TRM-UAP: Enhancing the Transferability of Data-Free Universal Adversarial Perturbation via Truncated Ratio Maximization

Yiran Liu, Xin Feng, Yunlong Wang, Wu Yang, Di Ming*
School of Computer Science and Engineering, Chongqing University of Technology
Chongqing, China
lyr199804@qq.com, {xfeng, ylwang, yangwu, diming}@cqut.edu.cn

Abstract

Aiming at crafting a single universal adversarial perturbation (UAP) to fool CNN models for various data samples, universal attack enables a more efficient and accurate evaluation for the robustness of CNN models. Early universal attacks craft UAPs depending on data priors. For more practical applications, the data-free universal attacks that make UAPs from random noises have aroused much attention recently. However, existing data-free UAP methods perturb all the CNN feature layers equally via the maximization of the CNN activation, leading to poor transferability. In this paper, we propose a novel data-free universal attack without depending on any real data samples through truncated ratio maximization, which we term as TRM-UAP. Specifically, different from the maximization of the positive activation in convolution layers, we propose to optimize the UAP generation from the ratio of positive and negative activations. To further enhance the transferability of universal attack, TRM-UAP not only performs the ratio maximization merely on low-level generic features via the truncation strategy, but also incorporates a curriculum optimization algorithm that can effectively learn the diversity of artificial images. Extensive experiments on the ImageNet dataset verify that TRM-UAP achieves a state-of-the-art average fooling rate and excellent transferability on different CNN models as compared to other data-free UAP methods. Code is available at https://github.com/RandolphCarter0/TRMUAP.

1. Introduction

Early research [26] shows that tiny and imperceptible perturbations can seriously disturb the prediction results of deep neural networks (DNNs), especially for image recognition tasks [3]. Adversarial Examples (AEs), crafted by adding tiny perturbations deliberately to benign samples, are not only imperceptible in computer vision tasks but also prone to transfer among the DNN models. Therefore, AEs have been regarded as a serious threat to DNN models since the development of deep learning [1].

To explore the impact of AEs, many methods are proposed to craft the adversarial perturbation that is highly transferable to various DNN models (*i.e.*, with a high fooling rate on other DNN models). However, the AEs generated by these works are explicitly designed for some specific samples and often fail to perturb other samples that are even from the same dataset. Different from the aforementioned image-specific attack, the universal attack that generates image-agnostic universal adversarial perturbation (UAP) was proposed in [16]. The UAP in a universal attack setting is trained from prior knowledge, such as substitute data, surrogate model, *etc*. By adding the UAP to benign samples, universal attacks can generate numerous AEs all together in a very short period. Furthermore, the universal attack can fool most DNN models that are trained from similar datasets, and can greatly reduce the computational cost of crafting AEs, making adversaries more applicable to real scenarios than the image-specific attacks [4, 6, 34].

Whereas, no matter the image-specific attacks or the universal attacks, well-annotated training data (*e.g.*, [21, 9, 7]) or substitute data (*e.g.*, [16, 12, 22, 19]) is required to generate AEs. In practice, obtaining a well-labeled large-scale dataset is challenging and costly, especially for some applications with critical security demand, where less prior knowledge is available. Recently, researchers have investigated data-free universal attack methods [18, 17, 14, 20, 33], where AEs are generated directly from random noises rather than data priors. Compared with previous methods that disturb the gradient or maximize the classification loss in data-dependent scenarios, current data-free UAP methods explore the feature-based perturbation, which tries to maximize the activation (*e.g.*, ReLU) of convolutional neural network (CNN) features. It is shown that feature-based UAP methods achieved highly efficient and applicable universal attack without using any data priors [4, 6, 34]. How-

*Corresponding Author: Di Ming

ever, current data-free UAP methods only consider the positive activation [18, 17] and perturb all layers of CNN features equally [20, 14]. As a result, UAPs crafted by surrogate models are hard to transfer to target models [6, 34].

Towards enhancing the transferability of data-free UAPs, we propose a novel data-free universal attack method called TRM-UAP, which formulates the UAP generation as a truncated ratio maximization problem. In particular, TRM-UAP attempts to over-fire positive neurons so that extracted features from multiple CNN layers can be fully disrupted. Besides the maximization of positive neurons, the proper activation on negative neurons may also be helpful. Hence, TRM-UAP performs a ratio maximization of positive and negative activations, which facilitates some negative neurons transiting to be positive. Moreover, some feature-based adversarial attacks [29, 35] reveal that multiple layers of CNN features are not equally important for disturbing the image classification [11, 28]. To further improve the transferability of universal attack, TRM-UAP performs ratio maximization only on parts of CNN activation layers. Specifically, a truncation strategy is proposed to truncate the activation of high-level convolutions and retain the activation of shallow convolutions to train perturbations. The intuition is that low-level CNN features are shown to provide more generic (data-independent) feature representations than high-level semantic features [30]. Thus the disturbed low-level features are expected to be able to transfer attacks across different CNN models, leading to highly transferable UAPs. Furthermore, a curriculum learning [2] based optimization algorithm is introduced to utilize artificial images and explore the diversity of input.

In general, the main contributions of our work can be summarized as follows:

- We propose a novel data-free universal attack method to craft image-agnostic adversarial perturbations without utilizing any real data samples during training.
- To the best of our knowledge, we are the first to formulate the data-free universal attack as a truncated ratio maximization problem. The proposed TRM-UAP enhances the transferability of UAPs from three perspectives: (i) the maximization on CNN features for crafting UAPs is enhanced by both maximization of positive activation and minimization of negative activation. (ii) the ratio maximization on truncated feature layers improves the generalizability of universal attacks to CNN models. (iii) artificial images are incorporated into curriculum optimization algorithm to strengthen the UAP dominance on feature activations.
- Extensive experiments on the ImageNet dataset show that TRM-UAP obtains the highest average fooling rate compared with other data-free UAP methods, indicating that truncated ratio maximization can greatly improve the transferability of universal attack.

2. Related Work

Image-specific attacks: The attacks that try to craft perturbations for corresponding images by utilizing training or substitute data are termed as image-specific attacks. Typical image-specific attacks, such as Fast Gradient Sign Method (FGSM) [9], I-FGSM and MI-FGSM [7], Project Gradient Descent (PGD) [15], and Output Diversified Sampling (ODS) [27], *etc*., usually make use of the gradient information from well-trained target CNN model (*i.e*., white-box attack) or surrogate model (*i.e*., black-box attack) to disturb images. However, since gradient-based attacks often require a number of iterations, the computational cost of making AEs becomes extremely expensive as the number of examples increases. Recently, feature-based attacks that try to craft AEs from the perspective of CNN features have received broad concern [29, 8, 35]. For example, Neuron Attribution based Attack (NAA) [35] conducted feature-level attacks by deploying a neuron attribution method to estimate the importance of neurons at multiple CNN layers.

Data-dependent universal attacks: The universal attack aims at training a single perturbation that is independent of any specific sample, *i.e*., UAP. By adding a single UAP to various benign samples, numerous AEs can be generated very efficiently. Until recently, most of the universal attacks are still data-dependent, meaning that the perturbations are learned from data priors, such as training data, substitute data, *etc*. For example, the primary universal attack [16] tried to seek a universal perturbation for a set of training samples. On the other hand, there are some works [22, 19] applying generative adversarial models to craft perturbations. Recently, Zhang *et al*. [32] proposed to generate UAPs by training perturbations on the proxy datasets.

Data-free universal attacks: The data-free universal attack makes UAP via random noise rather than prior knowledge from data directly. It is considered the most applicable attack for real applications in the adversarial attack field. However, since the information of both data and target models is unknown, the data-free universal attack is extremely challenging and only a few recent works focus on it. Mopuri *et al*. [18] first propose a data-free universal attack, *i.e*., the Fast Feature Fool (FFF) method, by maximizing the feature activations across all the CNN layers. In addition, they further propose a Generalizable Data-free UAP (GD-UAP) to improve the FFF attack via optimizing the UAP training process with a saturation check strategy [17]. Besides, AAA [20] learns a generative model to make UAPs based on the pre-defined class impression. However, as the number of image categories increases, the cost of acquiring class impressions will also broaden. Prior-Driven Uncertainty Approximation (PD-UA) [14] proposes to craft UAPs by maximizing the uncertainty approximation of the model. Cosine-UAP [33] introduces the cosine similarity to train UAPs in a self-supervised way. Because of the shortage of

prior knowledge, current data-free UAP methods generally present poor transferability across various CNN models.

3. Methodology

In this section, we first introduce the motivation of maximizing the ratio of activations. After that, we present the details of the proposed truncated ratio maximization method to craft the UAP in a strictly data-free fashion.

3.1. Preliminary

To fool CNN models f, UAP [16] tries to find the universal perturbation $\boldsymbol{v}$ which maximizes the classification loss $\mathcal{L}$ (*e.g.*, cross entropy) over the data distribution $\mathbb{D}$:

$$\max_{\boldsymbol{v}\sim\mathbb{S}} \mathop{\mathbb{E}}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathbb{D}} [\mathcal{L}(f(\boldsymbol{v}+\boldsymbol{x}),\boldsymbol{y}))] \qquad \text{s.t. } \|\boldsymbol{v}\|_p \leq \epsilon \quad (1)$$

where $\mathbb{S}$ is the perturbation space, $(\boldsymbol{x},\boldsymbol{y})$ are the data and the class label (or one-hot ground truth) sampled from $\mathbb{D}$, $f(\cdot)$ is the output of CNN models, and ϵ is the constant to constrain the perturbation $\boldsymbol{v}$ in the ℓ_p-norm bound (*e.g.*, $p = 0, 1, 2, \infty$). To be consistent with previous works [18, 17, 32], we adopt the ℓ_∞-norm to realize the imperceptibility of perturbation for all analyzes throughout this paper.

However, the training data of the target model is usually unavailable and hard to access, making the data-dependence approaches not suitable for practical adversarial attacks. As a consequence, Mopuri *et al*. [18] propose the Fast Feature Fool (FFF) method to maximize the CNN activations in a strictly data-free fashion:

$$\begin{aligned} \max_{\boldsymbol{v}} \quad & ||\mathcal{A}^{(i)}(\boldsymbol{v})||_2, \qquad \text{for } i = 1, 2, \cdots, L \\ \text{s.t.} \quad & ||\boldsymbol{v}||_\infty \leq \epsilon \end{aligned} \quad (2)$$

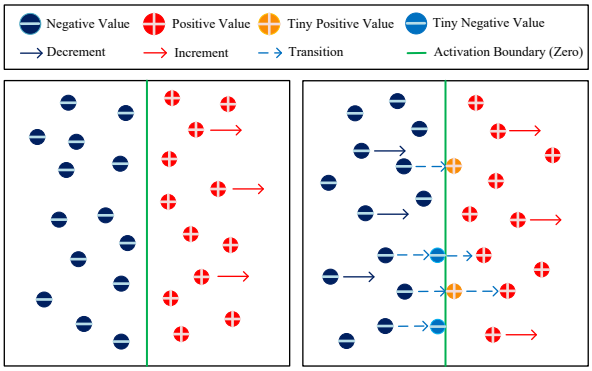
where $\mathcal{A}^{(i)}(\boldsymbol{v}) = \text{Activation}(\mathcal{C}^{(i)}(\boldsymbol{v}))$ (*e.g.*, ReLU activation), $\mathcal{C}^{(i)}(\boldsymbol{v})$ is the output of the i-th convolution layer, and L is the total number of convolution layers. It can be seen that this typical data-free universal attack method aims at accumulating the errors gradually through multiple convolution layers to enlarge the classification loss, finally leading to the misclassification of perturbed samples $\boldsymbol{v} + \boldsymbol{x}$.

3.2. Maximizing the Ratio of Activations

Different from previous data-free universal methods [18, 17] that only maximize the activation of convolution layer (see Fig. 1 (a)), we propose to craft the universal adversarial perturbation $\boldsymbol{v}$ via maximizing the ratio of activations:

$$\begin{aligned} \max_{\boldsymbol{v}} \quad & \frac{||\mathcal{C}_+^{(i)}(\boldsymbol{v})||_2}{||\mathcal{C}_-^{(i)}(\boldsymbol{v})||_2}, \qquad \text{for } i = 1, 2, \cdots, L \\ \text{s.t.} \quad & ||\boldsymbol{v}||_\infty \leq \epsilon \end{aligned} \quad (3)$$

where positive activation and negative activation in i-th convolution layer are defined as $\mathcal{C}_+^{(i)}(\boldsymbol{v}) = \max(\mathcal{C}^{(i)}(\boldsymbol{v}), 0)$ and $\mathcal{C}_-^{(i)}(\boldsymbol{v}) = \min(\mathcal{C}^{(i)}(\boldsymbol{v}), 0)$ respectively, and their magnitudes are measured by ℓ_2-norm.



(a) Positive Maximization (b) Ratio Maximization

Figure 1. The illustration of previous positive maximization [18, 17] and our proposed ratio maximization.

Towards further enlarging the fooling probability, this ratio formulation (3) tries to increase the magnitude and the number of positive activation via maximizing positive activation and minimizing negative activation simultaneously. Specifically speaking, ratio maximization is a dynamic process. Firstly, minimizing the negative activation could lead to the change of sign in some negative features whose values are close to zero, *i.e.*, jumping from a tiny negative value to a tiny positive value, as shown in Fig. 1 (b). Afterward, at the next iteration, maximizing the positive activation will continue to increase the magnitude of above-mentioned features from a tiny positive value to a large positive value.

During training, with the decrease in the number of negative activation and the increase in the number of positive activation, our proposed formulation (3) can achieve a higher ratio of positive activation as compared to previous data-free universal attacks, *e.g.*, the formulation (2). Meanwhile, it also indicates that our ratio maximization could craft the UAP with more significant attack intensity so as to increase the fooling probability further.

3.3. Truncated Ratio Maximization

Features extracted from shallow to deep convolution layers behave differently in crafting the adversarial perturbation. LAFEAT [31] analyzes the influence of each intermediate convolution layer on the misclassification, and attacks features only in a specific layer. On the other hand, GD-UAP [17] thoroughly studied the relative change in feature activations of adversarial examples as compared to original images at multiple CNN layers. However, GD-UAP attacks the feature activations extracted from all convolution layers together, without taking layer-wise feature patterns into consideration. According to [30], low-level features in shallow layers have generic patterns, while high-level features in deep layers have specific patterns. It can be seen that

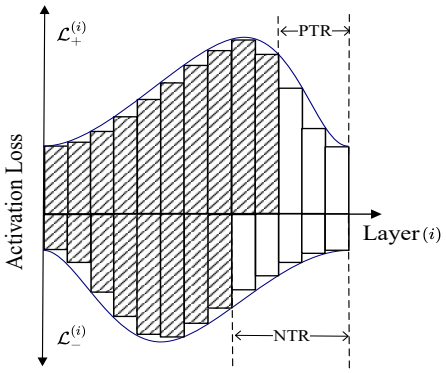


Figure 2. Truncated positive and negative activation losses, where cumulative shaded areas are used to craft the UAP.

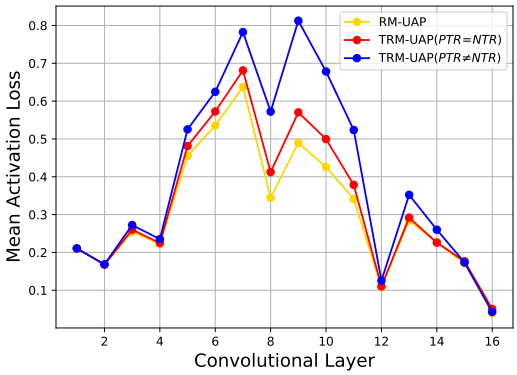


Figure 3. The mean positive activation loss in various convolution layers with different truncation settings.

not all convolution layers contribute positively to increasing the positive activation and the objective loss. Therefore, based on ratio maximization defined in formulation (3), we propose a truncated ratio maximization method to further enhance the intensity and the transferability of UAPs.

For notational simplicity, we begin by defining the ratio maximization problem in i-th convolution layer as $\mathcal{L}^{(i)}(\boldsymbol{v}) = \mathcal{L}_+^{(i)}(\boldsymbol{v})/\mathcal{L}_-^{(i)}(\boldsymbol{v})$, where $\mathcal{L}_+^{(i)}(\boldsymbol{v}) = ||\mathcal{C}_+^{(i)}(\boldsymbol{v})||_2$ and $\mathcal{L}_-^{(i)}(\boldsymbol{v}) = ||\mathcal{C}_-^{(i)}(\boldsymbol{v})||_2$. To maximize the ratio of activations in convolution layers altogether, the overall loss function of ratio maximization is reformulated as

$$\mathcal{L}(\boldsymbol{v}) = \sum_{i=1}^{L} \log \mathcal{L}^{(i)}(\boldsymbol{v}) \tag{4}$$

where log rescales the activation to prevent gradient explosions. Different from other methods that craft UAPs based on all convolution layers or a specific convolution layer, we propose to truncate the computation of positive and negative activations in deep convolution layers, since specific feature patterns are not beneficial for improving the quality of UAPs. Assuming that the truncation rate (TR) is defined as $TR = \lfloor (L-l)/L \rfloor \%$, we make truncation on $(l+1)$-th convolution layer by seting $\mathcal{L}_+^{(i)}(\boldsymbol{v}) = \tau \ (i > l)$ and $\mathcal{L}_-^{(i)}(\boldsymbol{v}) = \tau \ (i > l)$, where τ is a small positive number such as $1e-9$. As a result, $\mathcal{L}(\boldsymbol{v})$ can be equivalently rewritten as $\mathcal{L}(\boldsymbol{v}) = \sum_{i=1}^{L} \log \mathcal{L}^{(i)}(\boldsymbol{v}) = \sum_{i=1}^{l} \log \mathcal{L}^{(i)}(\boldsymbol{v})$.

However, $\mathcal{L}_+^{(i)}(\boldsymbol{v})$ and $\mathcal{L}_-^{(i)}(\boldsymbol{v})$ not only have different scales but also behave differently in various layers for maximizing the ratio loss, which could lead to a degradation of the attack performance. To resolve these problems, we first reformulate the ratio loss in i-th convolution layer as

$$\mathcal{L}_\alpha^{(i)}(\boldsymbol{v}) = \frac{\mathcal{L}_+^{(i)}(\boldsymbol{v})}{(\mathcal{L}_-^{(i)}(\boldsymbol{v}))^\alpha}, \tag{5}$$

where α is a scaling hyperparameter to adjust the relative importance between positive and negative activation losses. Secondly, positive and negative truncation rates are defined as $PTR = \lfloor (L-l')/L \rfloor \%$ and $NTR = \lfloor (L-l'')/L \rfloor \%$ respectively. That is to say, we make truncation on different layers for positive and negative activation losses by setting $\mathcal{L}_+^{(i)}(\boldsymbol{v}) = \tau \ (i > l')$ and $\mathcal{L}_-^{(i)}(\boldsymbol{v}) = \tau \ (i > l'')$, see Fig. 2. Finally, the overall loss function $\mathcal{L}(\boldsymbol{v})$ can be reduced to

$$\begin{aligned} \mathcal{L}(\boldsymbol{v}) &= \textstyle\sum_{i=1}^{L} \log \mathcal{L}_\alpha^{(i)}(\boldsymbol{v}) \\ &= \textstyle\sum_{i=1}^{l'} \log \mathcal{L}_+^{(i)}(\boldsymbol{v}) - \alpha \cdot \sum_{i=1}^{l''} \log \mathcal{L}_-^{(i)}(\boldsymbol{v}) + c \\ &\propto \textstyle\sum_{i=1}^{l'} \log \mathcal{L}_+^{(i)}(\boldsymbol{v}) - \alpha \cdot \sum_{i=1}^{l''} \log \mathcal{L}_-^{(i)}(\boldsymbol{v}), \end{aligned} \tag{6}$$

where $c = ((1-\alpha) \cdot L - l' + \alpha \cdot l'') \cdot \log \tau$, *i.e.*, a constant.

Consequently, for data-free universal adversarial attack, we propose a truncated ratio maximization method (TRM-UAP) to craft the perturbation $\boldsymbol{v}$ satisfying

$$\begin{aligned} \max_{\boldsymbol{v}} \quad & \sum_{i=1}^{l'} \log ||\mathcal{C}_+^{(i)}(\boldsymbol{v})||_2 - \alpha \cdot \sum_{i=1}^{l''} \log ||\mathcal{C}_-^{(i)}(\boldsymbol{v})||_2 \\ \text{s.t.} \quad & ||\boldsymbol{v}||_\infty \leq \epsilon \end{aligned} \tag{7}$$

where $\boldsymbol{v}$ is constrained by ℓ_∞-norm and bound ϵ. To verify the previous claim, we compute the mean positive activation loss in various convolution layers via feeding UAPs into CNN models (*e.g.*, VGG19). As shown in Fig. 3, TRM-UAP ($PTR \neq NTR$, *i.e.*, $l' \neq l'' < L$) generates a larger cumulative positive activation loss than RM-UAP ($l' = l'' = L$) and TRM-UAP ($PTR = NTR$, *i.e.*, $l' = l'' < L$), which could lead to a further increase in fooling rates.

3.4. Curriculum Optimization Algorithm

For purpose of improving the diversity of inputs, we utilize Gaussian noise and jigsaw image [33] to generate artificial images, which are fed into surrogate models together with the perturbation. To resemble training samples in real-world datasets, the mean filter is used to smooth the edge between different regions in jigsaw images. However, it

will be difficult for the optimization algorithm to converge if artificial images with complex patterns are provided at the beginning of the training procedure. Inspired by curriculum learning [2], we start the training with simple artificial images and then gradually feed more complex artificial images into surrogate models. Specifically, the generation of artificial images is controlled by the distribution parameter and the training iteration, defined as:

$$\begin{aligned} &D_1 \prec D_2 \prec \cdots \prec D_n, \\ &D_t = \{\boldsymbol{x} | \boldsymbol{x} \sim P(\theta_0, t)\}, \end{aligned} \tag{8}$$

where D_t is a set of artificial images $\boldsymbol{x}$ sampled at t-th iteration ($t = 1, 2, \cdots, n$), P is the distribution of artificial samples with the default parameter θ_0, and $D_a \prec D_b$ denotes that the pattern of artificial images in D_b is more complex than D_a. In detail, as the number of training iterations t increases, the distribution parameter θ_0 (*e.g.*, the standard deviation of Gaussian distribution and the frequency of jigsaw images) is gradually increasing. For t-th iteration, our curriculum optimization algorithm is maximizing

$$\mathcal{L}_t = \frac{1}{|D_t|} \sum_{\boldsymbol{x} \in D_t} \Big(\sum_{i=1}^{l'} \log \mathcal{L}_+^{(i)}(\boldsymbol{v}+\boldsymbol{x}) - \alpha \cdot \sum_{i=1}^{l''} \log \mathcal{L}_-^{(i)}(\boldsymbol{v}+\boldsymbol{x}) \Big) \tag{9}$$

where $|D_t|$ is the number of artificial images in the set D_t.

The entire procedure of our proposed TRM-UAP method is summarized in Algorithm 1. During the training process, the perturbation vector $\boldsymbol{v}$ is randomly initialized by the uniform distribution $\mathcal{U}$, and a surrogate dataset is used to validate the convergence of optimization for every H iterations. l' and l'' represent the retained layers of positive and negative activations corresponding to truncation rates PTR and NTR respectively. The scaling hyperparameter α adjusts the ratio between truncated positive and negative activation losses. The algorithm is converged only if either the iteration number reaches the maximum T or the fooling rate test reaches the threshold F_{max}. Similar to GD-UAP [17], we run a saturation rate test for each iteration to dynamically compress the perturbation vector only if the saturation rate $\hat{r}$ is smaller than the predefined threshold r.

3.5. Connection to Other Data-Free UAP Methods

When positive and negative truncation rates are set as $PTR = 0\%$ and $NTR = 100\%$ (*i.e.*, $l' = L, l'' = 0$), this means all the negative activations are truncated and all the positive activations are retained. Thus, the proposed TRM-UAP defined in Eq. (7) is reduced to GD-UAP [17]. It can be seen that GD-UAP is only a special case of TRM-UAP under the data-free universal attack setting. As compared to GD-UAP, our TRM-UAP improves both intensity and transferability of crafted UAPs via truncated ratio maximization. The detail of their differences will be further discussed in the experiments.

Algorithm 1 Curriculum optimization algorithm for solving truncated ratio maximization problem to craft universal adversarial perturbations

Input: surrogate CNN model f, limitation value ϵ, learning rate η, positive and negative truncation rates PTR, NTR, scaling hyperparameter α, maximum iteration number T, convergence threshold F_{max}, validation test hyperparameter H, saturation threshold r

Output: universal adversarial perturbation $\boldsymbol{v}$

Initialize $\boldsymbol{v}_0 \sim \mathcal{U}(-\epsilon, \epsilon)$, $t = 0$, $F = 0$
while $t < T$ *and* $F < F_{max}$ **do**
  $t = t + 1$
  Generate the artificial image set D_t via Eq. (8)
  Compute the gradient $\nabla \mathcal{L}_t$ of the loss $\mathcal{L}_t$ in Eq. (9)
  Update $\boldsymbol{v}_t = \boldsymbol{v}_{t-1} + \eta \cdot \nabla \mathcal{L}_t$
  Clip $\boldsymbol{v}_t = \min(\epsilon, \max(\boldsymbol{v}_t, -\epsilon))$
  Compute the saturation rate $\hat{r}$ and adjust $\boldsymbol{v}_t$ if $\hat{r} < r$
  Conduct the fooling rate test FR if $t\%H == 0$ and $F = F + 1$ if FR not the best fooling rate
end while
return $\boldsymbol{v}_t$

4. Experiments

Setup: By following the experiment setup in existing data-free universal methods [18, 17, 20, 14], we evaluated the proposed TRM-UAP method on the validation set of ImageNet [23] with the classical pre-trained CNN models including AlexNet [13], VGG16 [24], VGG19 [24], ResNet152 [10] and GoogleNet [25].

Evaluation Criteria: We used the fooling rate (FR) proposed by data-free universal methods [18, 17] as the evaluation metric. Particularly, a higher fooling rate represented higher attack success rate and better transferability. In addition, we used the logit loss of C&W attack [5] to further evaluate the transferability of data-free UAPs.

Comparative Methods: We made a comparison with data-free universal approaches, including FFF [18], AAA [20], GD-UAP [17], PD-UA [14] and Cosine-UAP [33]. Note that GD-UAP [17] trained models under different setups, for a fair comparison, we reproduced the results of GD-UAP [17] by making the setup consistent with TRM-UAP.

Implementation Details: All of our experiments were implemented on PyTorch with a single NVIDIA GeForce RTX 3090Ti GPU. Following the common setting [18, 17], we set $\epsilon = 10/255$ to restrict changes of the perturbation. The maximum iteration T was set as 10000, and the saturation threshold r was set to 0.001%. The value range of hyperparameters $PTR, NTR \in [0, 1]$ denotes the truncation rate of positive and negative activations respectively. We set the scaling hyperparameter α and truncation rates with appropriate values for different CNN models.

Attack	AlexNet	VGG16	VGG19	ResNet152	GoogleNet	Average
FFF	80.92	47.10	43.62	-	56.44	-
AAA	89.04	71.59	72.84	60.72	75.28	73.89
GD-UAP	85.24	90.01	87.34	45.96	45.87	64.65
PD-UA	-	70.69	64.98	46.39	67.12	-
Cosine-UAP	91.07	89.48	86.81	65.35	**87.57**	84.08
TRM-UAP(Ours)	**93.53±0.07**	**94.30±0.15**	**91.35±0.30**	**67.46±0.35**	85.32±0.04	**86.39**

Table 1. The fooling rate (FR) of our proposed TRM-UAP and other data-free universal methods. We show the mean and standard deviation of FR with five runs.

	AlexNet	VGG16	VGG19	ResNet152	GoogleNet
AlexNet	93.53±0.07	60.10±0.24	57.08±0.15	27.31±0.30	32.70±0.22
VGG16	47.53±0.51	94.30±0.12	89.68±0.14	61.43±0.40	53.95±0.59
VGG19	46.01±0.44	89.82±0.15	91.35±0.30	47.19±0.66	46.48±0.78
ResNet152	53.56±0.75	77.20±0.35	73.30±0.41	67.46±0.35	57.54±0.50
GoogleNet	60.10±1.16	79.66±0.95	79.98±1.06	58.85±1.94	85.32±0.04

Table 2. The transferability of the perturbation crafted by our TRM-UAP method (rows: target models; columns: surrogate models). The fooling rate (FR) on the diagonal represents white-box attacks (*i.e.*, UAP is crafted by the target model itself), and other off-diagnoal values are black-box attacks (*i.e.*, UAP is crafted by non-target models.). All results of mean and standard deviation are computed on average with five turns.

4.1. Main Results

Our proposed TRM-UAP is applied to craft UAPs on five CNN models and create AEs to attack these CNN models separately on the ImageNet validation set. We set $\alpha \in \{1.0, 1.0, 0.5, 1.5, 1.3\}$, $PTR \in \{0.0, 0.2, 0.2, 0.7, 0.4\}$, $NTR \in \{0.8, 0.3, 0.2, 0.8, 0.3\}$, corresponding to AlexNet, VGG16, VGG19, ResNet152, GoogleNet. The attack performance compared with other data-free universal methods is shown in Table 1. Note that the results of GD-UAP were reproduced by our best effort on PyTorch for a fair comparison. As compared to other methods, our TRM-UAP method improves the fooling rate in most CNN models. Although Cosine-UAP achieves a higher FR in GoogleNet, TRM-UAP improves the FR around $1 \sim 5\%$ on other four models and further improves the average FR of all five models more than 2%. Besides, it is shown in Fig. 4 that the visualization of crafted UAPs has clearly particular image patterns with abundant features. We also visualized AEs of TRM-UAP attack in Fig. 6 and found that high probabilities of AEs were given to the labels of incorrect class.

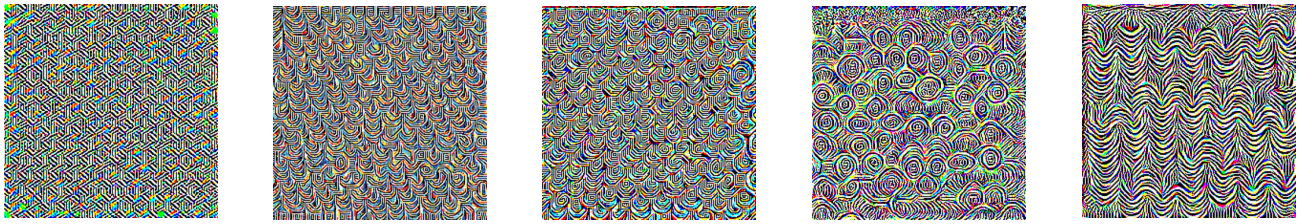
We further used the UAP made by surrogate models to attack other target models to verify the transferability. Table 2 represented the attack performance in different attack settings. The models in rows denote target models, and the models in columns denote surrogate models for crafting the UAP. The results on the diagonal denote the white-box attack, while the rest are transferable attacks in the black-box setting. Most TRM-UAP attacks perform excellently in transferability, which indicates that the universal perturbations crafted by TRM-UAP are generalized well to transfer attacks on target models.

4.2. Evaluating the Transferability

To further compare the effectiveness of universal attacks, we used the logit loss of C&W attack [5] to evaluate the transferability of GD-UAP, Cosine-UAP, and TRM-UAP. The logit loss is defined as $loss = (\max_{j \neq t} F(\boldsymbol{v} + \boldsymbol{x})_j) - F(\boldsymbol{v} + \boldsymbol{x})_t$, where $F(\cdot)_j$ is the j-th output in the logits layer of network and t is the index of ground truth. A larger logit loss value represents better transferability of attacks to CNN models. The logit loss of original images is calculated as a baseline. Note that the results of GD-UAP and Cosine-UAP are reproduced with our best effort on PyTorch. As shown in Fig. 5, TRM-UAP achieves a higher logit loss than GD-UAP and Cosine-UAP in most cases. Additionally, the discrepancy among GD-UAP, Cosine-UAP, and TRM-UAP becomes significant in deep CNN models (*e.g.*, ResNet152 and GoogleNet). Therefore, TRM-UAP achieves relatively higher logit losses in both white-box and black-box attack settings, which indicates a better transferability across various CNN models.

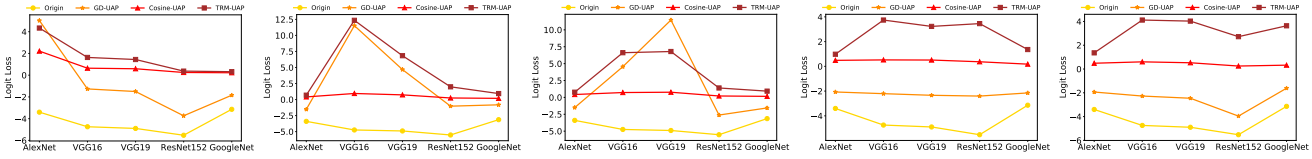
4.3. Parameter Study

First, the influence of ratio maximization on attack performance is studied. Here we mainly focused on validating the effectiveness of truncated negative activation in deep CNN models (*e.g.*, ResNet152 and GoogleNet). The exper-



(a) AlexNet (b) VGG16 (c) VGG19 (d) ResNet152 (e) GoogleNet

Figure 4. The visualization of UAPs crafted by our TRM-UAP method. For the best visualization in color, we enhance the value of the pixel to [0, 255].



(a) AlexNet (b) VGG16 (c) VGG19 (d) ResNet152 (e) GoogleNet

Figure 5. The logit loss of original samples and adversarial samples crafted by GD-UAP, Cosine-UAP, our TRM-UAP methods. In each sub-figure, the subtitle indicates the surrogate model for crafting UAPs, and target models are shown on the horizontal axis. Original samples are classified correctly by the target model and exploited to craft AEs for comparisons. For each attack setting, we compute the average of the logit loss over all the samples.

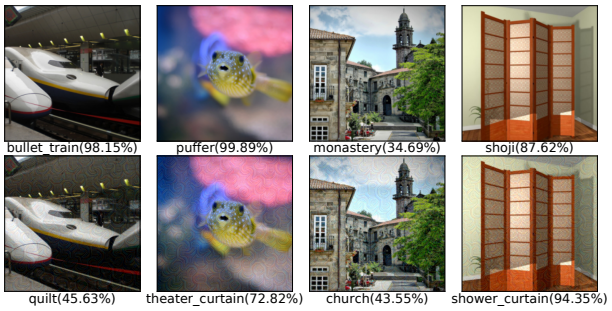
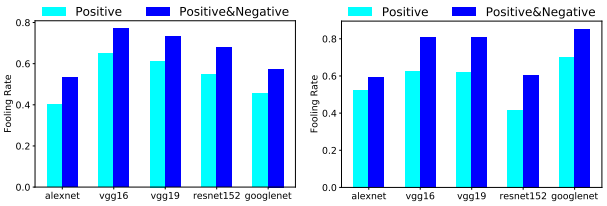


Figure 6. Examples of our TRM-UAP attack tested on VGG16 and the perturbation crafted by VGG19 (top: original examples; bottom: adversarial examples). The annotation represents the classification label and the probability predicted by CNNs.



(a) ResNet152 (b) GoogleNet

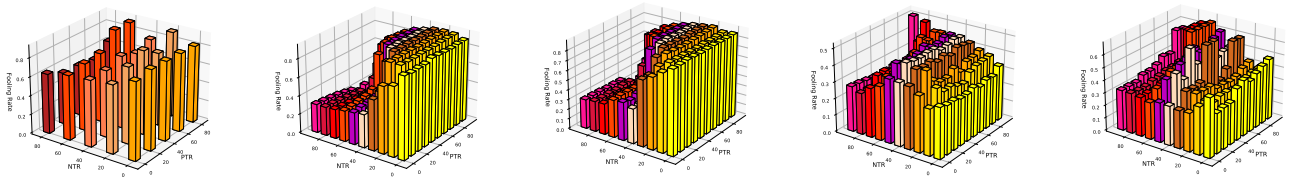
Figure 7. The ablation study of truncated negative activations on CNN models. Positive: the attack only uses the truncated positive activation. Positive&Negative: the attack integrates truncated positive and negative activations.

imental setting is the same as Section 4.1. Based on this, we also crafted UAPs by truncated positive activation only, *i.e.*, PTR remains unchanged and $NTR = 100\%$. As shown in Fig. 7, the ablation experiment indicates that truncated negative activation combined with truncated positive activation improves the FR by a large margin (around $8 \sim 18\%$) as compared to standalone truncated positive activation, thus enhancing both intensity and transferability of UAPs significantly.

To explore the influence of truncation strategy, we computed the fooling rate of attacks in CNN models by varying truncation rates. Specifically, the proportion of retained activation increased by 10% for each step. That means UAP learned the feature information from shallow to deep convolution layers gradually. Due to space limitations, UAPs were only tested in white-box settings. The results of all CNN models were presented in Fig. 8, where the parameter space was explored w.r.t. NTR and PTR. As can be seen, the optimal result was obtained on the attack setting of the truncated activation rather than the whole activation. Especially in deep CNN models, only part of positive and negative activations were retained to craft UAPs. From the analysis of truncation in Fig. 8, we observe that the shallow convolution layer extracts generic features which are better at transferring the attack across different models.

4.4. Exploring the Property of UAP

In the following, the property of universal perturbations is thoroughly investigated. We studied the distinction between original example and adversarial examples made by image-specific attack (MI-FGSM) and data-free universal attacks (GD-UAP and TRM-UAP). As shown in Fig. 9, AEs



(a) AlexNet (b) VGG16 (c) VGG19 (d) ResNet152 (e) GoogleNet

Figure 8. The parameter study of truncation rates on CNN models in white-box attack setting. The x, y, and z axis are NTR, PTR, and FR, respectively. Note that there are only five convolution layers in AlexNet, truncation results (*e.g.*, 10%, 30%, 50%, *etc.*) are omitted since these values are consistent with the ones at their right coordinate.

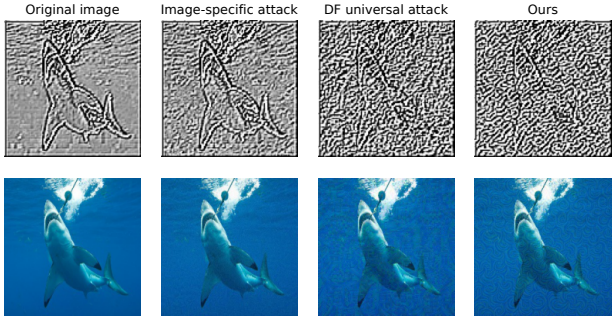
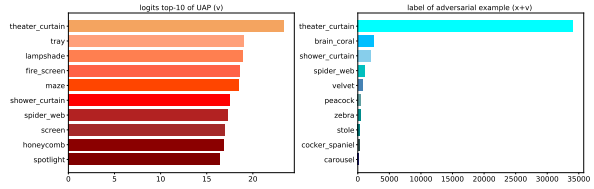
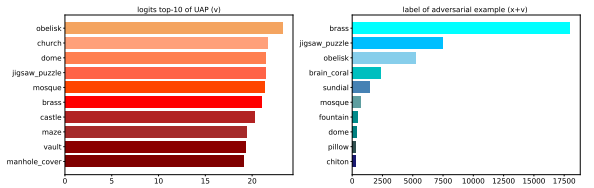


Figure 9. The distinction of feature maps among original image, image-specific attack, and universal attacks (top: feature maps extracted from a shallow convolution layer; bottom: original example and AEs made by VGG19).



(a) UAP from low-level (former 80%) convolution layers



(b) UAP from high-level (latter 20%) convolution layers

Figure 10. The analysis of the label distribution of UAPs and AEs (both made by VGG16). In each subfigure, the left is the top-10 logits value of UAP ($\boldsymbol{v}$), and the right is the distribution of top-10 predicated labels of AEs ($\boldsymbol{v} + \boldsymbol{x}$).

are indistinguishable compared with original example, and the imperceptible noise added to the input effectively perturbed feature maps in a shallow convolution layer. Furthermore, data-free universal attacks disturbed the feature map seriously in contrast to the image-specific attack. Nevertheless, the feature map of TRM-UAP had better regularity and repeatability than GD-UAP. We can conclude that features extracted by shallow convolution layers profoundly impact the final classification of CNN models and are beneficial for universal attack to craft the transferable perturbation.

To further verify the above viewpoints, we trained UAPs in completely opposite truncation strategies (*i.e.*, retaining former 80% layers versus retaining latter 20% layers), and counted the distribution of class labels of AEs. The result in Fig. 10 (a) was consistent with the hypothesis [33] that the top-1 logits prediction of UAP dominates the prediction of AEs. Both labels were *theater curtain*. Contrarily, it was shown in Fig. 10 (b) that the phenomenon of top-1 label domination disappeared, where the UAP label was *obelisk* and the AE label was *brass*. Thus, the shallow convolution layer is more beneficial to transfer universal attack with generic features through a variety of models.

5. Conclusion

In this paper, we have proposed a novel data-free universal attack, called TRM-UAP, to reformulate the UAP generation task as a truncated ratio optimization problem. Compared with previous methods, our TRM-UAP method integrates positive activation maximization with negative activation minimization. Towards further improving the transferability of UAPs, we propose a truncation strategy that computes positive and negative activation losses from low-level convolutions, as well as a curriculum optimization algorithm to fully mine the diversity of artificial images. Experimental results on the ImageNet dataset validate the better transferability of our TRM-UAP attack than other data-free UAP attacks on different CNN models. Additionally, UAPs learned from generic features in shallow convolution layers can dominate the model prediction and transfer the universal attack.

Acknowledgement. This work is partially supported by the Scientific Research Foundation of Chongqing University of Technology under Grant No.2022ZDZ026, Natural Science Foundation of Chongqing, China under Grant No.CSTB2022NSCQ-MSX0493, the Key project of Chongqing Technology Innovation and Application Development under Grant No.cstc2021jscx-dxwtBX0018.

References

[1] Naveed Akhtar and Ajmal Mian. Threat of adversarial attacks on deep learning in computer vision: A survey. *IEEE Access*, 6:14410–14430, 2018.

[2] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48, 2009.

[3] Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 387–402. Springer, 2013.

[4] Battista Biggio and Fabio Roli. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, 84:317–331, 2018.

[5] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy*, pages 39–57, 2017.

[6] Ashutosh Chaubey, Nikhil Agrawal, Kavya Barnwal, Keerat K Guliani, and Pramod Mehta. Universal adversarial perturbations: A survey. *arXiv preprint arXiv:2005.08087*, 2020.

[7] 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.

[8] Aditya Ganeshan, Vivek BS, and R Venkatesh Babu. Fda: Feature disruptive attack. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8069–8079, 2019.

[9] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

[10] 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.

[11] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. *Advances in Neural Information Processing Systems*, 32, 2019.

[12] Valentin Khrulkov and Ivan Oseledets. Art of singular vectors and universal adversarial perturbations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8562–8570, 2018.

[13] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25, 2012.

[14] Hong Liu, Rongrong Ji, Jie Li, Baochang Zhang, Yue Gao, Yongjian Wu, and Feiyue Huang. Universal adversarial perturbation via prior driven uncertainty approximation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2941–2949, 2019.

[15] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

[16] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1765–1773, 2017.

[17] Konda Reddy Mopuri, Aditya Ganeshan, and R Venkatesh Babu. Generalizable data-free objective for crafting universal adversarial perturbations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(10):2452–2465, 2018.

[18] Konda Reddy Mopuri, Utsav Garg, and R Venkatesh Babu. Fast feature fool: A data independent approach to universal adversarial perturbations. In *Proceedings of the British Machine Vision Conference*, 2017.

[19] Konda Reddy Mopuri, Utkarsh Ojha, Utsav Garg, and R Venkatesh Babu. Nag: Network for adversary generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 742–751, 2018.

[20] Konda Reddy Mopuri, Phani Krishna Uppala, and R Venkatesh Babu. Ask, acquire, and attack: Data-free uap generation using class impressions. In *Proceedings of the European Conference on Computer Vision*, pages 20–35, 2018.

[21] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *2016 IEEE European Symposium on Security and Privacy*, pages 372–387, 2016.

[22] Omid Poursaeed, Isay Katsman, Bicheng Gao, and Serge Belongie. Generative adversarial perturbations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4422–4431, 2018.

[23] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[24] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

[25] 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.

[26] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014.

[27] Yusuke Tashiro, Yang Song, and Stefano Ermon. Diversity can be transferred: Output diversification for white-and black-box attacks. *Advances in Neural Information Processing Systems*, 33:4536–4548, 2020.

[28] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *International Conference on Learning Representations*, 2019.

[29] Zhibo Wang, Hengchang Guo, Zhifei Zhang, Wenxin Liu, Zhan Qin, and Kui Ren. Feature importance-aware transferable adversarial attacks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7639–7648, 2021.

[30] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *Advances in Neural Information Processing Systems*, 27, 2014.

[31] Yunrui Yu, Xitong Gao, and Cheng-Zhong Xu. Lafeat: piercing through adversarial defenses with latent features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5735–5745, 2021.

[32] Chaoning Zhang, Philipp Benz, Tooba Imtiaz, and In So Kweon. Understanding adversarial examples from the mutual influence of images and perturbations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14521–14530, 2020.

[33] Chaoning Zhang, Philipp Benz, Adil Karjauv, and In So Kweon. Data-free universal adversarial perturbation and black-box attack. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7868–7877, 2021.

[34] Chaoning Zhang, Philipp Benz, Chenguo Lin, Adil Karjauv, Jing Wu, and In So Kweon. A survey on universal adversarial attack. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pages 4687–4694, 2021.

[35] Jianping Zhang, Weibin Wu, Jen-tse Huang, Yizhan Huang, Wenxuan Wang, Yuxin Su, and Michael R Lyu. Improving adversarial transferability via neuron attribution-based attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14993–15002, 2022.