

Improving Generalization of Universal Adversarial Perturbation via Dynamic Maximin Optimization

Yechao Zhang¹, Yingzhe Xu¹, Junyu Shi¹, Leo Yu Zhang²,
Shengshan Hu¹, Minghui Li¹, Yanjun Zhang³

¹Huazhong University of Science and Technology

²Griffith University

³University of Technology Sydney

{ycz, hynxyz, string1, hushengshan, minghuili}@hust.edu.cn, leocityu@outlook.com, yanjun.j.zhang@gmail.com

Abstract

Deep neural networks (DNNs) are susceptible to universal adversarial perturbations (UAPs). These perturbations are meticulously designed to fool the target model universally across all sample classes. Unlike instance-specific adversarial examples (AEs), generating UAPs is more complex because they must be generalized across a wide range of data samples and models. Our research reveals that existing universal attack methods, which optimize UAPs using DNNs with static model parameter snapshots, do not fully leverage the potential of DNNs to generate more effective UAPs. Rather than optimizing UAPs against static DNN models with a fixed training set, we suggest using dynamic model-data pairs to generate UAPs. In particular, we introduce a dynamic maximin optimization strategy, aiming to optimize the UAP across a variety of optimal model-data pairs. We term this approach DM-UAP. DM-UAP utilizes an iterative max-min optimization framework that refines the model-data pairs, coupled with a curriculum UAP learning algorithm to examine the combined space of model parameters and data thoroughly. Comprehensive experiments on the ImageNet dataset demonstrate that the proposed DM-UAP markedly enhances both cross-sample universality and cross-model transferability of UAPs. Using only 500 samples for UAP generation, DM-UAP outperforms the state-of-the-art approach with an average increase in fooling ratio of 12.108%.

Code — <https://github.com/yechao-zhang/DM-UAP>

1 Introduction

Deep Neural Networks (DNNs) represent the cutting edge of computer vision (Krizhevsky, Sutskever, and Hinton 2012; Simonyan and Zisserman 2014; He et al. 2016). A key strength of DNNs is their ability to identify subtle patterns beyond human perception. However, this capability of DNNs can unintentionally enable adversarial attacks, where deceptively minor adversarial perturbations of the input—imperceptible to humans—create adversarial examples (AE) that lead to mispredictions of DNNs (Szegedy et al. 2014).

Adding to this concern is the emergence of a notable class of adversarial techniques known as universal adversarial perturbation (UAP), which are designed to induce

mispredictions against universal input samples (Moosavi-Dezfooli et al. 2017; Khruikov and Oseledets 2018; Poursoeed et al. 2018; Shafahi et al. 2020). In addition to this characteristic of *cross-sample universality*, UAPs could also possess another element of generalization in adversarial attacks, termed *cross-model transferability*. This denotes the capacity of (AEs) to preserve their efficacy across different DNN models (Hu et al. 2022; Li et al. 2024). Together, these capabilities enable UAPs to facilitate the generation of transferable AEs at scale, compromising various DNN architectures and multiple machine learning tasks simultaneously (Zhong and Deng 2022; Xie et al. 2021; Ding et al. 2021; Wang et al. 2023; Zhou et al. 2023a,b, 2024). Consequently, UAPs have garnered significant attention and have become a focal point of extensive research in the field.

Recent research aimed at improving the generalization of UAPs has explored different aspects. Some focus on analyzing the training data used for UAP generation (Zhang et al. 2020b; Li et al. 2022), while others reinterpret the gradient aggregation within and between the training batch (Shafahi et al. 2020; Liu et al. 2023). In addition, refinements to training loss have also been explored (Zhang et al. 2021). However, these existing works fall short by only refining UAPs against fixed DNN parameters, thereby limiting the generalization to static instances of the model without fully harnessing the adaptive capabilities inherent in the model.

In this work, we highlight that utilizing the ever-evolving optimized model parameter configurations can enable the generation of stronger and more transferable UAPs. The intuition behind this is that exposing the generation process to a wider range of model variance, *even within a fixed architecture*, fosters UAPs to capture a wider spectrum of vulnerabilities between models. This, in turn, allows UAPs to generalize to models on which they were not directly optimized. Furthermore, we embed the objectives of *cross-sample universality* and *cross-model transferability* into a unified maximin formulation, which involves the composite inner minimization of the model parameters and training data and an outer maximization of the UAP. In other words, we generate UAPs on a wide range of parameter-data pairs that have jointly minimized the classification loss *w.r.t.* their respective labels. In this way, the generated UAP is optimized to maximize the loss in more antagonistic scenarios from a combined landscape of both parameter and input space.

To effectively address the optimization challenges associated with this maximin formulation, we develop an iterative max-min-min optimization framework with dynamic adaptation designs. We decouple the inner composite optimization into a two-stage min-min optimization process, where we dynamically optimize a parameter-data pair within each minibatch iteration. Additionally, we introduce curriculum UAP learning, in which we first proceed the maximization against “easy” parameter-data pairs, and gradually increase their optimization spaces along with the training epoch, to ensure a smoother UAP optimization. In general, the main contributions of our work can be summarized as follows:

- We propose a novel maximin formulation, which adapts the model parameters during the optimization process to improve the generalization of UAP. To the best of our knowledge, this is the first exploration of the model parameter landscape for UAP generation.
- To effectively solve the maximin formulation, we propose a dynamic iterative max-min-min optimization framework, which decouples the composite minimization into a two-stage min-min optimization and leverages curriculum learning to ensure a smooth UAP update.
- Extensive experiments on the ImageNet dataset demonstrate the superior generalization of UAPs generated by the proposed DM-UAP compared to the state-of-the-art methods under various attack settings.

2 Related Work

2.1 Instance-specific Attacks

Instance-specific attacks generate AEs by adding perturbations to individual samples. In a white-box setting, the adversary has full knowledge of the target model (architecture and parameters), constructing AEs using directly computed gradients, in a single-step (Goodfellow, Shlens, and Szegedy 2014) or iterative manner (Madry et al. 2017). In a black-box setting, the adversary lacks direct access to the target model. Alternatively, AEs are crafted against surrogate models, expecting effectiveness against the black-box target model. Since the discovery of instance-specific AE (Szegedy et al. 2013), research has extensively studied their transferability, exploring loss function refinement (Zhao, Liu, and Larson 2021), input transformations (Xie et al. 2019), and model ensembles (Zhang et al. 2024) to improve the effectiveness.

2.2 Universal Adversarial Attacks

The seminal work (Moosavi-Dezfooli et al. 2017) of UAPs, denoted as UAP, proposed aggregating perturbation vectors obtained from instance-specific attack methods to generate UAPs. Subsequent work, such as GAP (Poursaeed et al. 2018) and NAG (Mopuri et al. 2018), introduced the use of generative models for UAP generation.

In a later development, SPGD (Shafahi et al. 2020) formulated the generation of UAPs as a maximization of the average prediction loss on universally perturbed training data and used the stochastic gradient method with mini-batch training to solve it. Since then, the mini-batch training paradigm has become a common approach to UAP generation in subsequent works (Co et al. 2021; Zhang et al. 2020a;

Benz et al. 2020; Liu et al. 2023; Hu et al. 2021). Among these, DF-UAP (Zhang et al. 2020b) found that UAPs dominate DNN prediction, and the original samples behave somewhat like random noise after superimposing UAPs. This finding inspired AT-UAP (Li et al. 2022), a more robust UAP achieved by integrating image-specific attacks and universal attacks, to enhance such a dominant effect. Recently, SGA (Liu et al. 2023) proposed to aggregate the noisy gradients obtained from multiple small-batch as a gradient estimation of a large-batch. However, these methods have not yet sought to improve the generalization of UAPs from the perspective of manipulating the underlying model.

3 Methodology

In this section, we first introduce existing UAP problem formulations and present our novel maximin formulation. After that, we present the details of our proposed max-min-min optimization framework for solving this formulation.

3.1 Preliminaries

Suppose \mathbf{S} contains a set of samples $(x, y) \in \mathcal{D}$ drawn from the data distribution \mathcal{D} , where $x \in \mathcal{X}$ represents the image feature and $y \in \mathcal{Y}$ denotes its corresponding label. The DNN model is depicted as f_θ , where f is a DNN function and θ represents the model parameters. In general, θ are optimized through empirical risk minimization (ERM) as follows:

$$\min_{\theta} \frac{1}{\|\mathbf{S}\|} \sum_{i=1}^{\|\mathbf{S}\|} \mathcal{L}(f_\theta(x_i), y_i), \quad (1)$$

where \mathcal{L} is the loss function (*e.g.*, cross-entropy) for training.

Empirical Risk Maximization. In a universal adversarial attack, the objective is to craft a single perturbation δ to fool the well-trained DNN model f_θ for most images within \mathcal{X} . In practice, existing work aims to optimize δ which maximizes the averaged prediction loss for universally perturbed data as follows:

$$\max_{\|\delta\|_\infty \leq \epsilon} \frac{1}{n} \sum_{i=1}^n \mathcal{L}(f_\theta(x_i + \delta), y_i), \quad (2)$$

where $x_1, \dots, x_n \in \mathcal{X}$ are training data used for optimizing the δ , y_i is the corresponding label of the unperturbed input x_i predicted by $y_i = \arg \max f_\theta(x_i)$, and UAP δ is a small perturbation constrained by ℓ_∞ -norm of ϵ .

Optimal-Data UAP Maximin. In addition to the standard empirical risk maximization, (Li et al. 2022) introduces a maximin optimization that leverages image-specific adversarial attacks to produce more effective training inputs. This leads to crafting a UAP based on a tailored training dataset. Formally, this maximin optimization is depicted as follows:

$$\begin{aligned} & \max_{\|\delta\|_\infty \leq \epsilon} \frac{1}{n} \sum_{i=1}^n \mathcal{L}(f_\theta(x_i^* + \delta), y_i), \\ & \text{s.t. } \forall i, x_i^* = \arg \min_{\|x' - x_i\|_2 \leq r} \mathcal{L}(f_\theta(x'), y_i). \end{aligned} \quad (3)$$

This maximin optimization searches for optimal data point x_i^* in the r -bounded ℓ_2 -norm vicinity of each original input

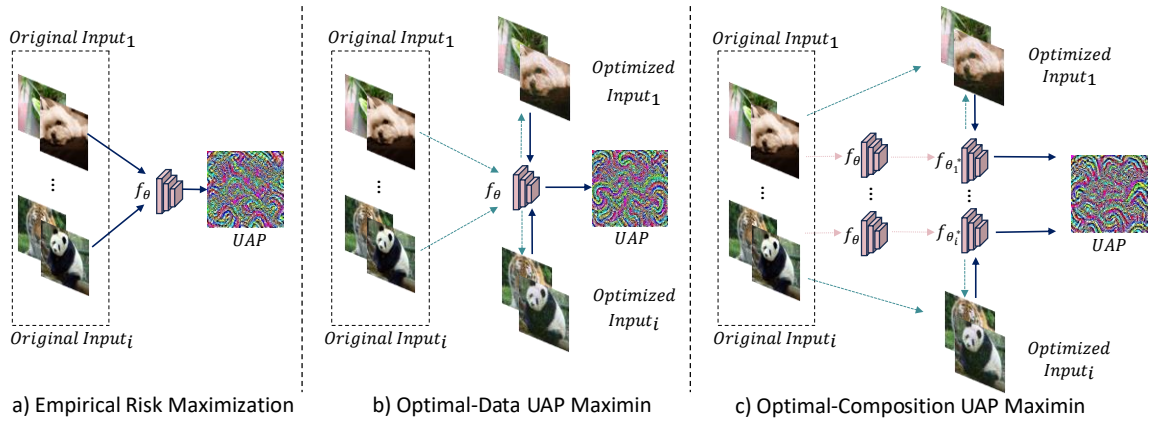


Figure 1: The illustration of different optimization flows: a) Eq. (2) use the original data and model for UAP generation; b) Eq. (3) use the model and original inputs to obtain optimized inputs, then use model and optimized inputs for UAP generation; c) Eq. (5) use original model and original inputs to obtain optimized models and inputs, then use them for UAP generation.

x_i , which minimizes the loss for its corresponding label y_i first. Then the UAP δ is optimized to maximize the loss over these optimal data points. Compared to Eq. (2), this formulation constructs antagonistic training data for maximization, compelling the UAP to avoid trivial solutions that may not effectively deceive the model f_θ .

3.2 Problem Formulation

In contrast to previous formulations (Eqs. (2) and (3)) that optimize UAP against a static model f_θ —whether for a direct white-box attack or through a surrogate model in transfer-based attacks—we propose optimizing across multiple variants of the model parameter θ to enhance the potency of the UAP.

Optimal-Parameters UAP Maximin. Inspired by the optimal-data UAP maximin concept in Eq. (3), we introduce a parallel optimization as follows:

$$\begin{aligned} \max_{\|\delta\|_\infty \leq \epsilon} \frac{1}{n} \sum_{i=1}^n \mathcal{L}(f_{\theta_i^*}(x_i + \delta), y_i), \\ \text{s.t. } \forall i, \theta_i^* = \arg \min_{\|\theta' - \theta\|_2 \leq \rho} \mathcal{L}(f_{\theta'}(x_i), y_i). \end{aligned} \quad (4)$$

Unlike Eq. (3), this optimization finds the optimal model parameters θ_i^* that minimize the loss for each data point (x_i, y_i) within a ρ -bounded ℓ_2 -norm vicinity of the original parameters θ . The UAP δ is then optimized to maximize the average loss across all perturbed data points $(x_i + \delta, y_i)$, evaluated using their respective model variants $f_{\theta_i^*}$. This ensures UAP effectively increases loss across different model variations, promoting a robust adversarial effect expected to generalize better across models.

Optimal-Composition UAP Maximin. Building on the optimal-parameters UAP maximin, we propose a more advanced formulation that includes a composite inner mini-

mization over both model parameters and data points:

$$\begin{aligned} \max_{\|\delta\|_\infty \leq \epsilon} \frac{1}{n} \sum_{i=1}^n \mathcal{L}(f_{\theta_i^*}(x_i^* + \delta), y_i), \\ \text{s.t. } \forall i, (\theta_i^*, x_i^*) = \arg \min_{\substack{\|\theta' - \theta\|_2 \leq \rho \\ \|x' - x_i\|_2 \leq r}} \mathcal{L}(f_{\theta'}(x'), y_i). \end{aligned} \quad (5)$$

In this context, each (θ_i^*, x_i^*) pair is simultaneously optimized to reduce the loss for y_i , while remaining within their respective neighborhoods defined by ρ and r . This sets up a more challenging condition for the UAP δ , which requires it to maximize loss against the best-modified models and their corresponding inputs. Both Eqs. (3) and (4) represent special cases of this formulation. Fig. 1 illustrate the optimization flow of Eqs. (2), (3) and (5).

3.3 Dynamic Maximin UAP: An Iterative Max-Min-Min Optimization Framework

Tackling Eq. (5) is a non-trivial task. First, the minimization process involves a dual-layered optimization that requires accurate navigation through two high-dimensional spaces. Second, the maximization phase must be flexible enough to cope with a wide range of combined landscapes concerning both data and the model parameters, which poses greater optimization challenges compared to Eqs. (2) to (4). To overcome these difficulties, we introduce an iterative max-min-min optimization framework. This framework separates the composite minimization into a two-stage min-min optimization and employs curriculum learning for a smoother UAP maximization, as detailed below.

Dynamic Strategy for mini-batch Training. Intuitively, Eq. (5) suggests a straightforward implementation: find all optimal pairs $(x_i^*, f_{\theta_i^*})$, use them as a fixed training set to update UAP through mini-batch training with stochastic gradient ascent. However, this is highly inefficient due to the significant storage and memory required for forward-backward propagation across multiple models $f_{\theta_i^*}$. To address this, we adopt a dynamic strategy: during each mini-batch iteration,

we optimize a mini-batch of images with a shared model on the fly, *i.e.*, each random mini-batch of images X_B with the initial model f_θ induces a composition (X_B^*, f_{θ^*}) . After that, the perturbed mini-batch data $X_B^* + \delta$ are propagated forward and backward in parallel through f_{θ^*} to obtain the update gradient for δ .

Sequential Model and Data Optimization. As discussed earlier, within each mini-batch iteration, we decouple the composite minimization of the data and model parameters into sequential min-min optimization steps. Specifically,

- We first perform model optimization on the mini-batch images X_B to obtain the optimized model f_{θ^*} .
- Then, we use the optimized model θ^* to perform data optimization to obtain X_B^* .

The rationale for this “optimize model first, data second” approach is to ensure that the model is always updated using unperturbed data. This is because adversarially constructed images X_B^* may not belong to the correct feature distribution \mathcal{X} . Overfitting the model to adversarially constructed images X_B^* could result in significant underfitting of the original distribution (Dong et al. 2020; Xu et al. 2022).

Model Optimization. To ensure the effectiveness of resulting UAP against the original model f_θ , the optimized model parameters θ^* should not deviate much from the origin θ . However, standard gradient optimizers cannot guarantee precise adherence to the ℓ_2 -norm constraint. To mitigate this, we leverage the normalized gradient of parameters when updating θ^* . Concretely, we update θ^* as follows:

$$\theta^* \leftarrow \theta^* - \alpha_m \frac{\nabla_{\theta} \frac{1}{B} \mathcal{L}(f_{\theta^*}(X_B), Y_B)}{\|\nabla_{\theta} \frac{1}{B} \mathcal{L}(f_{\theta^*}(X_B), Y_B)\|_2}, \quad (6)$$

where Y_B are the corresponding labels of images X_B , α_m is the step size of model optimization. Given the adversarial budget ρ_t at the t -th epoch, we update θ^* for K_m steps and set the step size α_m as $\frac{\rho_t}{K_m}$ to ensure that θ^* remain within the ρ_t -bounded neighborhood of θ .

Data Optimization. As discussed earlier, the data optimization is conducted in condition to the optimized model f_{θ^*} inside each mini-batch iteration. Specifically, we update the data using the standard ℓ_2 -norm targeted PGD as follows:

$$x^* \leftarrow \Pi_{r_t}(x^* - \alpha_d \cdot \frac{\nabla_x \mathcal{L}(f_{\theta^*}(x^*), y)}{\|\nabla_x \mathcal{L}(f_{\theta^*}(x^*), y)\|_2}), \quad (7)$$

where x^* denotes a sample in mini-batch X_B^* , and $\Pi(\cdot)$ is the projection function that ensures the data perturbation does not exceed the budget of r_t . We set the step size of data optimization α_d as $1.25 \times \frac{r_t}{K_d}$. Eq. (7) is executed for all the samples in X_B^* in parallel, and repeated for K_d steps.

Curriculum UAP Learning. The outer maximization of updating UAP is challenging as it requires accumulating gradients from various parameters and inputs, all of which are constantly changing in high-dimensional neighborhoods. To address this, we leverage the concept of curriculum learning (Bengio et al. 2009). We adopt a progressive approach by initiating inner-min-min optimization of θ^* and X_B^* in smaller neighborhoods during the early training epochs.

Algorithm 1: Dynamic Maximin UAP (DM-UAP) with Curriculum Learning

Input: Data $X = x_1, \dots, x_n$, model f with parameters θ , number of epochs T , mini-batch size B .

Input: UAP maximum perturbation magnitude ϵ , initial learning rate γ .

Input: Model maximum neighborhood size ρ , number of model optimization steps K_m .

Input: Data maximum neighborhood size r , number of data optimization steps K_d .

Output: Universal Adversarial Perturbation (UAP) δ

- 1 Initialize $\delta \sim \mathcal{U}(-\epsilon, \epsilon)$;
- 2 **for** $t = 1$ to T **do**
- 3 Compute the perturbation magnitude $\rho_t = t \times \frac{\rho}{T}$ and step size $\alpha_m = \frac{\rho_t}{K_m}$ for model optimization in epoch t ;
- 4 Compute the perturbation magnitude $r_t = t \times \frac{r}{T}$ and step size $\alpha_d = 1.25 \times \frac{r_t}{K_d}$ for data optimization in epoch t ;
- 5 **for** mini-batch $X_B \in X$ **do**
- 6 $Y_B = \arg \max f_\theta(X_B)$;
- 7 Initialize $\theta^* \leftarrow \theta$;
- 8 Initialize $X_B^* \leftarrow X_B$;
- 9 **for** $k = 1$ to K_m **do**
- 10 Update model parameters with Eq. (6);
- 11 **end**
- 12 **for** $k = 1$ to K_d **do**
- 13 Update data with Eq. (7) (*in parallel*);
- 14 **end**
- 15 Update UAP with Eq. (8) ;
- 16 **end**
- 17 **end**
- 18 **return** δ ;

Then, we gradually increase the neighborhood sizes to encompass larger regions over the course of the training. In each epoch t out of a total of T epochs, we set the neighborhood sizes as $\rho_t = t \times \frac{\rho}{T}$ and $r_t = t \times \frac{r}{T}$, respectively.

To better update UAP in a dynamic environment, instead of using SGD with the sign function as in previous works (Liu et al. 2023), we use Adam as the optimizer. Since Adam adjusts learning rates and momentum to adapt to the optimization landscapes, which makes it suitable for such non-stationary objectives, the UAPs are updated adaptively to the evolving dynamics of the model and data. Formally, the UAP is updated in each iteration as follows:

$$\begin{aligned} \delta &= \text{Adam}(\nabla_{\delta} \mathcal{L}(f_{\theta^*}(X_B^* + \delta), Y_B), \gamma) \\ \delta &= \min(\max(\delta, -\epsilon), \epsilon) \end{aligned} \quad (8)$$

where γ is the initial learning rate, and the min-max operation is conducted to ensure the valid ℓ_∞ -norm constraint on the generated UAP. The entire framework of our proposed DM-UAP method is summarized in Algorithm 1.

Method	ANet	GNet	V16	V19	R152	Avg
UAP	93.30	78.90	78.30	77.80	84.00	82.46
NAG	96.44	90.37	77.57	83.78	87.24	87.08
GAP	-	82.70	83.70	80.10	-	82.17
DF-UAP	96.17	88.94	94.30	94.98	90.08	92.89
Cos-UAP	96.50	90.50	97.40	96.40	90.20	94.20
AT-UAP	97.01	90.82	97.51	97.56	91.52	94.88
SGA	96.99	90.64	97.83	96.56	92.86	94.98
Ours	97.19	93.28	98.43	97.81	92.90	95.92

Table 1: The fooling ratio (%) in the white-box setting by various UAP attack methods. The UAPs are crafted on the AlexNet, GoogleNet, VGG16, VGG19, and ResNet152.

4 Experiment

Setup: Following (Moosavi-Dezfooli et al. 2017; Liu et al. 2023), we randomly select 10 images from each category in the ImageNet training set, resulting in a total of 10,000 images, for UAP generations. In addition, we also consider a data-limit setting, in which only 500 random images from the training set are sampled. Aligning with previous work, we evaluate our method on the ImageNet validation set, which contains 50,000 images, using classical pre-trained CNN models AlexNet, GoogleNet, VGG16, VGG19, and ResNet152 as target models.

Evaluation Metrics: We employ the widely used fooling ratio metric, which calculates the variation proportion of model predictions when applying the UAP, for evaluation.

Comparative Methods: In the white-box attack scenario, we compare our method with the following existing approaches: UAP, NAG, GAP, DF-UAP, Cos-UAP, AT-UAP, and SGA. For other settings, we compare DM-UAP with UAP, GAP, SPGD, AT-UAP, and SGA. SPGD is considered the baseline, as AT-UAP, SGA, and our proposed DM-UAP all build upon its standard mini-batch training. Note that SGA is the current state-of-the-art method.

Hyper-parameters Setting: We set the maximum perturbation budget ϵ of all methods as $10/255$. Following SGA (Liu et al. 2023), the number of training epochs T is 20, and the batch size B is 125. The step numbers for inner model optimization K_m and data optimization K_d in our method are both 10, with default neighborhood size $\rho = 1$ and $r = 32$.

4.1 Generalization Performance of UAPs

We perform universal adversarial attacks under the white-box and black-box settings respectively and evaluate the overall performance of our proposed DM-UAP with baselines on the ImageNet validation set.

White-box Attack. We present the results of white-box attacks on five models using our DM-UAP approach, compared with other methods in Sec. 4. For SGA, we used its official source code and followed the settings from (Liu et al. 2023) with 10,000 samples. For other methods, we used results from their respective papers. Our method achieves the highest attack performance across all models. DM-UAP shows a notable improvement of over 2% on GoogleNet. These indicate UAPs generated by our method can better generalize to unknown samples.

Method	Samples	ANet	V16	GNet	V19	R50	R152	Avg
SPGD	500	52.34*	58.45*	22.80	45.78	24.22	19.72	37.22
M-SPGD		72.38*	86.13*	33.14	70.46	35.78	28.12	54.34 ^(+17.12)
AT-UAP		80.41*	92.11*	42.28	80.29	43.59	33.49	62.03 ^(+24.81)
SGA		82.34*	90.70*	51.07	78.79	50.95	40.76	65.77 ^(+28.55)
M-SGA		85.93*	91.42*	47.69	78.41	48.77	37.95	65.03 ^(+27.81)
Ours		91.39*	92.83*	56.23	82.74	55.01	42.62	70.14 ^(+32.92)
SPGD	10000	71.02*	96.48*	50.21	89.48	53.90	42.75	67.31
M-SPGD		73.35*	97.47*	52.18	91.42	56.19	44.07	69.11 ^(+1.80)
AT-UAP		82.33*	97.23*	56.83	91.22	58.58	46.81	72.17 ^(+4.86)
SGA		82.38*	97.30*	60.46	91.49	62.21	51.26	74.18 ^(+6.87)
M-SGA		80.48*	97.83*	60.23	92.46	61.86	50.51	73.89 ^(+6.58)
Ours		91.35*	96.91*	66.34	91.33	63.88	51.48	76.88 ^(+9.57)

Table 2: The fooling ratio (%) in the ensemble-model setting by different UAP generation methods. The UAPs are crafted on the ensemble models, *i.e.*, AlexNet and VGG16.

Black-box Attack. We evaluate transfer attacks with 10,000 and 500 training images. UAPs are generated for five considered models, and AEs are transferred between them. As shown in Tab. 3, DM-UAP outperforms others across all models. Compared to AT-UAP and SGA, DM-UAP consistently achieves the highest average fooling ratio improvement over SPGD for all surrogate models, while AT-UAP and SGA are sometimes inferior to SPGD. With 10,000 images, DM-UAP’s average fooling ratio improvement over SPGD ranges from 2.76% to 11.15%. With 500 images, DM-UAP’s improvement ranges from 8.22% to 45.19%, compared to AT-UAP’s -1.51% to 38.00% and SGA’s 0.77% to 22.55%. DM-UAP outperforms SGA with an average increase in fooling ratio of 12.108%. This manifests the importance of a dynamic model landscape proposed in our formulation in the limited samples scenario.

4.2 Scalability Performance of UAPs

In this subsection, we analyze the scalability performance of the proposed method from various aspects, including ensemble model setting, diverse sample setting, Transformer-to-CNN setting, and the attack-under-defense setting.

Ensemble-Model Setting. Following (Liu et al. 2023), we implement the model ensemble method using the averaged loss functions of two models, *i.e.*, AlexNet and VGG16. In this experiment, we still use SPGD as the baseline method, and compare the improvement of DM-UAP with those of others. For more comparison, we also integrate SPGD and SGA with the momentum (Dong et al. 2018) method, denoted as M-SPGD and M-SGA. In addition to the commonly used five models, we also test the transferability on ResNet50. The results are reported in Sec. 4.1. DM-UAP still outstrips the others by a clear superiority, which verifies the effectiveness of our dynamic maximin formulation still suffices in such a more complex optimization landscape.

Diverse-Sample Setting. Furthermore, we investigate the impact of varying the number of training samples on the attack performance. The results are depicted in Fig. 2. Our method consistently achieves the best performance with any number of training samples. However, we observe that once the number of samples exceeds a certain threshold, the corresponding increase in attack performance plateaus. The phenomenon underscores the limitations of excessive increase

Model	Method	10,000 samples						500 samples					
		AlexNet	GoogleNet	VGG16	VGG19	ResNet152	Average	AlexNet	GoogleNet	VGG16	VGG19	ResNet152	Average
AlexNet	UAP	84.28*	31.12	39.17	37.24	22.11	42.78	48.67*	18.13	23.07	22.59	14.28	25.35
	GAP	88.55*	35.52	52.85	49.22	28.81	50.99	83.96*	32.99	48.26	45.17	26.98	47.47
	SPGD	96.30*	54.16	61.39	58.89	36.78	61.50	89.20*	40.79	51.15	49.53	28.05	51.74
	AT-UAP	96.74*	48.86	62.23	58.80	33.36	60.00 _(-1.50)	93.80*	31.78	51.98	48.57	25.00	50.23 _(-1.51)
	SGA	96.99*	46.62	65.57	59.81	34.30	60.66 _(-0.84)	94.91*	34.88	55.27	50.85	26.64	52.51 _(+0.77)
	Ours	97.19*	53.95	68.20	63.03	38.91	64.26 _(+2.76)	96.13*	48.11	63.28	58.38	33.92	59.96 _(+8.22)
GoogleNet	UAP	39.55	55.01*	49.82	49.11	29.66	44.63	23.83	15.91*	19.45	18.94	11.78	17.98
	GAP	53.31	80.21*	72.56	70.62	49.85	65.31	37.59	33.32*	35.63	35.41	20.81	32.55
	SPGD	50.47	86.09*	66.79	65.93	44.66	62.79	31.28	55.35*	29.50	28.69	17.89	32.54
	AT-UAP	54.13	92.55*	79.13	76.54	53.51	71.17 _(+8.38)	49.49	80.63*	65.03	63.23	41.14	59.90 _(+27.36)
	SGA	59.22	88.46*	77.10	74.80	54.97	70.91 _(+8.12)	50.77	68.10*	60.18	59.13	37.28	55.09 _(+22.55)
	Ours	57.62	93.28*	81.45	80.05	57.29	73.94 _(+11.15)	57.29	88.07*	76.88	74.57	51.43	69.65 _(+37.11)
VGG16	UAP	33.33	36.19	75.36*	64.09	31.33	48.06	22.87	14.97	26.67*	22.84	12.72	20.01
	GAP	35.90	50.05	84.59*	76.49	38.97	57.20	34.65	26.39	57.82*	44.69	21.39	37.00
	SPGD	40.67	43.46	92.81*	83.18	44.44	60.91	34.57	24.10	73.10*	56.04	21.06	41.77
	AT-UAP	43.72	42.39	96.68*	88.13	34.97	61.18 _(+0.27)	43.34	33.48	90.13*	75.06	28.84	54.17 _(+12.40)
	SGA	44.48	52.53	97.83*	92.36	48.30	67.10 _(+6.19)	42.85	41.96	92.64*	80.83	36.24	58.90 _(+17.13)
	Ours	49.35	57.18	98.43*	94.12	49.97	69.81 _(+8.90)	45.99	49.96	96.78*	89.03	43.57	65.07 _(+23.30)
VGG19	UAP	33.98	36.62	64.57	74.77*	30.48	48.08	23.59	15.26	25.13	26.09*	12.77	20.57
	GAP	45.23	43.69	73.89	82.31*	30.02	55.03	38.86	30.36	52.12	60.83*	23.19	41.07
	SPGD	40.76	47.84	84.26	92.90*	45.43	62.24	33.38	22.30	52.50	65.49*	19.79	38.69
	AT-UAP	45.42	43.72	90.61	95.47*	40.05	63.05 _(+0.81)	42.82	34.20	77.99	88.02*	29.89	54.58 _(+15.89)
	SGA	45.89	55.53	92.62	96.56*	49.90	68.10 _(+5.86)	42.70	42.90	84.42	90.88*	36.69	59.52 _(+20.83)
	Ours	49.81	59.91	95.51	97.81*	55.75	71.76 _(+9.52)	47.24	50.97	91.33	96.55*	41.57	65.53 _(+26.84)
ResNet152	UAP	35.15	36.04	49.19	47.50	57.36*	45.05	27.28	18.06	25.02	24.10	17.97*	22.49
	GAP	47.70	56.08	68.63	66.51	73.80*	62.54	39.60	38.23	46.90	46.24	46.66*	43.53
	SPGD	46.12	55.01	77.33	74.19	90.33*	68.60	30.23	19.01	27.33	26.40	18.87*	24.37
	AT-UAP	47.69	57.94	77.88	75.09	92.07*	70.13 _(+1.53)	44.67	47.62	71.56	67.51	80.49*	62.37 _(+38.00)
	SGA	49.30	62.40	80.53	78.18	92.86*	72.65 _(+4.05)	40.53	35.57	49.45	47.51	43.01*	43.21 _(+18.84)
	Ours	50.09	63.50	81.49	79.10	92.90*	73.42 _(+4.82)	51.26	59.17	78.05	75.40	83.92*	69.56 _(+45.19)

Table 3: The fooling ratio (%) on five models in the black-box setting by different UAP attack methods. The UAPs are crafted on AlexNet, GoogleNet, VGG16, VGG19, and ResNet152, respectively. We conduct the evaluation using 10,000 and 500 images for training, respectively. The respective last column represents the average improvement or deterioration.

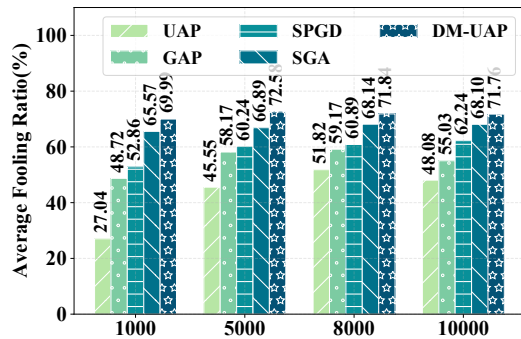


Figure 2: Average fooling ratio (%) on five models in the diverse-sample training scenarios, using VGG19.

in the number of training samples.

Transformer-to-CNN Setting. Unlike CNN models, transformer models employ self-attention mechanisms rather than convolutional blocks. Studies indicate that transformer models demonstrate reduced cross-model transferability to other architectures, particularly CNN. We adapt transformer models from DeiT (Touvron et al. 2021) and ViT (Dosovitskiy et al. 2020) family as surrogate models and evaluate the cross-model performance of different UAP methods. We can see from Sec. 4.2, that compared to SPGD, AT-UAP and SGA, our method consistently achieves the highest fooling ratio against all black-box CNN models.

Model	UAP Method	ANet	GNet	V16	V19	R152	Avg
ViT-B	SPGD	51.65	30.10	49.39	47.66	23.86	40.53
	SGA	54.31	37.86	54.18	52.26	28.93	45.51 _(+4.98)
	AT-UAP	56.24	44.88	55.47	54.19	32.31	48.62 _(+8.09)
	Ours	59.01	51.63	62.35	61.62	38.31	54.58 _(+14.05)
ViT-L	SPGD	35.94	24.93	36.01	34.83	18.55	30.05
	SGA	43.34	34.69	47.18	46.35	24.60	39.23 _(+9.18)
	AT-UAP	51.38	40.14	53.73	51.47	28.63	45.07 _(+15.02)
	Ours	54.94	46.55	59.50	56.26	32.52	49.95 _(+19.90)
DeiT-S	SPGD	34.38	21.76	41.09	37.47	19.53	30.85
	SGA	48.35	40.69	49.63	47.43	28.37	42.89 _(+12.04)
	AT-UAP	54.29	46.59	57.36	55.11	32.21	49.11 _(+18.26)
	Ours	56.60	50.09	62.00	59.64	37.35	53.14 _(+22.29)
DeiT-B	SPGD	39.66	26.72	40.02	38.04	19.66	32.82
	SGA	43.30	27.81	44.65	42.52	21.18	35.89 _(+3.07)
	AT-UAP	41.12	27.37	48.41	44.44	21.74	36.62 _(+3.80)
	Ours	46.32	29.56	51.26	47.12	23.22	39.50 _(+6.68)

Table 4: Fooling ratio (%) in the **Transformer-to-CNN** setting using different generation methods.

4.3 Ablation Study

We conduct ablation studies to evaluate the impact of curriculum learning, the model and data optimization order, and optimization neighborhood sizes on the performance of our proposed framework, using 500 training images by default.

On the curriculum UAP learning. We examine the impact of curriculum learning by gradually increasing the model neighborhood size ρ from 0 to 4 and the data neighborhood size r from 0 to 32 throughout the training epochs.

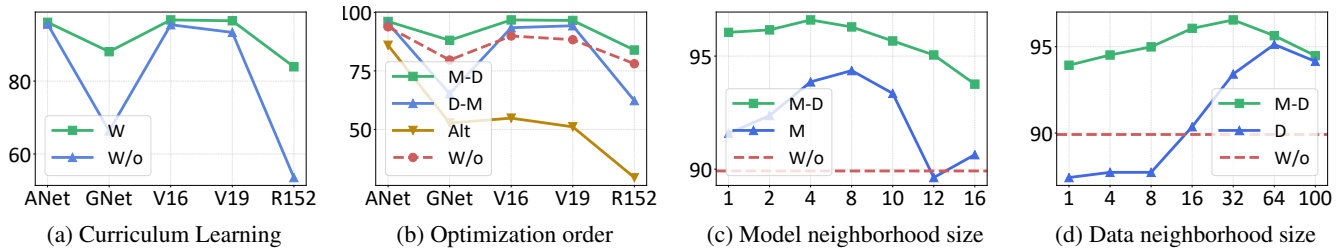


Figure 3: Ablation study on model and data optimization. “M” means “Model Optimization” and “D” means “Data Optimization”, “M-D” denotes “Model-Data Optimization”, “D-M” denote “Data-Model Optimization”. (a) Fooling ratios of DM-UAP with/without curriculum learning, *i.e.*, increasing neighborhood sizes. (b) Fooling ratio by different optimization orders in white-box setting for five models. (c) Fooling ratios of DM-UAP with/without data optimization for different model neighborhood sizes. (d) Fooling ratios of DM-UAP with/without model optimization for different data neighborhood sizes.

For comparison, fixed neighborhood sizes of $\rho = 4$ and $r = 32$ are also used. As shown in Fig. 3a, the results underscore the importance of using an easy-to-hard antagonistic approach as in curriculum learning for optimizing UAP, which significantly enhances its effectiveness.

On the optimization order of model and data. We explore the influence of different optimization orders to highlight the “optimize model first, data second” sequencing adopted in DM-UAP. In Fig. 3b, we use “Model-Data Optimization” to denote the process of 10 model optimization steps followed by 10 data optimization steps as in DM-UAP. In contrast, “Data-Model Optimization” reverses this order. “Alternating Optimization” means alternating single optimization steps between the model and data 20 times. “Without Optimization” means both model and data are not optimized, which reduces to the formulation of Eq. (2). The results show “Model-Data Optimization” consistently outperforms the “Without Optimization” baseline. “Data-Model Optimization” sometimes, while “Alternating Optimization” consistently underperforms. These confirm our hypothesis that the model should be optimized using unperturbed data.

On the model neighborhood size. We explore the influence of the maximum model neighborhood size, with data neighborhood size set as $r = 32$ for “Model-Data Optimization”. We tested ρ values of 1, 2, 4, 8, 10, 12, and 16, as depicted in Fig. 3c. For comparison, the formulation of optimal-parameters UAP maximin in Eq. (4) is also assessed, denoted as “Model Optimization”. The results indicate that “Model-Data Optimization” consistently surpasses “Model Optimization”, and their optimal ρ value differ.

On the data neighborhood size. Similarly, we explore the influence of the maximum data neighborhood size r , with the model neighborhood size set as $\rho = 4$ for “Model-Data Optimization”. We test r values of 1, 4, 8, 16, 32, 64, and 100, illustrated in Fig. 3d. “Data Optimization”, the formulation of optimal-data UAP maximin in Eq. (3) is also examined by varying r . The results also demonstrate that “Model-Data Optimization” invariably outperforms “Data Optimization”, their optimal r value differ.

5 Limitation and Future Work

Our DM-UAP introduces a novel maximin formulation to enhance the generalization of UAP by dynamically optimizing both the model and data during the UAP generation process. As a result, this incurs additional computational expenses. Specifically, for crafting UAPs on VGG16, the time expense for DM-UAP is approximately 1.6 times that of SGA and twice that of AT-UAP (see Tab. 5). However, considering the universal nature of UAP, these computational costs can be deemed negligible. Once the UAPs are created, there is no need for additional computations, as the off-the-shelf UAPs can be readily utilized to generate AEs at scale. For future studies on crafting UAPs against large models that require more memory, we think adaptively selecting part of the parameters for optimization may be worth exploring.

Method	Time consumption	Memory consumption	Average attack success rate(%)
SPGD	27min’58s	19620 MiB	60.91
AT-UAP	1h’12min’34s	20210 MiB	61.18
SGA	1h’28min’44s	10382 MiB	67.10
Ours	2h’21min’26s	20714 MiB	69.81

Table 5: The optimization expenses of different UAP methods. UAPs are obtained from the VGG16 model.

6 Conclusion

In this paper, we present DM-UAP, a novel approach for generating UAPs with a dynamic maximin optimization strategy. DM-UAP not only optimizes the data used for training but also takes into account dynamic model parameters. Specifically, DM-UAP incorporates an iterative maximin-min optimization framework that dynamically minimizes classification loss to obtain model-data pairs, as well as a curriculum learning algorithm to thoroughly explore the combined landscape of model parameters and data. Extensive experiments on ImageNet demonstrate the superior performance of DM-UAP over state-of-the-art methods, significantly improving the generalization of generated UAPs.

Acknowledgements

Minghui’s work is supported by the National Natural Science Foundation of China (Grant No.62202186). Shengshan’s work is supported by the National Natural Science Foundation of China (Grant No.62372196). Minghui Li is the corresponding author.

References

- Bengio, Y.; Louradour, J.; Collobert, R.; and Weston, J. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, 41–48.
- Benz, P.; Zhang, C.; Imtiaz, T.; and Kweon, I. S. 2020. Double targeted universal adversarial perturbations. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*.
- Co, K. T.; Muñoz-González, L.; Kanthan, L.; Glocker, B.; and Lupu, E. C. 2021. Universal adversarial robustness of texture and shape-biased models. In *2021 IEEE International Conference on Image Processing (ICIP)*, 799–803. IEEE.
- Ding, W.; Wei, X.; Ji, R.; Hong, X.; Tian, Q.; and Gong, Y. 2021. Beyond universal person re-identification attack. *IEEE transactions on information forensics and security*, 16: 3442–3455.
- Dong, Y.; Deng, Z.; Pang, T.; Zhu, J.; and Su, H. 2020. Adversarial distributional training for robust deep learning. *Advances in Neural Information Processing Systems*, 33: 8270–8283.
- Dong, Y.; Liao, F.; Pang, T.; Su, H.; Zhu, J.; Hu, X.; and Li, J. 2018. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 9185–9193.
- Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.
- Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.
- He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.
- Hu, S.; Liu, X.; Zhang, Y.; Li, M.; Zhang, L. Y.; Jin, H.; and Wu, L. 2022. Protecting Facial Privacy: Generating Adversarial Identity Masks via Style-Robust Makeup Transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15014–15023.
- Hu, S.; Zhang, Y.; Liu, X.; Zhang, L. Y.; Li, M.; and Jin, H. 2021. AdvHash: Set-to-set Targeted Attack on Deep Hashing with One Single Adversarial Patch. In *Proceedings of the 29th ACM International Conference on Multimedia, MM ’21*, 2335–2343. New York, NY, USA: Association for Computing Machinery. ISBN 9781450386517.
- Khrulkov, V.; and Oseledets, I. 2018. Art of singular vectors and universal adversarial perturbations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8562–8570.
- Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.
- Li, M.; Wang, J.; Zhang, H.; Zhou, Z.; Hu, S.; and Pei, X. 2024. Transferable Adversarial Facial Images for Privacy Protection. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM’24)*.
- Li, M.; Yang, Y.; Wei, K.; Yang, X.; and Huang, H. 2022. Learning universal adversarial perturbation by adversarial example. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 1350–1358.
- Liu, X.; Zhong, Y.; Zhang, Y.; Qin, L.; and Deng, W. 2023. Enhancing Generalization of Universal Adversarial Perturbation through Gradient Aggregation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4435–4444.
- Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.
- Moosavi-Dezfooli, S.-M.; Fawzi, A.; Fawzi, O.; and Frossard, P. 2017. Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR’17)*, 1765–1773.
- Mopuri, K. R.; Ojha, U.; Garg, U.; and Babu, R. V. 2018. Nag: Network for adversary generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 742–751.
- Poursaeed, O.; Katsman, I.; Gao, B.; and Belongie, S. 2018. Generative adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4422–4431.
- Shafahi, A.; Najibi, M.; Xu, Z.; Dickerson, J.; Davis, L. S.; and Goldstein, T. 2020. Universal adversarial training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 5636–5643.
- Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
- Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; and Fergus, R. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.
- Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I. J.; and Fergus, R. 2014. Intriguing properties of neural networks. In Bengio, Y.; and LeCun, Y., eds., *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.
- Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; and Jégou, H. 2021. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, 10347–10357. PMLR.
- Wang, D.; Yao, W.; Jiang, T.; and Chen, X. 2023. Improving Transferability of Universal Adversarial Perturbation with

Feature Disruption. *IEEE Transactions on Image Processing*.

Xie, C.; Zhang, Z.; Zhou, Y.; Bai, S.; Wang, J.; Ren, Z.; and Yuille, A. L. 2019. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2730–2739.

Xie, Y.; Li, Z.; Shi, C.; Liu, J.; Chen, Y.; and Yuan, B. 2021. Enabling fast and universal audio adversarial attack using generative model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 14129–14137.

Xu, X.; Zhang, J. Y.; Ma, E.; Son, H. H.; Koyejo, S.; and Li, B. 2022. Adversarially robust models may not transfer better: Sufficient conditions for domain transferability from the view of regularization. In *International Conference on Machine Learning*, 24770–24802. PMLR.

Zhang, C.; Benz, P.; Imtiaz, T.; and Kweon, I.-S. 2020a. Cd-uap: Class discriminative universal adversarial perturbation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 6754–6761.

Zhang, C.; Benz, P.; Imtiaz, T.; and Kweon, I. S. 2020b. Understanding adversarial examples from the mutual influence of images and perturbations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14521–14530.

Zhang, C.; Benz, P.; Karjauv, A.; and Kweon, I. S. 2021. Data-free universal adversarial perturbation and black-box attack. In *Proceedings of the IEEE/CVF international conference on computer vision*, 7868–7877.

Zhang, Y.; Hu, S.; Zhang, L. Y.; Shi, J.; Li, M.; Liu, X.; Wan, W.; and Jin, H. 2024. Why Does Little Robustness Help? A Further Step Towards Understanding Adversarial Transferability. In *2024 IEEE Symposium on Security and Privacy (SP)*, 3365–3384.

Zhao, Z.; Liu, Z.; and Larson, M. 2021. On Success and Simplicity: A Second Look at Transferable Targeted Attacks. In *Advances in Neural Information Processing Systems*.

Zhong, Y.; and Deng, W. 2022. Opom: Customized invisible cloak towards face privacy protection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3): 3590–3603.

Zhou, Z.; Hu, S.; Li, M.; Zhang, H.; Zhang, Y.; and Jin, H. 2023a. AdvCLIP: Downstream-agnostic Adversarial Examples in Multimodal Contrastive Learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, 6311–6320. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701085.

Zhou, Z.; Hu, S.; Zhao, R.; Wang, Q.; Zhang, L. Y.; Hou, J.; and Jin, H. 2023b. Downstream-agnostic adversarial examples. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV'23)*, 4345–4355.

Zhou, Z.; Song, Y.; Li, M.; Hu, S.; Wang, X.; Zhang, L. Y.; Yao, D.; and Jin, H. 2024. Darksam: Fooling segment anything model to segment nothing. In *Proceedings of the 38th Annual Conference on Neural Information Processing Systems (NeurIPS'24)*.