This CVPR Workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Enhancing the Transferability of Adversarial Attacks with Stealth Preservation

Xinwei Zhang[1,2], Tianyuan Zhang[1,2,3], Yitong Zhang[1], Shuangcheng Liu[1,2,†]
[1] School of Computer Science and Engineering, Beihang University, Beijing, China
[2] State Key Lab of Software Development Environment, Beihang University, Beijing, China
[3] Shen Yuan Honors College, Beihang University, Beijing, China
{xinweizhang, zhangtianyuan, 22373337, 93777}@buaa.edu.cn

Abstract

*Deep neural networks are susceptible to attacks from adversarial examples in recent years. Especially, the black-box attacks cause a more serious threat to practical applications. However, while most existing black-box attacks have achieved a high success rate in deceiving models, they have not focused on the stealthiness of adversarial examples, often exhibiting suspicious visual appearances. To address this issue, this paper proposes the Mask Momentum Iterative Attack (*MMIA*), which introduces a masking mechanism and adopts an optimal perturbation strategy to identify regions of an image most vulnerable to attacks. This approach effectively ensures the transferability and stealthiness of adversarial examples. Simultaneously, by integrating image enhancement techniques and temporal and spatial momentum terms into the iterative process of the attack, we prevent the attack from getting stuck in local optima, further improving the transferability of adversarial examples. To enhance the success rate of black-box attacks, we apply* MMIA *to a model ensemble using a joint optimization strategy. We demonstrate that adversarially trained models with a strong defense ability are also susceptible to our black-box attacks. We conduct extensive experiments on classification tasks using common vision models, and our results significantly demonstrate the superiority of our method over state-of-the-art approaches when considering both transferability and stealthiness.*

1. Introduction

As Deep Neural Networks (DNNs) continue to excel in a broad spectrum of applications, including computer vision [24], natural language processing [46], and acoustics [41], the challenges to their security, *e.g.*, adversarial attacks [16, 30, 31, 33, 38, 47, 54] and the infiltration of backdoor trojans [9, 17, 37], are gradually being unveiled. Adversarial examples, crafted by subtly introducing imperceptible noise into clean instances, pose a formidable security challenge by easily inducing erroneous predictions from DNNs. This phenomenon represents a potent threat to the performance of deep learning applications.

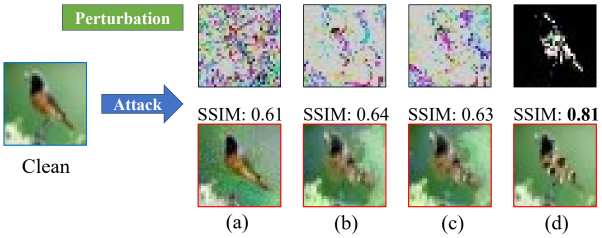


Figure 1. We show four types of iterative algorithms to generate adversarial examples on the CIFAR-10 dataset with the ResNet-50 model. **First row**: the perturbations generated by each adversarial algorithm. **Second row**: the corresponding adversarial examples. (a) PGD. (b) MI-FGSM. (c) SMI-FGSM. (d) **MMIA**. Our method specifically generates perturbation in the most sensitive regions for classification model, reaching a high SSIM score.

In the past few years, a series of works have been proposed to conduct adversarial attacks under different conditions and settings [1, 14, 28]. Generally, adversarial attacks can be categorized into white-box attacks and black-box attacks. In white-box attacks, the attacker has complete access to the target model, including its structure, parameters, and training data. This allows the attacker to leverage such information to generate more effective adversarial examples, including gradient-based attacks [16], optimization-based attacks [4], *etc*. In black-box attacks, the attacker can only observe and interact with the target model in a limited manner, possessing little or no knowledge about the specific details of the target model, including query-based attacks [7], transfer-based attacks [14], *etc*. In this paper, we primarily focus on the more challenging black-box attack, which is more relevant for the practical deployment of deep learning applications.

[†] Corresponding author.

Though several attempts have been adopted to perform black-box attacks, existing works often overlook the factor of maintaining image structure stability [14, 15, 52, 56], *e.g.* the stealthiness of black-box attacks. Consequently, their practical applications are far from satisfactory. It is worth noting that gradient-based attacks have shown promising results by focusing on improving the transferability of adversarial examples, thereby achieving success in black-box attacks. There is an interesting observation that in a white-box environment, iterative attacks outperform single-step attacks, whereas in a black-box environment, the opposite is true [26]. Therefore, to generate adversarial examples with high success rates in both white-box and black-box settings, most researchers currently adopt the strategy of designing algorithms based on iterative attacks. For instance, MI-FGSM [14] introduces momentum to prevent iterations from falling into local peaks, while SMI-FGSM [52] utilizes spatial momentum to enhance the transferability of adversarial examples. However, they often neglect the factor of stealthiness. After multiple iterations of attacking the entire image, the image content may suffer considerable damage and distortion, as shown in Figure1 (a), (b) and (c).

To address the aforementioned issues, this paper proposes the Mask Momentum Iterative Attack (*MMIA*), which introduces a masking mechanism and spatial momentum at each iteration step. Inspired by attention mechanisms that models focus more on salient regions when classifying images [10], we design an optimal perturbation strategy to identify the areas in the image most susceptible to attacks and generate masks. Simultaneously, to ensure the transferability of adversarial examples, we create diverse input patterns by applying transformations [20, 56] such as resizing, cropping, and rotation to prevent network overfitting. Moreover, throughout the iterative process of the attack, we maintain temporal [14] and spatial [52] momentum term continuously. This prevents the attack from getting stuck in local optima, further enhancing the transferability of adversarial examples. We demonstrate that the adversarial examples generated by *MMIA* achieve high success rates in both white-box and black-box attacks, while maintaining high SSIM [55] scores. This ensures that the images of adversarial examples generated by *MMIA* are closer to clean samples, as shown in Figure1 (d).

To further enhance the transferability of adversarial examples, we propose a multi-model joint optimization strategy to apply *MMIA* to model ensemble. Existing research indicates that if an adversarial example deceives multiple models, it is more likely to maintain adversarial characteristics against other black-box models [39, 51]. Specifically, we fuse gradient information from various models at each iteration step and combine it with a maske mechanism to generate more robust perturbations. We demonstrate that adversarial examples generated by *MMIA* under the multi-model joint optimization strategy can successfully deceive robust models obtained through ensemble adversarial training in a black-box manner.

To the best of our knowledge, we are the first to introduce the maske mechanism to ensure the stealthiness of adversarial images and incorporate spatial momentum term and fusing gradients from multiple models to generate more robust black-box adversarial samples. Extensive experiments were conducted on CIFAR-10 and ImageNet datasets, covering a variety of common visual models and adversarial models for classification tasks. Our results demonstrate that our approach not only achieves high attack effectiveness but also maintains elevated SSIM score, showcasing its effectiveness across different datasets and models.

2. Releated Works

2.1. Adversarial examples

Adversarial examples are specially designed samples that possess features not easily perceivable by humans but can lead to incorrect prediction by DNNs. In recent years, a series of studies on adversarial attacks have been proposed [22, 29, 32, 34–36, 50, 53, 57, 58]. Adversarial attack methods are typically classified into several categories. For example, based on whether there is a specified target for the attack, they can be classified as targeted attacks or non-targeted attacks. Based on the scope of the attack, adversarial attacks can be divided into digital world attacks and physical world attacks. Based on whether the attacker has sufficient knowledge of the model being attacked, they can be classified as white-box attacks or black-box attacks, *etc*.

White-box attacks refer to attacks where the attacker has complete internal information and access rights, allowing them to fully understand the structure, parameters, algorithms, and training data of the target model. Szegedy *et al*. [47] first proposed the concept of adversarial examples and used the L-BFGS method for generation. By utilizing the gradient information of the target model, Goodfellow *et al*. [16] proposed a fast gradient algorithm called the Fast Gradient Sign Method (FGSM) for rapidly generating adversarial examples. Building upon FGSM, Kurakin *et al*. [27] designed iterative versions called the Basic Iterative Method (BIM) and the Iterative Least-Likely Class Method (ILCM). Furthermore, Madry *et al*. [40] extended the concept of clipping to a projection process and added random perturbation during initialization, creating the Projective Gradient Descent (PGD) attack method. It is considered the strongest first-order attack method. Although the aforementioned methods have achieved significant success in white-box attacks, there has been a substantial decrease in performance when conducting black-box attacks.

Black-box attacks refer to situations where the attacker lacks complete internal information and access rights to the target model, and can only observe the model's behavior through inputs and outputs. They can be divided into two categories: 1) query-based attacks, 2) transfer-based attacks. Query-based attacks estimate gradients based on confidence scores [8, 12, 42] or decision [2, 6, 11] information output by the black-box model in order to conduct attacks. However, this type of attack still requires access to the target model, albeit limited to querying its outputs. Transfer-based attacks stem from the transferability of adversarial examples [47]. They conduct white-box attacks on substitute models and then transfer these adversarial examples to the target model to achieve the attack effect. Therefore, transfer-based attacks pose a higher threat in the real world. To enhance transferability, Dong *et al.* [14] improved the BIM method based on the idea of momentum, proposing the Momentum Iterative Fast Gradient Sign Method (MI-FGSM). Furthermore, Xie *et al.* [56] proposed the Diverse Inputs Iterative Fast Gradient Sign Method (DI^2-FGSM), which generates adversarial examples by creating diverse input patterns, further enhancing transferability. Based on translation invariance, Dong *et al.* [15] proposed the Translation-Invariant Fast Gradient Sign Method (TI-FGSM). To stabilize the direction of gradient updates, Wang *et al.* [52] introduced the spatial domain gradient in images, proposing the Spatial Momentum Iterative Fast Gradient Sign Method (SMI-FGSM). Although these methods have made significant improvements in black-box transfer attacks, they still focus on perturbing the entire image, leading to a decrease in the stealth of the image.

2.2. Gradient-based Attack Methods

Several gradient-based methods have been proposed to generate the adversarial examples. In this section, we provide a brief review of them.

Fast Gradient Sign Method (FGSM) [16] generate an adversarial example by performing a single-step update that increases the model's loss for the given image $\boldsymbol{x}$:

$$\boldsymbol{x}^{adv} = \boldsymbol{x} + \epsilon \cdot sign(\nabla_{\boldsymbol{X}} J(\boldsymbol{x}^{adv}, y)), \tag{1}$$

where $\nabla_{\boldsymbol{X}} J$ the gradient of the loss function with respect to the image $\boldsymbol{x}$ and $sign(\cdot)$ is the sign function ensuring that the generated perturbation satisfies the sign of the L_∞ norm distace.

Projected Gradient descent (PGD) [40] is a variant of the Iterative version of FGSM(I-FGSM), initializing with uniformly random noise, and stands out as one of the most powerful first-order attack methods. It iteratively applies gradient updates with a small step size α, projecting the perturbation into a specified range at each iteration.

$$\boldsymbol{x}_{t+1}^{adv} = \Pi_\epsilon(\boldsymbol{x}_t^{adv} + \alpha \cdot sign(\nabla_{\boldsymbol{x}} J(\boldsymbol{x}_t^{adv}, y))), \tag{2}$$

where $\boldsymbol{x}_0^{adv} = \boldsymbol{x}$, and Π_ϵ is a clip function that projects perturbation into the specified range ϵ when exceed during the number of iterations T.

Momentum Iterative Fast Gradient Sign Method (MI-FGSM) [14] stabilizes the update direction in iterative attacks by incorporating a temporal momentum term, preventing convergence to local optima and enhancing the transferability of adversarial examples.

$$g_{t+1} = \mu \cdot g_t + \frac{\nabla_{\boldsymbol{x}} J(\boldsymbol{x}_t^{adv}, y)}{\|\nabla_{\boldsymbol{x}} J(\boldsymbol{x}_t^{adv}, y)\|_1}, \tag{3}$$

$$\boldsymbol{x}_{t+1}^{adv} = \boldsymbol{x}_t^{adv} + \alpha \cdot sign(g_{t+1}), \tag{4}$$

where g_t is the accumulated gradient up to the t-th and μ is the decay factor.

Spatial Momentum Iterative Fast Gradient Sign Method (SMI-FGSM) [52] introduces the spatial momentum iteration term in iterative attacks to stabilize the update direction. It integrates multiple gradients from random transformations of the same image, utilizing information from the contextual region to generate a stable gradient. The formulation is as follows:

$$g_{t+1}^s = \sum_{i=1}^{N} \lambda_i \nabla_{\boldsymbol{x}} J(H_i(\boldsymbol{x}_t^{adv}), y), \tag{5}$$

$$\boldsymbol{x}_{t+1}^{adv} = \boldsymbol{x}_t^{adv} + \alpha \cdot sign(g_{t+1}^s), \tag{6}$$

where $H_i(\cdot)$ transforms $\boldsymbol{x}_t^{adv}$ by adding random padding around the image and resizing it to the original size.

3. Approach

In this paper, we propose a Mask Momentum Iterative Algorithm (*MMIA*), which can generate adversarial examples deceiving both white-box and black-box models with minimal disruption to the image. In this section, we initially articulate the problem definition and then provide a detailed overview of the proposed algorithm framework, as illustrated in Figure 2. We first elucidate the optimal perturbation strategy by introducing a masking mechanism to perturb the most sensitive regions of the image at each iteration step. Subsequently, we expound on the proposed *MMIA* algorithm, enhancing the transferability of adversarial examples by integrating image augmentation techniques, as well as the momentum term into the iterative process of the attack. Finally, we employ a joint optimization approach for model ensembles, leveraging attack information from multiple models to update perturbations effectively. The adversarial examples discussed in this paper satisfy the L_∞ norm restriction in the non-targeted attack fashion.

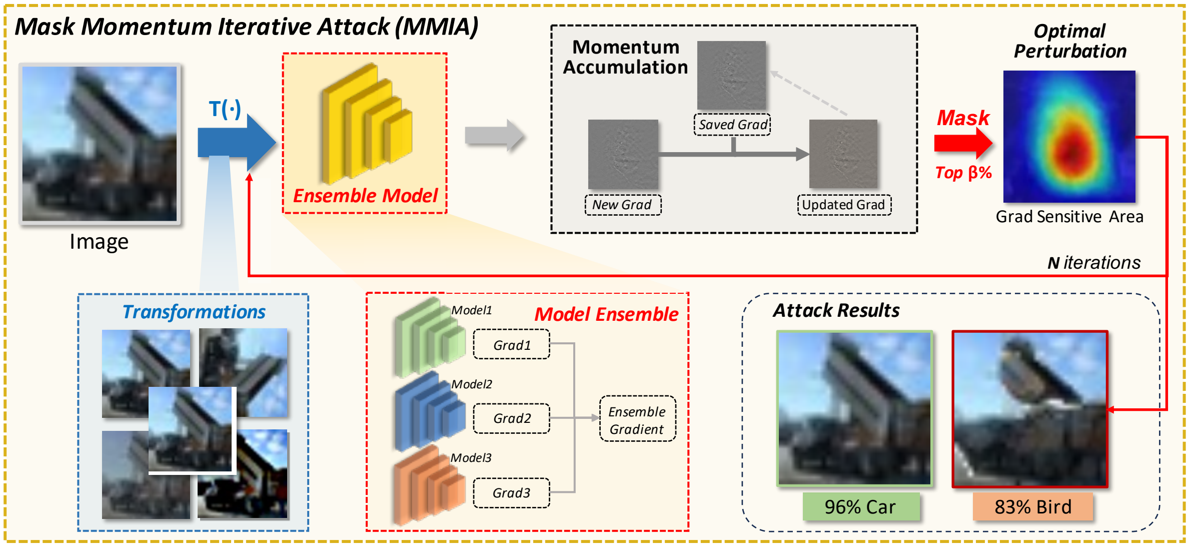


Figure 2. The framework of our MMIA method. We employ an optimal perturbation strategy, identifying the most sensitive regions of the image for perturbation at each iteration step through the introduction of a maske mechanism. Then, we enhance the transferability by integrating image enhancement techniques, along with temporal and spatial momentum terms into the iterative process of the attack. Simultaneously, we propose a model ensemble joint optimization approach, making full use of attack information from multiple models.

3.1. Problem Definitions

Given a deep neural network classifier $\mathcal{F}$ and an input image $\boldsymbol{x}$ with a true label y, an adversarial sample $\boldsymbol{x}^{adv}$ can lead to a wrong prediction by the model *i.e.* $\mathcal{F}(\boldsymbol{x}) \neq y$. For adversarial example generation, the objective is to maximize the loss function $J(\boldsymbol{x}^{adv}, y)$ of the classifier $\mathcal{F}$. Therefore, the constrained optimization problem can be expressed as:

$$\arg\max_{\boldsymbol{x}_{adv}} J(\boldsymbol{x}^{adv}, y) \quad s.t.\ eqqqqqqqqqqqqq\ \ \|\boldsymbol{x}-\boldsymbol{x}^{adv}\|_\infty < \epsilon, \tag{7}$$

where ϵ is the size of adversarial perturbation and $\|\cdot\|_\infty$ is the distance metric used to quantify the distance between two inputs $\boldsymbol{x}$ and $\boldsymbol{x}^{adv}$ under the constraint of the L_∞ norm sufficiently small.

Meanwhile, we choose **SSIM** (Structural Similarity Index) to measure the similarity between two images, considering brightness, contrast, and structure. The SSIM value ranges from 0 to 1, with higher values indicating greater similarity between images. Its calculation is as follows:

$$\operatorname{SSIM}(\boldsymbol{x}, \boldsymbol{x}_{adv}) = \frac{(2\mu_{\boldsymbol{x}}\mu_{\boldsymbol{x}_{adv}} + c_1)(2\sigma_{\boldsymbol{x}\boldsymbol{x}_{adv}} + c_2)}{(\mu_{\boldsymbol{x}}^2 + \mu_{\boldsymbol{x}_{adv}}^2 + c_1)(\sigma_{\boldsymbol{x}}^2 + \sigma_{\boldsymbol{x}_{adv}}^2 + c_2)}, \tag{8}$$

where μ and σ are the mean and standard deviation of the image, and c_1 and c_2 are constants.

The objective is to obtain adversarial samples that successfully deceive the model while maintaining a high level of concealment, *i.e.*, a high structural similarity. The problem can be defined as follows:

$$\arg\max_{\boldsymbol{x}_{adv}} J(\boldsymbol{x}^{adv}, y) + \operatorname{SSIM}(\boldsymbol{x}, \boldsymbol{x}_{adv}),\ s.t.\ \ \|\boldsymbol{x}-\boldsymbol{x}^{adv}\|_\infty < \epsilon, \tag{9}$$

3.2. Mask Momentum Iter Attack

Although some methods have made breakthroughs in the transferability of black-box attacks, achieving high attack success rates, they often neglect the property of concealment when generating adversarial examples. The obtained adversarial examples frequently exhibit significant differences from the original images, i.e., visible distortions, posing obstacles to practical applications.

To address this issue, we propose a *MMIA* (Masked Momentum Iterative Method), outlined in the framework shown in Figure 2. Inspired by the varying focus of CNNs on different regions in an image, as demonstrated by methods like GradCAM, we introduce a masking technique into each iteration step. We present an optimal perturbation strategy aimed at perturbing only the most sensitive regions of the classifier during each optimization step, rather than disturbing the entire image. This approach aims to maintain the structural similarity of the adversarial image consistently. In addition, we have also considered image enhancement techniques and temporal and spatial momentum techniques to enhance the effectiveness of the attack.

```
Algorithm 1 MMIA Algorithm
Input: Classifier $f$ with loss function $J$, real example $\boldsymbol{x}$, ground-truth label $y$, outer iterations $t$, inner iterations $n$, transformation probability $p$, perturbation size $\beta$, max perturbation $\varepsilon$, perturbation step size $\alpha$, decay factor $\mu$;
Output: Adversarial example $\boldsymbol{x}^*$;
Init transformations function $T(\cdot)$;
$Grad_0 = 0$; $\boldsymbol{x}^* = \boldsymbol{x}$;
for $i = 0$ to $t - 1$ do
    for $j = 0$ to $n - 1$ do
        $\boldsymbol{x}^*_{ij} = T(\boldsymbol{x}^*_i)$ with the probability $p$
        Get gradient $G_j$ of $\boldsymbol{x}^*_{ij}$ by Eq 10.
    end for
    Obtain spatial momentum gradient $g^s_i$ by
        $g^s_i = \frac{1}{n}\sum G_j$
    Update $Grad_{i+1}$ by
        $Grad_{i+1} = \mu Grad_i + \frac{g^s_i}{\|g^s_i\|_1}$
    $\boldsymbol{Mask} = SearchMask(Grad_{i+1}, \beta)$;
    Update $\boldsymbol{x}^*_{i+1}$ by Eq 13.
    Clip $\boldsymbol{x}^*_{i+1}$ according to $\varepsilon$;
end for
return $\boldsymbol{x}^* = \boldsymbol{x}^*_t$
```

3.2.1 Optimal Perturbation Strategy

To seek the perturbation region of an image, which involves finding a mask to restrict the pixel area to be modified during the optimization steps. In each iteration, we first utilize the following formula for backpropagation to obtain the gradient information for the entire image:

$$\boldsymbol{Grads} = \nabla_{\boldsymbol{x}} J(\boldsymbol{x}^{adv}_t, y), \tag{10}$$

This gradient information contains gradients for the RGB channels, and we sum their absolute values to merge them into a two-dimensional gradient matrix:

$$\boldsymbol{Grad} = \sum_{i=1}^{3} abs(\boldsymbol{Grads}_i), \tag{11}$$

where $abs(\cdot)$ is the absolute value function, and i respectively represent the RGB three channels.

Positions with larger gradient values are expected to have larger updates. Therefore, we prioritize selecting these positions for updates to obtain the region to be updated in each iteration. Assuming the perturbation area occupies a percentage $\beta\%$ of the total image area, we search for the top $\beta\%$ largest values in the gradient matrix, denoted as the $threshold$. The construction rules for the mask are as follows:

$$\boldsymbol{Mask} = \begin{cases} 1, & \boldsymbol{Grad}_{ij} > threshold \\ 0, & \boldsymbol{Grad}_{ij} \le threshold \end{cases} \tag{12}$$

3.2.2 Attack Algorithm

The Mask Momentum Iterative Attack (*MMIA*) is summarized in Algorithm 1. To enhance the effectiveness of the attack, we employ image enhancement and spatial momentum techniques to diversify the input images. Following the DI-FGSM, we introduce random transformations to the input image at each iteration, generating different input patterns to introduce randomness to the adversarial perturbation, thereby enhancing transferability. For spatial momentum techniques, we follow the concept of SMI-FGSM. We transform the image by adding random padding around it and resizing it to the original size, inducing pixel shifts. The final gradient $\boldsymbol{Grad}$ obtained by Eq 5 considers the accumulation of spatial momentum from multiple random transformations, achieving different gradient contributions from context pixels.

Regarding temporal momentum techniques, we draw inspiration from MI-FGSM, continuously updating and retaining gradient information generated in previous iterations to prevent falling into local optima. So, we further update the gradient $\boldsymbol{Grad}$ based on Eq 3. Ultimately, We employ an optimal perturbation strategy by Eq 12 to find the $\boldsymbol{Mask}$ for the obtained gradient information at each iteration step and update it according to the following formula:

$$\boldsymbol{x}^*_{i+1} = \boldsymbol{x}^*_i + \alpha \cdot sign(\boldsymbol{Grad}_{i+1}) \cdot \boldsymbol{Mask} \tag{13}$$

3.3. Model Ensemble Joint Optimization

In this section, we explore how to effectively use *MMIA* to attack model ensembles. To enhance performance and robustness, ensemble methods have been widely adopted in research [5, 18, 25]. Previous studies [39] indicate that if a sample remains adversarial across multiple models, it may capture an inherent direction that consistently deceives these models and is more likely to transfer to other models simultaneously. The ensemble approach can also be applied to adversarial attacks, thereby achieving potent black-box attacks.

In [14], it is suggested to attack the ensemble of model logits by merging multiple logit activations. [39] proposes averaging the predicted probabilities of each model during prediction. As *MMIA* involves selecting sensitive regions based on gradient information, we recommend to merge and attack the gradient information of the model ensemble. The specific approach is as follows: in each iteration, let $\boldsymbol{Grad}_k$ be the gradient information obtained after *MMIA* attacking the k-th model, and $\boldsymbol{Mask}_k$ be the corresponding mask

Table 1. The attack success and stealth score (AS^3) (%) of non-targeted adversarial attacks against 10 models on ImageNet dataset. * represent the white-box attacks. The complete results can be found in Appendix B.

	Attack	Inception v3	ResNet50	VGG16	MobileNet	DenseNet	GoogleNet	ResNet50$_{adv}$	MobileNet$_{adv}$	ShuffleNet$_{adv}$	RegNetX$_{adv}$
Inception v3	FGSM	45.95*	21.51	21.55	32.89	24.62	28.69	**6.81**	21.01	**19.52**	**6.24**
	MI-FGSM	63.23*	24.23	23.29	33.57	24.89	27.52	4.29	18.54	15.42	4.39
	SMI-FGSM	49.82*	19.00	16.37	25.80	18.47	21.35	4.30	21.07	9.73	3.65
	MMIA	**71.54***	**38.41**	**31.59**	**41.73**	**39.73**	**40.13**	5.08	**24.38**	11.86	4.72
ResNet50	FGSM	24.14	37.38*	24.52	31.66	27.48	26.77	**6.37**	21.13	**18.54**	**6.19**
	MI-FGSM	34.5	63.60*	38.55	41.93	44.53	35.68	5.04	20.81	14.27	5.19
	SMI-FGSM	22.94	49.95*	21.02	29.19	28.10	23.06	4.31	24.96	11.25	4.23
	MMIA	**57.76**	**71.67***	**56.46**	**58.31**	**65.42**	**55.12**	5.43	**30.78**	10.79	5.27
RegNetX$_{adv}$	FGSM	46.39	43.65	40.61	46.64	44.30	46.77	35.66	50.65	32.77	51.26*
	MI-FGSM	62.93	57.97	**53.35**	61.36	**59.10**	63.38	50.36	63.32	44.66	70.22*
	SMI-FGSM	29.12	24.15	23.84	33.79	25.61	31.25	30.07	42.72	33.57	42.07*
	MMIA	**66.00**	**58.97**	53.04	**61.66**	57.53	**67.85**	**56.26**	**69.28**	**50.63**	**74.27***

matrix obtained using the optimal perturbation strategy. We obtain the final gradient for updating perturbation using the following rule and generate perturbation.

$$\boldsymbol{Grad} = \sum_{i=1}^{K} \boldsymbol{Grad}_k \cdot \boldsymbol{Mask}_k, \tag{14}$$

where k is the number of the models. We summarize the *MMIA* algorithm for attacking model ensemble whose gradients are averaged in Appendix A.

4. Experiments

In this section, we present experimental results to demonstrate the effectiveness of the proposed method. We first specify the experimental setup in Sec. 4.1. Then, in Sec. 4.2 and Sec. 4.3, we discuss the results of attacking a single model and a ensemble of models respectively. Finally, a series of ablation experiments are conducted in Sec. 4.4.

4.1. Experimental Setup

Datasets. We perform experiments using the ImageNet [13] and CIFAR-10 [23] datasets, which are widely utilized in classification tasks. For evaluation purposes, we selected 1000 images from the validation sets of each dataset, ensuring a diverse representation of different categories.

Model Architectures. For each dataset, we investigate 6 normal trained models: ResNet50 [19], VGG16 [45], Inception V3 [49], DenseNet [21], MobileNet [44], and GoogleNet [48]. Additionally, for the ImageNet dataset, we reported results for four adversarially trained models [50]: ResNet50$_{adv}$, ShuffleNet$_{adv}$ [59], MobileNet$_{adv}$, and RegNetX$_{adv}$ [43].

Baseline. In our experiments, we compare our method against one-step gradient methods *i.e.* FGSM, and iterative methods including MI-FGSM and SMI-FGSM. Since optimization-based methods cannot explicitly control the distance between adversarial examples and their corresponding genuine examples, they were not directly compared to our approach. All experiments in this study were conducted based on non-targeted attacks under the L_∞ norm bound.

Hyper-Parameters. For all experiments, the maximum perturbation is set to 16, with pixel values ranging from [0, 255]. And for all iterative methods, we set the number of iterations to 15 with a step size of 1.6. For MI-FGSM, we followed the recommendation in [14] and set μ to 1.0. As for SMI-FGSM, we configure the transformation count n to be 12. For *MMIA* , we set the transformation probability p to 0.5 and perturbation size β to 50%.

Evaluation Metrics. For the assessment of attack effectiveness, we employ the Attack Success Rate (ASR). Evaluating the ASR is meaningful only when the model correctly classifies the original images. Therefore, we only consider images that the model correctly classifies for calculating the Attack Success Rate:

$$ASR = \frac{N_{attack}}{N_{correct}}, \tag{15}$$

where $N_{correct}$ is the number of images in the test set that the model correctly classifies, and N_{attack} is the number of images in the $N_{correct}$ set that attack successfully.

For assessing the stealthiness of the attacks, we use the Structural Similarity Index (SSIM) score, as indicated by Eq 8. In the end, we utilize the Attack Success and Stealth Score (AS^3) to compute the algorithm's overall performance by $AS^3 = ASR \cdot SSIM$.

4.2. Single-Model Attack

We report the AS^3 for various attacks on different models within the ImageNet dataset and the CIFAR-10 dataset in the Table 1 and Table 2 separately. The complete results can be found in Appendix B. The adversarial examples are generated for different normal and adversarial models using the 4 attack methods mentioned in the experimental setup. Evaluations are conducted separately under white-box and black-box scenarios.

From the table, we can observe that MMIA consistently achieves the highest AS^3 scores in most cases on two

Table 2. The attack success and stealth score (AS^3) (%) of non-targeted adversarial attacks against 4 models on CIFAR-10 dataset. * represent the white-box attacks. The complete results can be found in Appendix B.

	Attack	Inception v3	ResNet50	MobileNet	DenseNet
Inception v3	FGSM	58.72*	43.84	64.01	52.39
	MI-FGSM	76.55*	55.03	67.56	57.79
	SMI-FGSM	75.66*	54.58	66.41	55.78
	MMIA	**78.73***	**65.01**	**69.15**	**64.30**
ResNet50	FGSM	58.09	51.37*	62.49	50.13
	MI-FGSM	75.62	85.63*	71.57	77.23
	SMI-FGSM	76.05	86.10*	73.31	78.39
	MMIA	**80.98**	**87.10***	**77.52**	**83.72**
MobileNet	FGSM	70.18	60.39	71.96*	62.70
	MI-FGSM	80.34	63.02	86.13*	64.80
	SMI-FGSM	82.60	65.95	87.81*	67.92
	MMIA	**84.79**	**76.21**	**87.92***	**75.94**
DenseNet	FGSM	58.38	48.18	63.92	50.26*
	MI-FGSM	76.36	76.08	70.79	85.20*
	SMI-FGSM	72.75	76.49	72.60	83.80*
	MMIA	**79.37**	**82.30**	**75.21**	**87.09***

datasets, demonstrating the effectiveness and stealthiness of the algorithm. Analysis reveals that the one-step gradient method FGSM has the lowest SSIM score, indicating poorer stealthiness, despite its impressive performance in black-box attacks. Iterative methods MI-FGSM and SMI-FGSM show improved SSIM scores compared to FGSM, achieving strong attack effectiveness, but still causing significant structural damage to the entire image. In contrast, our MMIA approach perturbs the most sensitive regions of the image using a masking mechanism, enhancing stealthiness while maintaining a high attack success rate.

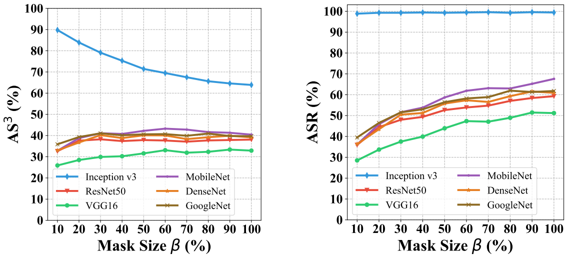
4.3. Model Ensemble Joint Optimization

In this section, we demonstrate the effectiveness of *MMIA* using a strategy of jointly optimizing model ensemble. We compare the AS^3 of *MMIA* when conducting black-box attacks on individual models versus using a strategy of joint optimization for attacks. Our study involves 5 models on the ImageNet dataset: Inception v3, ResNet50, VGG16, MobileNet, and DenseNet. The experimental results are shown in Table 3. We initially attack each individual model to generate adversarial samples and perform black-box testing on the remaining four models (see lines 2-6). For the joint optimization strategy attack, we generate adversarial samples by attacking with four models and conduct black-box testing on the remaining one model (see the last line).

It can be observed that the strategy of jointly optimizing model ensemble is highly effective in enhancing the transferability and stealthiness of adversarial samples. Compared to attacking individual models alone, attacking the model ensemble achieves optimal results for all scenarios.

Table 3. The attack success and stealth score (AS^3) (%) of non-targeted adversarial attacks against 5 single models and model ensembles on ImageNet dataset. The last row represents the joint optimization of attacking examples against 4 other models excluding the one in each column.

	Inception v3	ResNet50	VGG16	MobileNet	DenseNet
Inception v3	-	38.41	31.59	41.73	39.73
ResNet50	57.76	-	56.46	58.31	61.41
VGG16	49.08	54.22	-	57.84	58.82
MobileNet	60.56	58.88	58.30	-	61.63
DenseNet	50.98	56.87	51.17	55.66	-
Ensemble	**61.99**	**60.28**	**59.33**	**60.47**	**62.01**



(a) The results of AS^3.

(b) The results of ASR.

Figure 3. Ablation studies on different mask size β. The curve of Inception v3 corresponds to white-box attack, and the others represent black-box attacks.

4.4. Ablation Studies

In this section, we conducted a series of ablation experiments to study the impact of different parameters. We only consider attacking a single network on the ImageNet dataset: Inception v3. The maximum perturbation value for each pixel in all experiments is set to 16.

Mask Size β. The perturbation mask size β plays a crucial role in enhancing the AS^3 metric. If beta equals 100%, the *MMIA* method transforms into a conventional iterative attack based on momentum and image transformations. Therefore, we investigated the appropriate mask size value. We use the *MMIA* to attack the Inception v3 model, generating adversarial examples with mask sizes ranging from 10% to 100%. The AS^3 and ASR for attacking the Inception v3, ResNet50, VGG16, MobileNet, DenseNet and GoogleNet model are presented in Figure 3, where the blue curve corresponds to white-box attacks on Inception v3, and the others represent black-box attacks.

It can be observed that for the ASR, increasing mask size corresponds to a higher ASR. However, for the AS^3, larger mask sizes tend to impact stealthiness, *i.e.* compromising $SSIM$ scores. In Figure 3a, the peak of AS3 for black-box attacks appears to be around 50%. Therefore, we recommend setting the mask size to 50% to strike a balance between ASR and $SSIM$ scores.

Total iteration number t. We currently investigate the

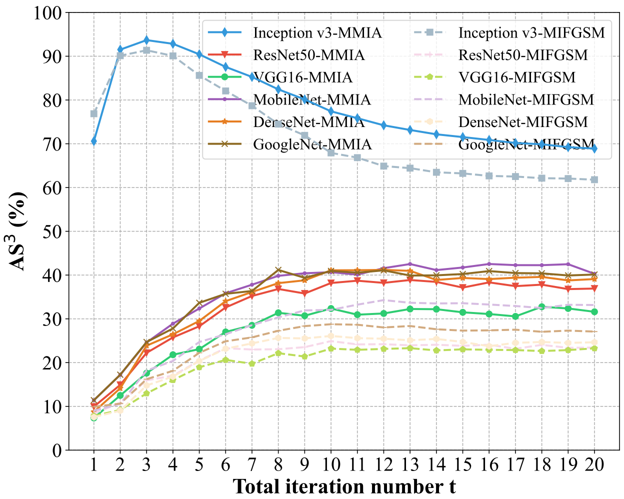


Figure 4. Ablation studies on different total iteration number t. The curve of Inception v3 corresponds to white-box attack, and the others represent black-box attacks.

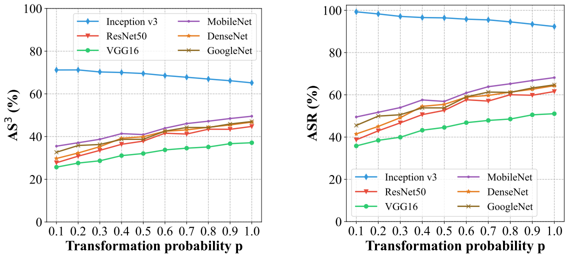
impact of the total iteration number t. We employ *MMIA* and MI-FGSM to attack the Inception v3 model with iteration numbers ranging from 1 to 20. Subsequently, we evaluate the AS^3 of adversarial examples against the Inception v3, ResNet50, VGG16, MobileNet, DenseNet and GoogleNet model. The results are shown in Figure 4.

It can be observed that the *MMIA* outperforms MI-FGSM in terms of AS^3 at both low and high iteration numbers. Furthermore, with the increase in iteration numbers, AS^3 against black-box models gradually increases and tends to stabilize, while high iteration numbers may lead to overfitting and a decline in performance against white-box models.

Transformation probability p**.** We further study the impact of the transformability probability p in *MMIA* . We discuss the scenario when p ranges from 0 to 1. The results are shown in Figure 5. We observe that with an increase in p, *MMIA* achieves higher AS^3 and ASR for black-box scenarios, while performance for white-box scenarios decreases. This trend provides valuable insights for building robust adversarial attacks in practice. Specifically, if you know that the black-box model is entirely different from any existing networks, setting $p = 1$ can maximize portability. If the black-box model is a hybrid of a new and existing network, choosing a moderate p value is advisable.

5. Case Study

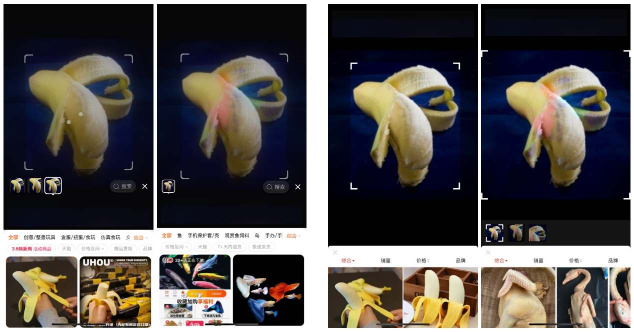
Here, we showcase the effectiveness of *MMIA* in practical attack scenarios. The application of *MMIA* is highly flexible. Under the rules of unconstrained adversarial attacks [3], we can set the mask size β to a very small value, increase the step size, and allow a wide range of perturbation. This allows the generation of highly effective adversarial examples with minimal damage to the image. To validate



(a) The results of AS^3.

(b) The results of ASR.

Figure 5. Ablation studies on different transformation probability p. The curve of Inception v3 corresponds to white-box attack, and the others represent black-box attacks.



(a) Taobao

(b) PinDuoDuo

Figure 6. Attack Taobao and PinDuoDuo platform with our adversarial example. The banana in (a) and (b) are identified incorrectly.

the effectiveness of *MMIA* , we conduct the model ensemble *MMIA* on an image of a banana. The mask size was set to only 2%, and the perturbation range was 100. As shown in Figure 6, this adversarial sample successfully attacks both the Taobao and PinDuoDuo platforms while visually preserving the original image, demonstrating the superiority of *MMIA* in terms of effectiveness and stealthiness.

6. Conclusions

In this paper, we propose the Mask Momentum Iterative Attack (*MMIA*) method. It effectively deceives both white-box and black-box models while maintaining stealthiness. Our method outperforms one-step gradient-based approaches and momentum iterative methods in terms of AS^3. We conduct extensive experiments to validate the effectiveness of the proposed method and perform ablation studies to investigate key influencing factors. To further enhance the transferability of adversarial examples, we recommend using the mask mechanism to attack model ensemble and fuse multiple gradients. Finally, we showcase the application of our attack in real-world scenario. We hope that our *MMIA* can inspire the development of more robust deep models.

Acknowledge

This work is supported by grant No. KZ46009501.

References

[1] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. In *International conference on machine learning*, pages 284–293. PMLR, 2018. 1

[2] Wieland Brendel, Jonas Rauber, and Matthias Bethge. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. *arXiv preprint arXiv:1712.04248*, 2017. 3

[3] Tom B Brown, Nicholas Carlini, Chiyuan Zhang, Catherine Olsson, Paul Christiano, and Ian Goodfellow. Unrestricted adversarial examples. *arXiv preprint arXiv:1809.08352*, 2018. 8

[4] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. Ieee, 2017. 1

[5] Rich Caruana, Alexandru Niculescu-Mizil, Geoff Crew, and Alex Ksikes. Ensemble selection from libraries of models. In *Proceedings of the twenty-first international conference on Machine learning*, page 18, 2004. 5

[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. 3

[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. 1

[8] 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. 3

[9] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017. 1

[10] Xuesong Chen, Xiyu Yan, Feng Zheng, Yong Jiang, Shu-Tao Xia, Yong Zhao, and Rongrong Ji. One-shot adversarial attacks on visual tracking with dual attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10176–10185, 2020. 2

[11] Minhao Cheng, Simranjit Singh, Patrick Chen, Pin-Yu Chen, Sijia Liu, and Cho-Jui Hsieh. Sign-opt: A query-efficient hard-label adversarial attack. *arXiv preprint arXiv:1909.10773*, 2019. 3

[12] Shuyu Cheng, Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Improving black-box adversarial attacks with a transfer-based prior. *Advances in neural information processing systems*, 32, 2019. 3

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6

[14] 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. 1, 2, 3, 5, 6

[15] 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. 2, 3

[16] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 1, 2, 3

[17] Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*, 2017. 1

[18] Lars Kai Hansen and Peter Salamon. Neural network ensembles. *IEEE transactions on pattern analysis and machine intelligence*, 12(10):993–1001, 1990. 5

[19] 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. 6

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 630–645. Springer, 2016. 2

[21] 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. 6

[22] Wei Jiang, Tianyuan Zhang, Shuangcheng Liu, Weiyu Ji, Zichao Zhang, and Gang Xiao. Exploring the physical-world adversarial robustness of vehicle detection. *Electronics*, 12 (18):3921, 2023. 2

[23] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 6

[24] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. 1

[25] Anders Krogh and Jesper Vedelsby. Neural network ensembles, cross validation, and active learning. *Advances in neural information processing systems*, 7, 1994. 5

[26] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*, 2016. 2

[27] 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. 2

[28] 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. 1

[29] Simin Li, Shuning Zhang, Gujun Chen, Dong Wang, Pu Feng, Jiakai Wang, Aishan Liu, Xin Yi, and Xianglong

Liu. Towards benchmarking and assessing visual naturalness of physical world adversarial attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12324–12333, 2023. 2

[30] Aishan Liu, Xianglong Liu, Jiaxin Fan, Yuqing Ma, Anlan Zhang, Huiyuan Xie, and Dacheng Tao. Perceptual-sensitive gan for generating adversarial patches. In *AAAI*, 2019. 1

[31] Aishan Liu, Jiakai Wang, Xianglong Liu, Bowen Cao, Chongzhi Zhang, and Hang Yu. Bias-based universal adversarial patch attack for automatic check-out. In *ECCV*, 2020. 1

[32] Aishan Liu, Xianglong Liu, Hang Yu, Chongzhi Zhang, Qiang Liu, and Dacheng Tao. Training robust deep neural networks via adversarial noise propagation. *TIP*, 2021. 2

[33] Aishan Liu, Jun Guo, Jiakai Wang, Siyuan Liang, Renshuai Tao, Wenbo Zhou, Cong Liu, Xianglong Liu, and Dacheng Tao. X-adv: Physical adversarial object attacks against x-ray prohibited item detection. In *USENIX Security Symposium*, 2023. 1

[34] Aishan Liu, Jun Guo, Jiakai Wang, Siyuan Liang, Renshuai Tao, Wenbo Zhou, Cong Liu, Xianglong Liu, and Dacheng Tao. X-adv: Physical adversarial object attacks against x-ray prohibited item detection. *arXiv preprint arXiv:2302.09491*, 1, 2023. 2

[35] Aishan Liu, Shiyu Tang, Xinyun Chen, Lei Huang, Haotong Qin, Xianglong Liu, and Dacheng Tao. Towards defending multiple lp-norm bounded adversarial perturbations via gated batch normalization. *International Journal of Computer Vision*, 2023.

[36] Aishan Liu, Shiyu Tang, Xinyun Chen, Lei Huang, Haotong Qin, Xianglong Liu, and Dacheng Tao. Towards defending multiple p-norm bounded adversarial perturbations via gated batch normalization. *International Journal of Computer Vision*, pages 1–18, 2023. 2

[37] Aishan Liu, Xinwei Zhang, Yisong Xiao, Yuguang Zhou, Siyuan Liang, Jiakai Wang, Xianglong Liu, Xiaochun Cao, and Dacheng Tao. Pre-trained trojan attacks for visual recognition. *arXiv preprint arXiv:2312.15172*, 2023. 1

[38] Shunchang Liu, Jiakai Wang, Aishan Liu, Yingwei Li, Yijie Gao, Xianglong Liu, and Dacheng Tao. Harnessing perceptual adversarial patches for crowd counting. In *ACM CCS*, 2022. 1

[39] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and blackbox attacks. *arXiv preprint arXiv:1611.02770*, 2016. 2, 5

[40] 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. 2, 3

[41] Abdel-rahman Mohamed, George E Dahl, and Geoffrey Hinton. Acoustic modeling using deep belief networks. *IEEE transactions on audio, speech, and language processing*, 20 (1):14–22, 2011. 1

[42] Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17:527–566, 2017. 3

[43] Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. Designing network design spaces. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10428–10436, 2020. 6

[44] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. 6

[45] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 6

[46] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27, 2014. 1

[47] 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. 1, 2, 3

[48] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. 6

[49] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 6

[50] Shiyu Tang, Ruihao Gong, Yan Wang, Aishan Liu, Jiakai Wang, Xinyun Chen, Fengwei Yu, Xianglong Liu, Dawn Song, Alan Yuille, et al. Robustart: Benchmarking robustness on architecture design and training techniques. *arXiv preprint arXiv:2109.05211*, 2021. 2, 6

[51] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204*, 2017. 2

[52] Guoqiu Wang, Huanqian Yan, and Xingxing Wei. Enhancing transferability of adversarial examples with spatial momentum. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pages 593–604. Springer, 2022. 2, 3

[53] Haodi Wang, Kai Dong, Zhilei Zhu, Haotong Qin, Aishan Liu, Xiaolin Fang, Jiakai Wang, and Xianglong Liu. Transferable multimodal attack on vision-language pre-training models. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 102–102. IEEE Computer Society, 2024. 2

[54] Jiakai Wang, Aishan Liu, Zixin Yin, Shunchang Liu, Shiyu Tang, and Xianglong Liu. Dual attention suppression attack: Generate adversarial camouflage in physical world. In *CVPR*, 2021. 1

[55] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 2

[56] 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. 2, 3

[57] Chongzhi Zhang, Aishan Liu, Xianglong Liu, Yitao Xu, Hang Yu, Yuqing Ma, and Tianlin Li. Interpreting and improving adversarial robustness of deep neural networks with neuron sensitivity. *IEEE Transactions on Image Processing*, 2021. 2

[58] Tianyuan Zhang, Yisong Xiao, Xiaoya Zhang, Hao Li, and Lu Wang. Benchmarking the physical-world adversarial robustness of vehicle detection. *arXiv preprint arXiv:2304.05098*, 2023. 2

[59] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6848–6856, 2018. 6