.8**	**15.2**	**12.4**	**15.8**
TI-DIM	Ens [8]	2.4	2.4	2.0	2.3
	SVRE [45]	5.0	5.3	5.1	5.1
	AdaEA [6]	3.3	2.4	2.1	2.6
	SMER	**31.2**	**26.6**	**24.1**	**27.3**

Table 3. The ASR (%) of targeted attack of Ens, SVRE, AdaEA and SMER. The adversarial examples are generated on four ensemble surrogates: Inc-v3, IncRes-v2, Res-101 and Inc-v4, with four attack baselines respectively.

step size β and the internal decay factor are both the same as the outer step size and decay factor respectively. SGD optimizer with learning rate $lr = 0.02$ is employed for the optimization of ensemble weights.

4.2. Attacks based on Homogeneous CNN Ensemble

The vulnerability of conventional recognition systems to adversarial attacks is well-documented, hence we evaluate the effectiveness of our SMER on various defensive methods. There are two groups of homogeneous tests, $i.e.$ adversarial training and transformation-based method, and one group of heterogeneous tests, $i.e.$ robust vision transformers in this section. All adversarial examples are crafted on four commonly trained surrogates, $i.e.$, Inc-v3, IncRes-v2, Res-101 and Inc-v4. Additionally, we present the results of the homogeneous ViT ensemble in Section 4.5.

Adversarial Training. The adversarial examples generated by SMER from commonly trained surrogates against adversarial-trained victims outperform those from the other three methods for each basic attack, as shown in Table 1. Particularly noteworthy is the significant ASR increment observed with I-FGSM, averaging about 26.8%, 15.2%, and 20.6% compared to Ens, SVRE, and AdaEA, respectively.

Transformation-based Method. Input transformation aims to mitigate the impact of adversarial artifacts and preserve the original features. It is evident that SMER exhibits similar superiority when attacking transformation-based defenses, as shown in Table 1. Despite fluctuations across different evaluations, SMER achieves the largest average increment, reaching 27.8% on I-FGSM compared to Ens, while 16.1% and 17.6% increments are observed on MI-FGSM compared to SVRE and AdaEA, respectively. Notably, the most aggressive attack occurs with TI-DIM, consistent with previous experiments, indicating that SMER is capable of leveraging advanced attacks rather than compromising them.

Robust ViT. ViT has emerged as the new generation for classification, with robust vision transformers enhancing their corresponding robustness. In this section, we employ robust ViTs [26–28] to evaluate the transferability of heterogeneous attacks. It is observed that when integrated with TI-DIM, our approach achieves superior competitive ASR averages (up to 83.3%) compared to the other three methods in Table 1. SMER still outperforms other approaches on less advanced baselines by a remarkable margin. This suggests that SMER is capable of generating adversarial examples with higher transferability from CNNs to ViTs.

From the results above, it is evident that our SMER achieves superior performance in both homogeneous and heterogeneous attacks compared to other methods. This is primarily attributed to our approach leveraging diverse

Ensemble Type	Ensemble Models	Attack	CNN				ViT			
			Res-50	WRN50-2	Bit-101	Average	ViT-B	DeiT-B	Swin-B	Average
Only CNNs	Res-18	AdaEA	81.4	68.0	80.9	76.8	35.0	45.0	36.7	38.9
	Inc-v3, BiT-50	SMER	**90.6**	**79.8**	**92.1**	**87.5**	**43.4**	**49.4**	**39.3**	**44.0**
	Res-18, Inc-v3	AdaEA	83.1	74.6	79.0	78.9	40.5	51.6	36.7	42.9
	BiT-50, Inc-v4	SMER	**93.6**	**86.8**	**93.9**	**91.4**	**48.9**	**56.2**	**49.8**	**51.6**
Only ViTs	ViT-T	AdaEA	62.0	47.7	66.3	58.7	78.2	92.5	71.4	80.7
	DeiT-T, Swin-T	SMER	**77.5**	**60.1**	**75.1**	**70.9**	**89.9**	**97.1**	**93.7**	**93.6**
	ViT-T, Swin-T	AdaEA	66.3	50.0	69.4	61.9	93.6	94.5	75.8	88.0
	DeiT-T, ViT-S	SMER	**84.5**	**70.9**	**84.5**	**80.0**	**98.8**	**99.2**	**96.9**	**98.3**
Mix	Res-18	AdaEA	80.2	63.2	75.5	73.0	62.1	72.1	46.4	60.2
	ViT-T, Inc-v3	SMER	**88.5**	**75.0**	**82.1**	**81.9**	**75.3**	**80.1**	**52.0**	**69.1**
	Inc-v3	AdaEA	69.6	55.5	67.7	64.3	65.9	77.7	71.6	71.7
	ViT-T, Swin-T	SMER	**77.2**	**63.7**	**73.0**	**71.3**	**79.0**	**86.7**	**91.3**	**85.7**
	Res-18, Inc-v3	AdaEA	79.8	64.0	78.7	74.2	74.8	85.6	54.4	71.6
	ViT-T, DeiT-T	SMER	**91.6**	**78.8**	**87.0**	**85.8**	**88.0**	**95.6**	**66.7**	**83.4**

Table 4. The ASR (%) of AdaEA and SMER on three ensemble types, *i.e.* only CNNs, only ViTs and Mixed. The adversarial examples are generated on MI-FGSM baselines to achieve black-box attacks.

strengths from each surrogate, which facilitates optimal direction search. Additionally, the reweighing operation is delegated to the surrogates themselves, enabling accurate weight adjustments.

4.3. Attacks based on Heterogeneous Ensemble

The back-propagated gradients of CNNs and ViTs differ [6, 45], and we incorporate ViTs to assess adversarial generation in a heterogeneous ensemble, *i.e.* the Mix ensemble, to further evaluate the ability of our SMER.

The performance clearly demonstrates that the ASR of our algorithm surpasses the other three approaches on both CNNs and ViTs, as shown in Table 2. This contribution mainly stems from the unconstrained diversity in perturbations, which circumvents the challenges of mitigating discrepancies and maximizes the attack effectiveness for the current model at each perturbation step. Consequently, SMER is capable of generating highly transferable perturbations under a mixed ensemble.

4.4. Targeted Attacks

The evaluation for targeted attacks can also demonstrate versatility to some extent, though the three compared methods were not initially designed for this purpose. Targeted attacks are typically described as follows: $\hat{x}_{t+1} = \hat{x}_t - \alpha \cdot sign(\nabla_x J(f(\hat{x}_t), \hat{y}))$, where $\hat{y}$ denotes the targeted label. SMER exhibits remarkable performance on the advanced baseline TI-DIM, surpassing the other three counterparts by 25.0%, 22.2%, and 24.7%, as shown in Table 3. These results indicate that SMER can generate highly transferable perturbations for the targeted label without additional modification. It also suggests that mitigating diversity is incompatible with targeted attacks.

4.5. Further Analysis

Ensemble types. We further demonstrate our superior performance of attacks on different ensemble types. SMER outperforms counterparts on both homogeneous ensembles, i.e., only CNNs and only ViTs, and heterogeneous ensembles, i.e., the Mix, as shown in Table 4. Specifically, SMER integrated with only CNNs achieves the highest average ASR of 91.4%, which is 12.5% higher than AdaEA on CNN victims. Meanwhile, SMER achieves an average ASR of 98.3% against ViT victims when integrated with only ViTs, which is 10.3% higher than AdaEA. Additionally, SMER outperforms AdaEA by up to 12.5%, 18.1%, and 13.9% across the three ensemble types, respectively. The advantage of SMER lies in its emphasis on perturbation diversity, which avoids the challenges of constrained convergence. Interestingly, the general trend indicates that as the ensemble number increases, the ASR increment of SMER compared to AdaEA is higher. This is likely because more ensemble surrogates imply greater diversity, which positively impacts the black-box transferability of SMER.

Attention alteration. We introduce feature attention region alteration [32], quantified by the Structural Similarity Index Measure (SSIM), to assess the ability of attacks to impact the classifiers, as shown in Figure 3. A larger SSIM gap between the first and the last attention map indicates a larger difference. The numerical results show that the first iteratively generated examples share similar attention regions compared to the original attention maps in almost every

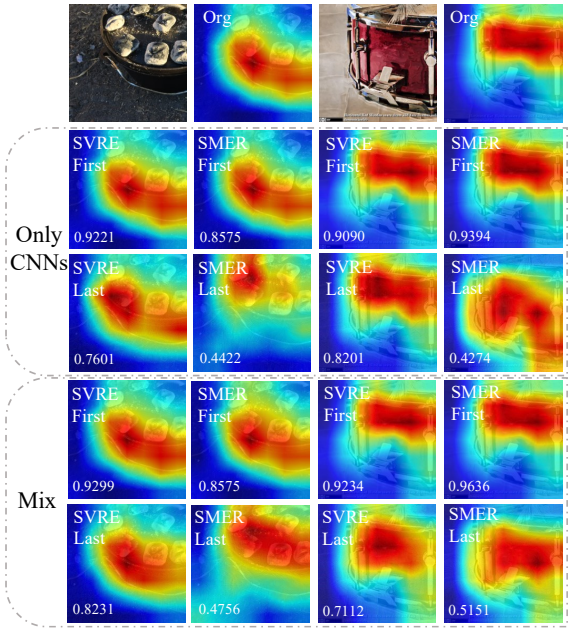

Figure 3. The hot maps of feature attention altering from the first perturbed images to the last perturbed images for SMER and SVRE with SSIM comparison. The redder region indicates more attention. Larger SSIM change indicates larger alteration in numerical value.

comparison, while the last output examples exhibit significant variation. The SSIM gaps of SMER exceed those of SVRE, indicating larger map alterations.

Discrepancy. We quantify the discrepancy by computing the reciprocal of the cosine similarity between the gradients generated by each method at each iteration and the average gradients from the entire ensemble for adversarial examples obtained at the same iteration, where a larger discrepancy indicates a larger difference from the ensemble average in Figure 4 (a). SMER witnesses a striking increase compared with the other two methods despite the initial decline. Interestingly, the discrepancy of the compared algorithms continuously rises while they propose to decrease it. It could be ascribed to the toughness of convergence under the bound and the limited steps.

Average loss. While ASR typically indicates attack performance, the average loss can also reflect the transferability of adversarial examples. In Figure 4 (b), SMER consistently shows a higher average loss compared to other approaches in each comparison, suggesting its ability to produce highly transferable adversaries.

Ablation study. To elucidate the effectiveness of each component in SMER, we conduct corresponding studies using adversarial examples from the ensemble employed in section 4.2. The results are averaged from Inc-v3$_{ens3}$, Inc-v3$_{ens4}$, and IncRes-v2$_{ens}$. In Figure 5 (a), we observe that

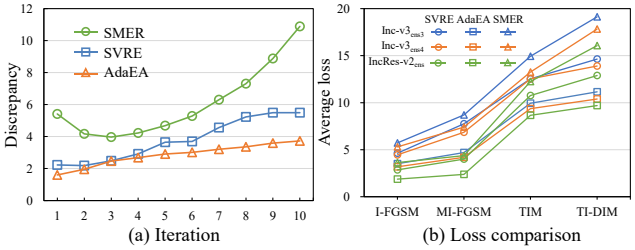

Figure 4. (a) The discrepancy between iterative perturbations and the corresponding ensemble average perturbations. (b) The average loss of three methods against three adversarial-trained models.

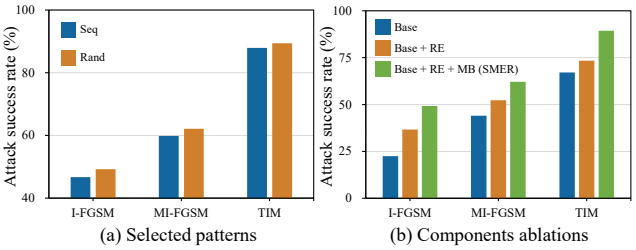

Figure 5. (a) The ASR of SMER with different selected patterns. (b) The ablation study of two components in SMER.

the efficacy of the random pattern (Rand) in SMER leads to about a 5% improvement compared to the sequential pattern (Seq). Figure 5 (b) demonstrates that the collaboration of reweighing (RE) and mini-batch (MB) operations can gradually increase the ASR. The base is the typical attack of the ensemble on logits. It is notable that the reweighing operation substitutes constant weights, which formerly compromised the diversity of ensemble surrogates contributing to the perturbation, and instead allows the surrogates to customize individual weights. Meanwhile, the mini-batch operation avoids fusing the devised customized outputs from the diverse ensemble, instead promoting individual influence, which further enhances customization.

5. Conclusion

In this paper, we introduce a novel method called SMER, which prioritizes diversity between models and generates adversarial examples by individual models with ensemble reweighing using reinforcement learning to customize model diversity. This approach aims to maximize attack loss and identify the optimal direction. Extensive experimental results demonstrate that SMER not only surpasses other methods with notable superiority across various black-box scenarios but also exhibits significant aggressiveness when integrated with ensemble surrogates featuring high discrepancy. In summary, SMER enhances the transferability of adversarial examples across both homogeneous and heterogeneous ensembles, effectively addressing both homogeneous and heterogeneous attacks.

References

[1] Naveed Akhtar, Mohammad A. A. K. Jalwana, Mohammed Bennamoun, and Ajmal Mian. Attack to fool and explain deep networks. *IEEE TPAMI*, 44(10):5980–5995, 2022. 1

[2] K Alex, Hamner Ben, and Goodfellow Ian. Nips 2017: Defense against adversarial attack, 2017. 5

[3] Dosovitskiy Alexey, Beyer Lucas, Kolesnikov Alexander, Weissenborn Dirk, Zhai Xiaohua, Unterthiner Thomas, Dehghani Mostafa, Minderer Matthias, Heigold Georg, Gelly Sylvain, Uszkoreit Jakob, and Houlsby Neil. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 1, 3, 5

[4] Zikui Cai, Xinxin Xie, Shasha Li, Mingjun Yin, Chengyu Song, Srikanth V. Krishnamurthy, Amit K. Roy-Chowdhury, and M. Salman Asif. Context-aware transfer attacks for object detection. In *AAAI*, 2022. 1

[5] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy (SP)*, 2017. 1

[6] Bin Chen, Jiali Yin, Shukai Chen, Bohao Chen, and Ximeng Liu. An adaptive model ensemble adversarial attack for boosting adversarial transferability. In *ICCV*, 2023. 1, 2, 3, 4, 5, 6, 7

[7] Szegedy Christian, Zaremba Wojciech, Sutskever Ilya, Bruna Joan, Erhan Dumitru, Goodfellow Ian, and Fergus Rob. Intriguing properties of neural networks. In *ICLR*, 2014. 3

[8] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *CVPR*, 2018. 1, 3, 4, 5, 6

[9] Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *CVPR*, 2019. 3, 5

[10] Lianli Gao, Qilong Zhang, Jingkuan Song, Xianglong Liu, and Heng Tao Shen. Patch-wise attack for fooling deep neural network. In *ECCV*, 2020. 1

[11] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and Harnessing Adversarial Examples. In *ICLR*, 2015. 1, 2, 3, 5

[12] Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens van der Maaten. Countering Adversarial Images using Input Transformations. In *ICLR*, 2018. 3, 5

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *ECCV*, 2016. 5

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5

[15] Xiaojun Jia, Xingxing Wei, Xiaochun Cao, and Hassan Foroosh. Comdefend: An efficient image compression model to defend adversarial examples. In *CVPR*, 2019. 5

[16] Xiaojun Jia, Yong Zhang, Baoyuan Wu, Ke Ma, Jue Wang, and Xiaochun Cao. Las-at: Adversarial training with learnable attack strategy. In *CVPR*, 2022. 1

[17] Jia Jinyuan, Cao Xiaoyu, Wang Binghui, and Gong Neil, Zhenqiang. Certified robustness for top-k predictions against adversarial perturbations via randomized smoothing. In *ICLR*, 2020. 3

[18] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *NeurIPS*, 2013. 2, 3

[19] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. In *ECCV*, 2020. 5

[20] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world, 2017. 1, 3

[21] Jiadong Lin, Chuanbiao Song, Kun He, Liwei Wang, and John E Hopcroft. Nesterov accelerated gradient and scale invariance for adversarial attaks. In *ICLR*, 2020. 1, 3, 5

[22] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into Transferable Adversarial Examples and Black-box Attacks. In *ICLR*, 2017. 1, 3

[23] Zihao Liu, Qi Liu, Tao Liu, Nuo Xu, Xue Lin, Yanzhi Wang, and Wujie Wen. Feature distillation: Dnn-oriented jpeg compression against adversarial examples. In *CVPR*, 2019. 3, 5

[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 *ICCV*, 2021. 5

[25] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards Deep Learning Models Resistant to Adversarial Attacks. In *ICLR*, 2018. 3

[26] Chengzhi Mao, Lu Jiang, Mostafa Dehghani, Carl Vondrick, Rahul Sukthankar, and Irfan Essa. Discrete representations strengthen vision transformer robustness. In *ICLR*. 1, 3, 5, 6

[27] Xiaofeng Mao, YueFeng Chen, Ranjie Duan, Yao Zhu, Gege Qi, Shaokai Ye, Xiaodan Li, Rong Zhang, et al. Enhance the visual representation via discrete adversarial training. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 3, 5

[28] Xiaofeng Mao, Gege Qi, Yuefeng Chen, Xiaodan Li, Ranjie Duan, Shaokai Ye, Yuan He, and Hui Xue. Towards robust vision transformer. In *CVPR*, 2022. 1, 3, 5, 6

[29] Muzammal Naseer, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Fatih Porikli. A self-supervised approach for adversarial robustness. In *CVPR*, 2020. 3

[30] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015. 5

[31] Simon Schrodi, Tonmoy Saikia, and Thomas Brox. Towards understanding adversarial robustness of optical flow networks. In *CVPR*, 2022. 1

[32] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *IJCV*, 128(2):336–359, 2020. 7

[33] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016. 5

[34] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*, 2017. 5

[35] 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 (ICML)*, 2021. 5

[36] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble Adversarial Training: Attacks and Defenses. In *ICLR*, 2018. 1, 3, 5

[37] Xiaosen Wang, Xuanran He, Jingdong Wang, and Kun He. Admix: Enhancing the transferability of adversarial attacks. In *ICCV*, 2021. 1

[38] Zhibo Wang, Hengchang Guo, Zhifei Zhang, Wenxin Liu, Zhan Qin, and Kui Ren. Feature importance-aware transferable adversarial attacks. In *ICCV*, 2021. 1

[39] Zheng Wang, Zhenwei Gao, Kangshuai Guo, Yang Yang, Xiaoming Wang, and Heng Tao Shen. Multilateral semantic relations modeling for image text retrieval. In *CVPR*, 2023. 1

[40] Zheng Wang, Zhenwei Gao, Guoqing Wang, Yang Yang, and Heng Tao Shen. Visual embedding augmentation in fourier domain for deep metric learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(10):5538–5548, 2023.

[41] Zheng Wang, Xing Xu, Guoqing Wang, Yang Yang, and Heng Tao Shen. Quaternion relation embedding for scene graph generation. *IEEE Transactions on Multimedia*, 25: 8646–8656, 2023. 1

[42] Zhipeng Wei, Jingjing Chen, Micah Goldblum, Zuxuan Wu, Tom Goldstein, and Yu-Gang Jiang. Towards Transferable Adversarial Attacks on Vision Transformers. In *AAAI*, 2022. 1

[43] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan Yuille. Mitigating adversarial effects through randomization. In *ICLR*, 2018. 5

[44] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan Yuille. Improving transferability of adversarial examples with input diversity. In *CVPR*, 2019. 3

[45] Yifeng Xiong, Jiadong Lin, Min Zhang, John E. Hopcroft, and Kun He. Stochastic variance reduced ensemble adversarial attack for boosting the adversarial transferability. In *CVPR*, 2022. 2, 3, 4, 5, 6, 7

[46] Weilin Xu, David Evans, and Yanjun Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. In *Network and Distributed System Security Symposium (NDSS)*, 2018. 3, 5

[47] Zheng Yuan, Jie Zhang, Yunpei Jia, Chuanqi Tan, Tao Xue, and Shiguang Shan. Meta gradient adversarial attack. In *ICCV*, 2021. 3

[48] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *BMVC*, 2016. 5

[49] Jianping Zhang, Weibin Wu, Jen-tse Huang, Yizhan Huang, Wenxuan Wang, Yuxin Su, and Michael R. Lyu. Improving adversarial transferability via neuron attribution-based attacks. In *CVPR*, 2022. 1