Towards Transferable Adversarial Attacks with Centralized Perturbation

Shangbo Wu[1], **Yu-an Tan**[1], **Yajie Wang**[1], **Ruinan Ma**[1], **Wencong Ma**[2], **Yuanzhang Li**[2*]

[1]School of Cyberspace Science and Technology, Beijing Institute of Technology
[2]School of Computer Science and Technology, Beijing Institute of Technology
{shangbo.wu, tan2008, wangyajie19, ruinan, wencong.ma, popular}@bit.edu.cn

Abstract

Adversarial transferability enables black-box attacks on unknown victim deep neural networks (DNNs), rendering attacks viable in real-world scenarios. Current transferable attacks create adversarial perturbation over the entire image, resulting in excessive noise that overfit the source model. Concentrating perturbation to dominant image regions that are model-agnostic is crucial to improving adversarial efficacy. However, limiting perturbation to local regions in the spatial domain proves inadequate in augmenting transferability. To this end, we propose a transferable adversarial attack with fine-grained perturbation optimization in the frequency domain, creating centralized perturbation. We devise a systematic pipeline to dynamically constrain perturbation optimization to dominant frequency coefficients. The constraint is optimized in parallel at each iteration, ensuring the directional alignment of perturbation optimization with model prediction. Our approach allows us to centralize perturbation towards sample-specific important frequency features, which are shared by DNNs, effectively mitigating source model overfitting. Experiments demonstrate that by dynamically centralizing perturbation on dominating frequency coefficients, crafted adversarial examples exhibit stronger transferability, and allowing them to bypass various defenses.

Introduction

DNNs demonstrate outstanding performance across various real-world applications (He et al. 2016; Dosovitskiy et al. 2021). However, DNNs remain susceptible to adversarial examples — malicious samples with carefully-crafted imperceptible perturbation that disrupt DNN functionalities (Zhang et al. 2023). The transferability of adversarial examples (Liu et al. 2017; Guo et al. 2019) allows for cross-model black-box attacks on even unknown victim DNNs, i.e., perturbation created on one model can fool another with no modification, posing a practical real-world threat.

Existing transferable adversarial attacks are gradient-based iterative attacks, stemming from FGSM (Goodfellow, Shlens, and Szegedy 2015) and PGD (Madry et al. 2018). By greedily accumulating gradient information obtained from the white-box source model, these attacks are able to generate transferable adversarial perturbation. However, as attacks try to search the entirety of the input space, the perturbation created tend to overfit the source model, producing excessive noise.

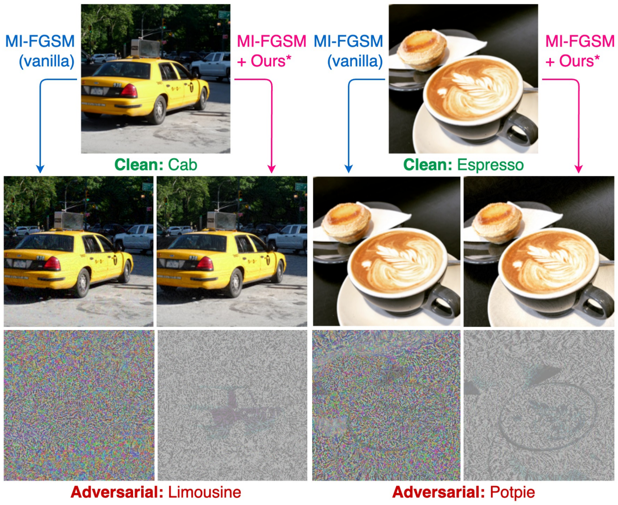

Figure 1: Adversarial examples crafted by vanilla MI-FGSM (blue arrow, left sample) and MI-FGSM boosted by our approach (pink arrow, right sample) on source model ResNet50 with $\varepsilon = 8/255$ and $T = 10$. Perturbation normalized for visualization (bottom row).

Yao et al. (2019) and Xu et al. (2019) explored perturbation constraints within the spatial domain, boasting an improved efficiency for white-box attacks and effectiveness against model interpretability methods. While these attempts fall short for transfer-based black-box attacks, the idea of concentrating perturbation is valuable.

Ehrlich and Davis (2019) first showed intriguing properties of the frequency features of images that can be exploited. Prior work attempts to craft low-frequency perturbation with a fixed frequency constraint, assuming that the fixed low-frequency region reflect higher DNN sensitivity. However, DNNs do not respond uniformly nor unchangeably to these regions. Sensitivity to frequency coefficients varies across images, only with lower frequencies tending to be more impactful. Our goal is to mitigate the effects of perturbation overfitting to the source model, to improve trans-

[*]Corresponding author.

ferability. Motivated by this insight, we propose to strategically optimize adversarial perturbation by identifying sensitive frequency regions per input, per iteration, targeting influential frequency coefficients dynamically.

In this paper, we propose to craft centralized adversarial perturbation, encompassing a shared frequency decomposition procedure, and two major strategies to regularize optimization. Frequency decomposition transforms data into frequency coefficient blocks with DCT (Discrete Cosine Transform). Perturbation centralization is performed via a quantization on these coefficients, omitting excessive perturbation, centralizing the optimization towards dominant regions. Our paramount contribution is a fine-grained quantization, controlled through a subsequent optimization of the quantization matrix, guaranteeing its direct alignment with regional sensitivity. Finally, we design our pipeline as a plug-and-play module, enabling seamless integration of our strategy into existing state-of-the-art gradient-based attacks.

Depicted in Figure 1, the effect of our proposed strategy to centralize perturbation is apparent. On the left side of the samples, perturbations generated by vanilla MI-FGSM exhibit obvious noise. In contrast, our centralized perturbation, while noticeably smaller in magnitude, yield equivalent adversarial potency. Our proposed optimization strategy eradicates model-specific perturbation, effectively avoiding source model overfitting. By purposely centralizing perturbation optimization within high-contribution frequency regions, improved adversarial effectiveness naturally arise. Our approach significantly enhances adversarial transferability by 11.7% on average, and boosts attack efficacy on adversarially defended models by 10.5% on average.

In summary, our contributions are as follows:

- We design a shared frequency decomposition process with DCT. Excessive perturbation is eliminated through quantization on blocks of frequency coefficients. As such, adversarial perturbation is centralized, effectively avoiding source model overfitting.
- We propose a systematic approach for precise control over the quantization process. Quantization matrix is optimized dynamically in parallel with the attack, ensuring a direct alignment with model prediction at each step.
- Through comprehensive experiments, we substantiate the efficacy of our proposed centralized perturbation. Our approach yields significantly enhanced adversarial effectiveness. Under the same perturbation budget, centralized perturbation achieve stronger transferability, and is more successful in bypassing defenses.

Related Work

Adversarial Attacks. Gradient-based attacks are extensively employed in both white-box and black-box transfer scenarios. The seminal work by Goodfellow, Shlens, and Szegedy (2015) introduced the single-step FGSM that exploits model gradients. Kurakin, Goodfellow, and Bengio (2016) improves optimization with the multi-step iterative BIM. Numerous endeavors have been made to enhance the transferability of iterative gradient-based attacks, such as MI-FGSM (Dong et al. 2018), TI-FGSM (Dong et al. 2019), DI-FGSM (Xie et al. 2019), SI-NI-FGSM (Lin et al. 2020), VMI-FGSM (Wang and He 2021), among others. We use gradient-based iterative solutions as our attack basis.

Frequency Optimizations. Ehrlich and Davis (2019) first explored benefits of frequency-domain deep learning. Guo, Frank, and Weinberger (2019), Sharma, Ding, and Brubaker (2019), and Deng and Karam (2020) exploit low-frequency image features to optimize perturbation within a predefined, constant low-frequency region. However, the assumption that DNN sensitivity maps to fixed frequency regions is disclaimed in Maiya et al. (2021), where different frequency regions yield varying sensitivity, with a higher sensitivity tendency towards lower coefficients. In contrast, our contribution lies in our effort to dynamically adjust frequency-domain perturbation optimization over back-propagation, augmenting attacks with centralized perturbation.

Methodology

Overview

Given a DNN classifier $\mathcal{F}(\cdot) : \boldsymbol{x} \in \mathbb{R}^m \mapsto y$, where $\boldsymbol{x}$ and y is the clean sample and ground truth label respectively. The adversary aims to create adversarial example $\boldsymbol{x}^{adv} = \boldsymbol{x} + \boldsymbol{\delta}$ with a minimal perturbation $\boldsymbol{\delta}$ such that $\mathcal{F}(\boldsymbol{x}^{adv}) \neq y$, where $\boldsymbol{x}^{adv}$ is restricted by ℓ_p-ball (ℓ_∞ in this work). Hence, the attack is an optimization which can be formulated as

$$\arg\max_{\boldsymbol{x}^{adv}} \mathcal{J}(\boldsymbol{x}^{adv}, y),\ s.t.\ \|\boldsymbol{x}^{adv} - \boldsymbol{x}\|_\infty \leqslant \varepsilon, \tag{1}$$

where $\mathcal{J}(\cdot, \cdot)$ is the cross-entropy loss. Various gradient-based approaches have been proposed to solve this optimization as stated before, which we adopt as our foundation.

To optimize and craft centralized perturbation in the frequency domain, we devise a three-fold approach as follows:

1. **Frequency coefficient decomposition:** Inspired by the JPEG codec, we design a shared differential data processing pipeline to decompose data into the frequency domain with DCT.
2. **Centralized perturbation optimization:** Within the frequency domain, excessive perturbation is omitted via quantizing each Y/Cb/Cr channel, centralizing perturbation optimization.
3. **Differential quantization optimization:** Quantization matrices are optimized through back-propagation, ensuring quantization aligns with dominant frequency regions.

We detail these approaches in the following sections.

Frequency Decomposition

We begin by addressing the shared frequency decomposition procedure in our work. We focus on 8-bit RGB images of shape $(3, 224, 224)$. Illustrated in Figure 2, let $\boldsymbol{X}$ be the data that we would like to *frequency decompose*:

1. We first convert $\boldsymbol{X}$ from RGB into the YCbCr color space, consisting of 3 color channels: the luma channel Y, and the chroma channels Cb and Cr.
2. After which, a channel-wise global DCT is applied to transform the data losslessly into the frequency domain.

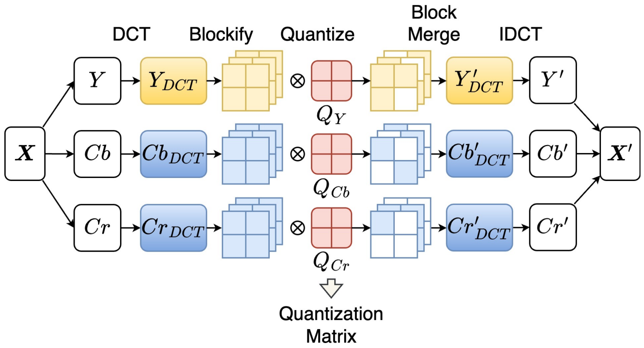

Figure 2: The frequency decomposition procedure.

3. Next, a "blockify" process is used to reshape the data into blocks of (8×8). The operation is applied to the last two dimensions of the data (width W and height H). As such, an image of shape (B, C, W, H) would be "blockified" into shape $(B, C, W \cdot H/64, 8, 8)$ (B: batch, C: channel).
4. Quantization is applied channel-wise via quantization matrices Qs, omitting excessive frequency coefficients.
5. Finally, inverse operations, namely "block merge" and "IDCT" (inverse-DCT) is performed, reconstructing coefficients back into RGB image $\boldsymbol{X'}$.

This procedure would decompose $\boldsymbol{X}$ into blocks of frequency coefficients in each Y/Cb/Cr channel. The sequential operation of "blockify" guarantees the lossless inversion of "block merge". Doing so, we ensure that every stage in our procedure is losslessly invertible, enabling reconstruction without information loss (with the exception of deliberate data quantization). The linearity of DCT and IDCT allow for the possibility of optimization within a reduced model space dimensionality, enabling the centralized perturbation and quantization to be optimized effectively.

This shared frequency decomposition procedure is shown in the green-bounded area in Figure 3. For the next two sections, we address our optimization strategy after acquiring gradient-based attack crafted perturbation $\boldsymbol{\delta}_t$ at iteration t.

Centralized Perturbation Optimization

Depicted in *Step 1* of Figure 3, first, the frequency decomposition is applied to $\boldsymbol{\delta}_t$ at iteration t, where it is separated into luma and chroma channels, then decomposed into frequency coefficients via DCT and "blockify". Resulting components are denoted as B_Y, B_{Cb}, B_{Cr}, where

$$\begin{aligned} B_Y &= \text{blockify}(DCT(Y)), \\ B_{Cb} &= \text{blockify}(DCT(Cb)), \\ B_{Cr} &= \text{blockify}(DCT(Cr)). \end{aligned} \tag{2}$$

Quantization matrices Q_Y, Q_{Cb}, Q_{Cr} are applied over each Y/Cb/Cr channel as

$$\begin{aligned} B'_Y &= B_Y \odot Q_Y, \\ B'_{Cb} &= B_{Cb} \odot Q_{Cb}, \\ B'_{Cr} &= B_{Cr} \odot Q_{Cr}. \end{aligned} \tag{3}$$

Lastly, inverse operations are taken as

$$\begin{aligned} Y' &= IDCT(\text{block-merge}(B'_Y)), \\ Cb' &= IDCT(\text{block-merge}(B'_{Cb})), \\ Cr' &= IDCT(\text{block-merge}(B'_{Cr})), \end{aligned} \tag{4}$$

and with an RGB conversion, reconstruction is complete.

Differing from the quantization matrix employed in the JPEG codec, ours is defined as $Q = (q_{ij}) \in \{0,1\}^{m \times m}$ where $m = 8$, and initialized as $Q_0 = \mathbb{1}$. Through the blockwise multiplication of $B \odot Q$, less impactful frequency coefficients are systematically discarded at every iteration. This process confines the perturbation optimization exclusively within frequency regions of greater impact over DNN predictions, as modeled by the pattern in Q.

At iteration t, perturbation $\boldsymbol{\delta}_t$ will be centralized via optimization, yielding $\boldsymbol{\delta}'_t$. We denote the entire process as

$$\boldsymbol{\delta}'_t = \mathcal{K}(\boldsymbol{\delta}_t; Q_t), \tag{5}$$

where $\mathcal{K}(\cdot; Q)$ is the frequency decomposition and quantization function with quantization matrix Q. Q_t is fixed at this stage per each iteration t.

Differential Quantization Matrix Optimization

Intuitively, quantization matrix Q holds pivotal significance in the process of centralizing frequency-sensitive perturbation. In this section, we delve into the modeling and optimization of Q to ensure its direct alignment with frequency-wise coefficient influence on DNN predictions.

Illustrated in *Step 2* of Figure 3, during this stage, the optimized $\boldsymbol{\delta}'_t$ is added onto $\boldsymbol{x}$ to get intermediate adversarial example $\boldsymbol{x}^{adv}_t$ at iteration t. $\boldsymbol{x}^{adv}_t$ goes through the identical shared frequency decomposition process, and is quantized in the same manner as *Step 1*, giving us $\boldsymbol{x}^{adv\prime}_t$. The quantized $\boldsymbol{x}^{adv\prime}_t$ is then fed back into source model $\mathcal{F}(\cdot)$, such that Q is updated subsequently through back-propagation.

We formulate the update of Q as an optimization problem by leveraging the gradient changes of the source model before and after the quantization of $\boldsymbol{x}^{adv}_t$. Our optimization goal is: Q should quantize $\boldsymbol{x}^{adv}_t$ in a way that, at each iteration, the source model should be less convinced that $\boldsymbol{x}^{adv}_t$ is y. As such, the loss is formulated as

$$\arg\max_{Q_t} \mathcal{J}(\mathcal{K}(\boldsymbol{x}^{adv}_t; Q_t), y), \tag{6}$$

where Q_t is the variable for optimization. Q consists of Q_Y, Q_{Cb}, Q_{Cr} for each Y/Cb/Cr channel respectively, and is collectively optimized by the Adam optimizer to solve this optimization. With this approach, we assure that (1) Q accurately reflects the impact of frequency coefficients on model predictions *per each iteration*, allowing for fine-grained control over the centralized perturbation quantization process, and (2) gradients from the source model accumulate across successive iterations, further boosting transferability.

Through the process of back-propagation, Q undergoes iterative updates via optimization. We denote the optimized result as matrix $P = (p_{ij}) \in \mathbb{R}^{m \times m}$. Following the update, a rounding function $\mathcal{R}(\cdot)$ is implemented before Q is applied

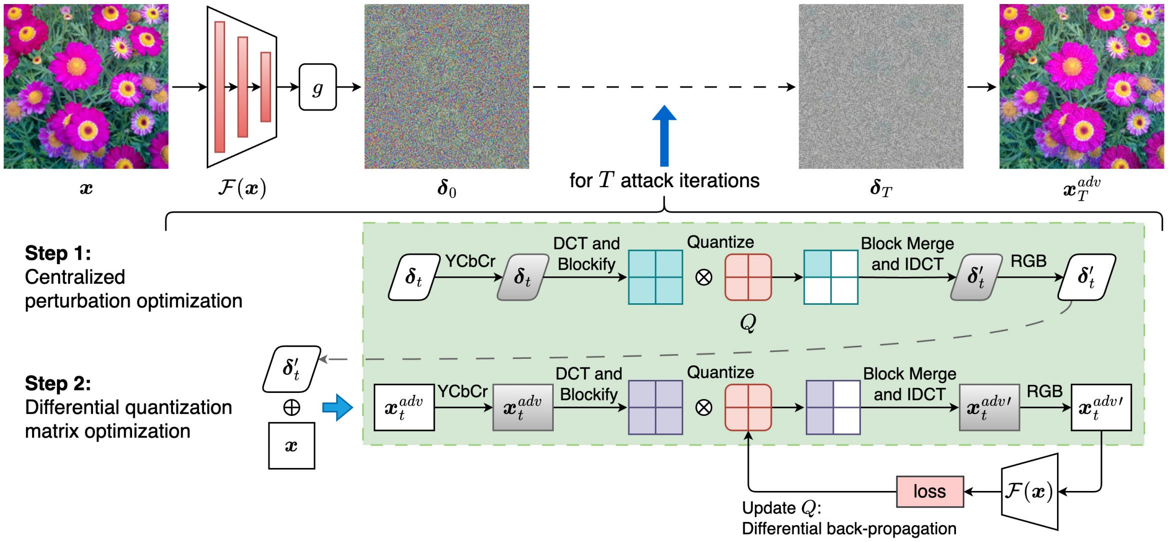

Figure 3: Pipeline of our strategy. Perturbation crafted with gradient-based attacks is centralized in the first step. Quantization matrix controlling the centralization is then optimized via back-propagation in the second step.

```
Algorithm 1: Centralized Adversarial Perturbation
Input: Original image $\boldsymbol{x}$, ground truth label $y$, source model $\mathcal{F}$ with loss $\mathcal{J}$.
Parameters: Iteration $T$, size of perturbation $\varepsilon$, learning rate $\beta$, quantization ratios $r_Y, r_{Cb}, r_{Cr}$ and corresponding quantization matrices $Q_Y, Q_{Cb}, Q_{Cr}$ (denoted as $Q$s for brevity) each Y/Cb/Cr channel.
Output: $\boldsymbol{x}_T^{adv}$.
Let step size $\alpha \leftarrow \varepsilon/T$, $Q_0 = \mathbb{1}$.
for $t = 0$ to $T - 1$ do
    Acquire gradient from $\mathcal{F}$ as $\nabla_{\boldsymbol{x}}\mathcal{J}(\boldsymbol{x}_t^{adv}, y)$.
    Perturbation $\boldsymbol{\delta}_t = \alpha \cdot \mathrm{sign}(\nabla_{\boldsymbol{x}}\mathcal{J}(\boldsymbol{x}_t^{adv}, y))$.
    Optimize $\boldsymbol{\delta}_t$ by Equation 5 and 7 to acquire $\boldsymbol{\delta}_t'$, and clip with respect to $\varepsilon$.
    Acquire intermediate $\boldsymbol{x}_t^{adv} = \boldsymbol{x} + \boldsymbol{\delta}_t'$.
    Update $Q_t$s by passing the quantized $\boldsymbol{x}_t^{adv}$ (through the same Equation 5 and 7) to $\mathcal{F}$ and solving the optimization via Equation 6.
end for
return $\boldsymbol{x}_T^{adv} = \boldsymbol{x} + \boldsymbol{\delta}_T'$.
```

to the quantization process. Specifically, for a quantization ratio of r where $0 \leqslant r \leqslant 1$ for each Y/Cb/Cr channel,

$$Q = \mathcal{R}(P; r) = \begin{cases} 1, & \text{where } p_{ij} \geqslant \rho, \\ 0, & \text{otherwise,} \end{cases} \tag{7}$$

where ρ is the r-th percentile of $\{p_{ij}\}$, i.e., the threshold for each Y/Cb/Cr channel. While Equation 7 is represented as a staircase (i.e., non-differentiable), $\mathcal{R}(\cdot)$ is implemented in a differentiable manner, which we detail in the Appendix.

The Algorithm

Finally, Algorithm 1 showcase a basic attack integrated with our strategy, crafting centralized adversarial perturbation (based on BIM). After the conventional step of creating intermediate perturbation from the source model's gradients, our approach of centralizing the perturbation in a frequency-sensitive manner, and updating the quantization matrix with respect to model predictions, is a streamline integration.

Our strategy guarantees that the centralized perturbation through frequency quantization is concentrated into frequency regions that hold greater significant to the model's judgements. Updating the quantization matrix per iteration allows for a more fine-grained control, making sure that each optimization aligns with the iteration-local gradients of the model. Lastly, as a bonus, at each iteration, gradients gets carried over and accumulated to the next through the directional optimization of the quantization matrix, improving adversarial transferability.

Experiments

Setup

Dataset. The dataset from the NeurIPS 2017 Adversarial Learning Challenge (Kurakin et al. 2018) is used, consisting of 1000 images from ImageNet with shape $(3, 224, 224)$.

Networks. We choose 3 source models as local surrogate models: ResNet-50 (Res-50) (He et al. 2016), VGG-19 (Simonyan and Zisserman 2015), Inception-v3 (Inc-v3) (Szegedy et al. 2016). All pretrained models are sourced from Wightman (2019).

Attacks. Our approach is integrated and evaluated against 5 state-of-the-art gradient-based attacks, namely MI-FGSM, DI-FGSM, TI-FGSM, VMI-FGSM, and SI-NI-FGSM, for a better presentation of transferability. The unmodified forms of these attacks serve as the baselines for our experiments.

Implementation Details. We intentionally drop excessive perturbation, centralizing optimization towards dominant

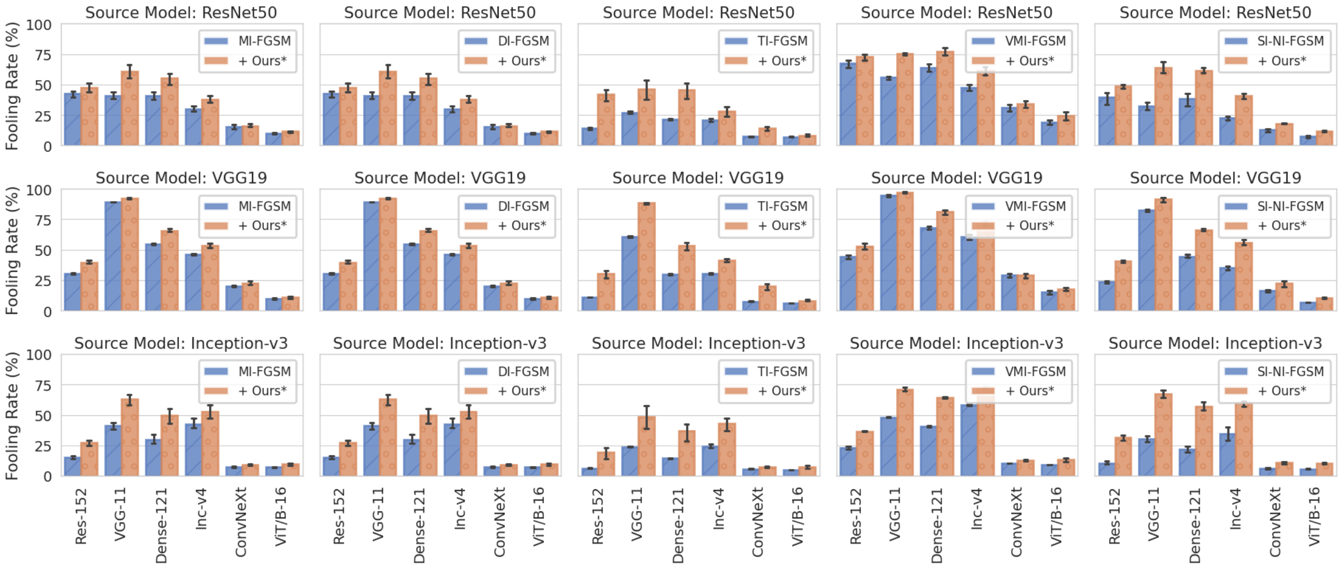

Figure 4: Evaluation of adversarial transferability. Transfer fooling rates are aggregated across attacks over T.

frequency regions. As such, to enable a fair comparison between our attack and baseline methods: for all three Y/Cb/Cr channels, each set at a quantization ratio of r_Y, r_{Cb}, r_{Cr}, the cumulative quantization rate is $(r_Y + r_{Cr} + r_{Cr})/3$; With a baseline attack under an ℓ_∞ constraint of ε_0, we equate our attack to the same perturbation budget by imposing a constraint of $\varepsilon = \varepsilon_0/((r_Y + r_{Cr} + r_{Cr})/3)$.

Hyper-parameters. Attacks are ℓ_∞-bounded. $\varepsilon = 8/255$. They run for iteration $T = 10, 20, 50$. Learning rate of the Adam optimizer $\beta = 0.1$. Quantization ratios are set to $r_Y = 0.9, r_{Cb} = 0.05, r_{Cr} = 0.05$, with a total quantization rate of $1/3$. All other parameters are kept exact to their original implementations.

Transferability

We initiate our evaluation with our primary focus: adversarial transferability. 6 normally trained models: ResNet-152 (Res-152), VGG-11, DenseNet-121 (Dense-121) (Huang et al. 2017), Inception-v4 (Inc-v4) (Szegedy et al. 2017), ConvNeXt (Liu et al. 2022), and ViT-B/16 (Dosovitskiy et al. 2021) are used as black-box target models.

Figure 4 presents the transfer fooling rates of the crafted adversarial examples used to attack the black-box models. Fooling rates are averaged over parameter: attack iteration T. We report a consistent improvement in transfer fooling rates when integrating our proposed frequency perturbation centralization strategy with vanilla gradient-based attacks. Displayed in orange, our approach achieves an average boost in adversarial transferability of 11.7% compared to the baseline attacks represented in blue.

We argue that the observed increase in transferability is the direct manifestation of the benefits that centralized perturbation brings in our method. Dominating frequency features of an image is learnt and shared by neural networks alike, while excessive perturbation that are crafted to fit model-specific features is what hinders the transferability of adversarial examples. By centralizing adversarial perturbation towards the shared dominant frequency regions, we effectively disrupt the model's judgment in a more generalized manner, boosting adversarial transferability.

Defense Evasion

Currently two types of defenses provide an extra layer of robustness guarantee to neural networks: (1) filter-based defenses, where perturbation is filtered through image transformations, and (2) adversarial training, where models are strengthened by training on adversarial examples.

Filter-based Defense. Guo et al. (2018) first explored using JPEG compression and bit-depth reduction as defenses against adversarial attacks. Since our approach also utilizes similar techniques (such as drawing inspiration from the JPEG codec for perturbation quantization), we are interested in investigating how our method withstands these defenses.

Shown in Table 1, we report the transfer fooling rates of attacks: MI-FGSM, VMI-FGSM and variants integrated with our approach. Highlighted in bold, we observe that attacks integrated with our strategy all outperform their baselines, with an increase of at least 3.5%. We posit that these defenses fail as they intend to preserve dominant image features to maintain visual quality, assuming that perturbation lies within unimportant regions to ensure imperceptibility. As such, they are unable to remove our perturbation designed to centralize around dominating regions, enabling our strategy to bypass these defenses.

Adversarial Training. 5 adversarially trained black-box models: Adv-Inc-v3, Inc-Res-v2$_{ens}$ (Tramèr et al. 2018), Adv-ResNet-50 (FastAT), Adv-ResNet-50 (FreeAT) (Shafahi et al. 2019), and EfficientNet-B0 (AdvProp) (Xie et al. 2020) are used for evaluating our approach against adversarial training as a defense mechanism.

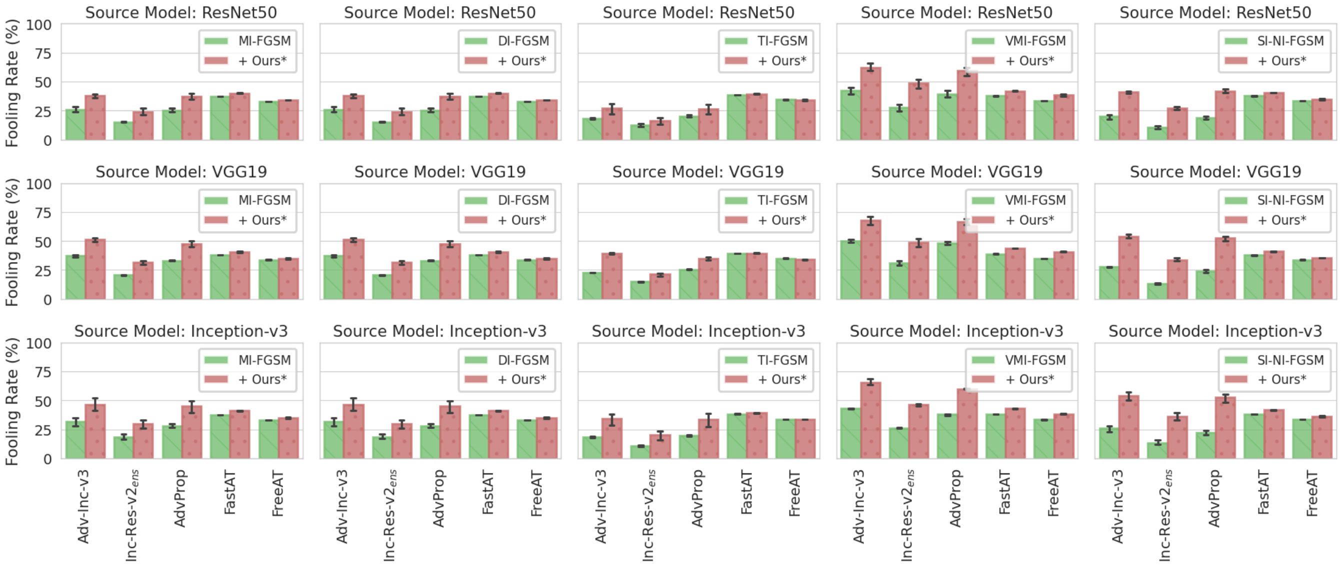

Figure 5: Evaluation of adversarial effectiveness against adversarial training-based defenses. Fooling rates aggregated over T.

Attack	Res-152	VGG-11	Dense-121	Inc-v4	Conv-NeXt	ViT-B/16
JPEG compression — quality level 75						
MI-FGSM	23.1% (±2.0)	59.2% (±13.8)	41.3% (±5.7)	35.0% (±6.7)	15.3% (±1.3)	12.3% (±0.5)
+Ours*	**34.8%** (±1.3)	**74.8%** (±11.0)	**55.8%** (±5.3)	**47.2%** (±6.2)	**22.3%** (±1.9)	**16.3%** (±1.0)
VMI-FGSM	33.1% (±6.0)	66.3% (±15.3)	51.7% (±6.8)	45.7% (±7.3)	22.0% (±2.8)	17.0% (±1.5)
+Ours*	**48.1%** (±5.7)	**81.0%** (±11.8)	**70.4%** (±5.0)	**62.8%** (±5.9)	**31.8%** (±3.5)	**23.8%** (±3.7)
Bit-depth reduction — to 3 bits						
MI-FGSM	17.9% (±2.8)	57.7% (±12.4)	37.8% (±6.2)	33.3% (±5.5)	14.3% (±0.9)	10.4% (±0.3)
+Ours*	**30.3%** (±0.8)	**71.6%** (±11.1)	**53.5%** (±6.0)	**43.3%** (±7.3)	**18.4%** (±1.4)	**13.9%** (±1.4)
VMI-FGSM	28.8% (±8.1)	63.9% (±14.4)	47.3% (±6.7)	42.3% (±7.4)	20.5% (±2.8)	13.5% (±1.5)
+Ours*	**44.7%** (±8.7)	**78.0%** (±12.5)	**69.0%** (±4.9)	**57.8%** (±6.8)	**27.0%** (±3.0)	**20.3%** (±2.5)

Table 1: Evaluation of adversarial effectiveness against filter-based adversarial defenses. Fooling rates aggregated with means and stds over attack iteration T.

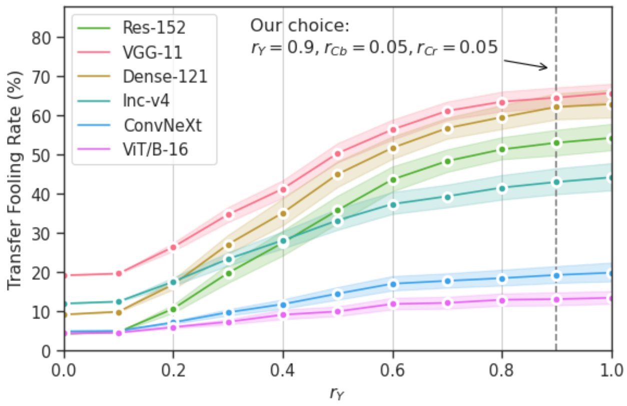

Figure 6: The effect of the quantization ratio of luma channel r_Y vs. chroma channels r_{Cb}, r_{Cr} over transferability.

We report our results in Figure 5, where we observe a steady boost of fooling rates by a maximum of 20% in some scenarios by our strategy. In all, results indicate that our proposed approach can successfully evade adversarial defenses that transfer-based attacks frequently fail against.

The Impact of Quantization Ratio Allocation

The quantization process is crucial for centralized perturbation optimization, controlled by quantization ratios rs. We explain our rationale for the choice of ratios through evaluating the impact of luma channel r_Y vs. chroma channels r_{Cb}, r_{Cr} on adversarial effectiveness.

We maintain a cumulative quantization rate of $1/3$, iterating the luma channel r_Y from 0 to 1 with step size 0.1, and allocating the rest equally to the chroma channels r_{Cb} and r_{Cr}. Figure 6 shows the transfer fooling rates of MI-FGSM integrated with our approach. As r_Y increases, the fooling rates show a consistent increase as well. We reason that this occurs because the luma channel contains more structural information, which DNNs commonly learn as more useful features compared to the chroma channels. However, changes in the luma channel are also more noticeable by the human visual system. As such, we choose $r_Y = 0.9$ for a balance of adversarial effectiveness and perturbation imperceptibility. Further investigation is addressed in the Appendix.

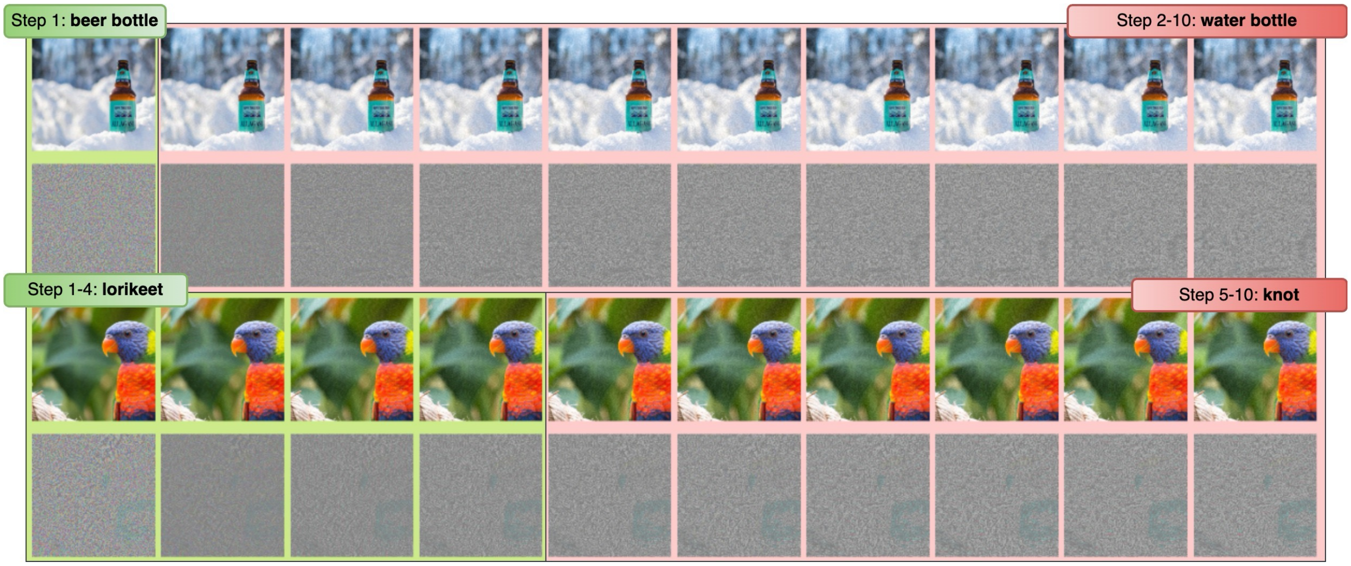

Figure 7: A visualization of perturbation over $T = 10$ optimization iterations (from left to right) of MI-FGSM attacking ResNet50. Correct predictions are framed in green, where incorrect ones are framed in red.

Visualization of Perturbation Optimization

In Figure 7, we visualize the perturbation optimized by MI-FGSM + our approach over 10 iterations, giving further insight to the centralized optimization process of our strategy.

On iteration 1 (step $t = 1$), quantization is initialized as $Q_0 = \mathbb{1}$. On $t = 2$, we observe an immediate drop in perturbation, indicating quantization is applied. Going further ($t = 3$ to 10), we find perturbation is gradually added at each iteration, centralized on more significant frequency regions. By ensuring a consistent optimization alignment with model judgement, our centralized perturbation succeeds in exploiting dominating frequency features of an image.

Ablation Study

We conduct an ablation study on our proposed centralized perturbation optimization process — the core of our strategy. We consider 4 other strategies: (1) *RandA*: randomly choose Qs fixed at the start, (2) *RandB*: randomly choose Qs at each iteration, (3) *Low*: preserving only low-frequency coefficients (which is what previous literature proposes to do), and (4) *High*: preserving only high-frequency coefficients. Details addressed in the Appendix.

In Table 2, we report the fooling rates of vanilla MI-FGSM and SI-NI-FGSM, and ones integrated with *RandA*, *RandB*, *Low*, and our strategy. (*High* is ignored as fooling rates mostly dropped to zero when it is integrated.) Highlighted in bold, we prove our strategy's ability as our approach consistently boosts transferability. In contrast, *RandA* and *RandB* both fail to achieve comparable adversarial effectiveness even with the baselines. While *Low* comes close to matching and occasionally surpasses us, our approach still maintains the upper hand across the majority of cases. We argue that *Low* and our approach both acknowledges that the low-frequency regions hold more dominating features. However, instead of brute-forcely constraining all perturbation towards consecutive low-frequency coefficients, our strategy's precise control over the centralization process yields an improved adversarial effectiveness.

Variant	**White-box**	**Res-152**	**VGG-11**	**Dense-121**	**Inc-v4**	**Conv-NeXt**	**ViT-B/16**
				MI-FGSM			
Vanilla	100%	44.6%	43.0%	45.3%	32.0%	16.6%	10.3%
RandA	86.2%	37.0%	52.1%	45.2%	31.4%	13.0%	9.4%
RandB	92.3%	37.6%	52.2%	45.0%	31.8%	13.7%	9.6%
Low	100%	50.3%	63.1%	56.2%	39.5%	18.9%	11.9%
Ours*	100%	**50.7%**	**65.7%**	**58.7%**	**41.0%**	17.8%	**12.0%**
				SI-NI-FGSM			
Vanilla	99.9%	43.2%	35.4%	42.9%	24.1%	13.4%	8.0%
RandA	88.4%	36.7%	61.5%	49.3%	34.8%	14.7%	10.1%
RandB	76.6%	28.2%	51.4%	42.0%	29.1%	11.6%	8.7%
Low	97.4%	45.5%	66.5%	57.9%	41.2%	18.6%	11.9%
Ours*	**98.7%**	**49.9%**	**68.9%**	**64.1%**	**42.6%**	17.6%	11.2%

Table 2: Ablation study of the centralized perturbation optimization strategy. Source model chosen as ResNet50.

Conclusion

We propose a novel approach to centralize adversarial perturbation to dominant frequency regions. Our technique involves dynamic perturbation optimization through frequency domain quantization, resulting in centralized perturbation that offers improved generalization. To ensure precise quantization alignment with the model's judgement, we incorporate a parallel optimization via back-propagation. Experiments prove that our strategy achieves remarkable transferability and succeeds in evading various defenses. Implications of our work prompt further exploration of defending centralized perturbation under a frequency context.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grants No. 62072037 and U1936218).

References

Deng, Y.; and Karam, L. J. 2020. Frequency-Tuned Universal Adversarial Attacks. *CoRR*, abs/2003.05549.

Dong, Y.; Liao, F.; Pang, T.; Su, H.; Zhu, J.; Hu, X.; and Li, J. 2018. Boosting Adversarial Attacks With Momentum. In *CVPR*, 9185–9193. Computer Vision Foundation / IEEE Computer Society.

Dong, Y.; Pang, T.; Su, H.; and Zhu, J. 2019. Evading Defenses to Transferable Adversarial Examples by Translation-Invariant Attacks. In *CVPR*, 4312–4321. Computer Vision Foundation / IEEE.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*. OpenReview.net.

Ehrlich, M.; and Davis, L. 2019. Deep Residual Learning in the JPEG Transform Domain. In *ICCV*, 3483–3492. IEEE.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2015. Explaining and Harnessing Adversarial Examples. In *ICLR (Poster)*.

Guo, C.; Frank, J. S.; and Weinberger, K. Q. 2019. Low Frequency Adversarial Perturbation. In *UAI*, volume 115 of *Proceedings of Machine Learning Research*, 1127–1137. AUAI Press.

Guo, C.; Rana, M.; Cissé, M.; and van der Maaten, L. 2018. Countering Adversarial Images using Input Transformations. In *ICLR (Poster)*. OpenReview.net.

Guo, F.; Zhao, Q.; Li, X.; Kuang, X.; Zhang, J.; Han, Y.; and Tan, Y. 2019. Detecting adversarial examples via prediction difference for deep neural networks. *Inf. Sci.*, 501: 182–192.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *CVPR*, 770–778. IEEE Computer Society.

Huang, G.; Liu, Z.; van der Maaten, L.; and Weinberger, K. Q. 2017. Densely Connected Convolutional Networks. In *CVPR*, 2261–2269. IEEE Computer Society.

Kurakin, A.; Goodfellow, I. J.; and Bengio, S. 2016. Adversarial examples in the physical world. *CoRR*, abs/1607.02533.

Kurakin, A.; Goodfellow, I. J.; Bengio, S.; Dong, Y.; Liao, F.; Liang, M.; Pang, T.; Zhu, J.; Hu, X.; Xie, C.; Wang, J.; Zhang, Z.; Ren, Z.; Yuille, A. L.; Huang, S.; Zhao, Y.; Zhao, Y.; Han, Z.; Long, J.; Berdibekov, Y.; Akiba, T.; Tokui, S.; and Abe, M. 2018. Adversarial Attacks and Defences Competition. *CoRR*, abs/1804.00097.

Lin, J.; Song, C.; He, K.; Wang, L.; and Hopcroft, J. E. 2020. Nesterov Accelerated Gradient and Scale Invariance for Adversarial Attacks. In *ICLR*. OpenReview.net.

Liu, Y.; Chen, X.; Liu, C.; and Song, D. 2017. Delving into Transferable Adversarial Examples and Black-box Attacks. In *ICLR (Poster)*. OpenReview.net.

Liu, Z.; Mao, H.; Wu, C.; Feichtenhofer, C.; Darrell, T.; and Xie, S. 2022. A ConvNet for the 2020s. In *CVPR*, 11966–11976. IEEE.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. In *ICLR (Poster)*. OpenReview.net.

Maiya, S. R.; Ehrlich, M.; Agarwal, V.; Lim, S.; Goldstein, T.; and Shrivastava, A. 2021. A Frequency Perspective of Adversarial Robustness. *CoRR*, abs/2111.00861.

Shafahi, A.; Najibi, M.; Ghiasi, A.; Xu, Z.; Dickerson, J. P.; Studer, C.; Davis, L. S.; Taylor, G.; and Goldstein, T. 2019. Adversarial training for free! In *NeurIPS*, 3353–3364.

Sharma, Y.; Ding, G. W.; and Brubaker, M. A. 2019. On the Effectiveness of Low Frequency Perturbations. In *IJCAI*, 3389–3396. ijcai.org.

Simonyan, K.; and Zisserman, A. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *ICLR*.

Szegedy, C.; Ioffe, S.; Vanhoucke, V.; and Alemi, A. A. 2017. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. In *AAAI*, 4278–4284. AAAI Press.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the Inception Architecture for Computer Vision. In *CVPR*, 2818–2826. IEEE Computer Society.

Tramèr, F.; Kurakin, A.; Papernot, N.; Goodfellow, I. J.; Boneh, D.; and McDaniel, P. D. 2018. Ensemble Adversarial Training: Attacks and Defenses. In *ICLR (Poster)*. OpenReview.net.

Wang, X.; and He, K. 2021. Enhancing the Transferability of Adversarial Attacks Through Variance Tuning. In *CVPR*, 1924–1933. Computer Vision Foundation / IEEE.

Wightman, R. 2019. PyTorch Image Models. https://github.com/huggingface/pytorch-image-models. Accessed: 2023-12-13.

Xie, C.; Tan, M.; Gong, B.; Wang, J.; Yuille, A. L.; and Le, Q. V. 2020. Adversarial Examples Improve Image Recognition. In *CVPR*, 816–825. Computer Vision Foundation / IEEE.

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 *CVPR*, 2730–2739. Computer Vision Foundation / IEEE.

Xu, K.; Liu, S.; Zhao, P.; Chen, P.; Zhang, H.; Fan, Q.; Erdogmus, D.; Wang, Y.; and Lin, X. 2019. Structured Adversarial Attack: Towards General Implementation and Better Interpretability. In *ICLR (Poster)*. OpenReview.net.

Yao, Z.; Gholami, A.; Xu, P.; Keutzer, K.; and Mahoney, M. W. 2019. Trust Region Based Adversarial Attack on Neural Networks. In *CVPR*, 11350–11359. Computer Vision Foundation / IEEE.

Zhang, Y.; Tan, Y.; Sun, H.; Zhao, Y.; Zhang, Q.; and Li, Y. 2023. Improving the invisibility of adversarial examples with perceptually adaptive perturbation. *Inf. Sci.*, 635: 126–137.