arXiv:2201.01102v1 [cs.CV] 4 Jan 2022

Towards Transferable Unrestricted Adversarial Examples with Minimum Changes

Fangcheng Liu[1], Chao Zhang[1] *, Hongyang Zhang[2] *
[1]Key Laboratory of Machine Perception (MOE), School of EECS, Peking University
[2]David R. Cheriton School of Computer Science, University of Waterloo & Vector Institute
{lfc@stu., c.zhang@}pku.edu.cn hongyang.zhang@uwaterloo.ca

Abstract

Transfer-based adversarial example is one of the most important classes of black-box attacks. However, there is a trade-off between transferability and imperceptibility of the adversarial perturbation. Prior work in this direction often requires a fixed but large ℓ_p-norm perturbation budget to reach a good transfer success rate, leading to perceptible adversarial perturbations. On the other hand, most of the current unrestricted adversarial attacks that aim to generate semantic-preserving perturbations suffer from weaker transferability to the target model. In this work, we propose a *geometry-aware framework* to generate transferable adversarial examples with minimum changes. Analogous to model selection in statistical machine learning, we leverage a validation model to select the optimal perturbation budget for each image under both the ℓ_∞-norm and unrestricted threat models. Extensive experiments verify the effectiveness of our framework on balancing imperceptibility and transferability of the crafted adversarial examples. The methodology is the foundation of our entry to the *CVPR'21 Security AI Challenger: Unrestricted Adversarial Attacks on ImageNet*, in which we ranked 1st place out of 1,559 teams and surpassed the runner-up submissions by 4.59% and 23.91% in terms of final score and average image quality level, respectively. Code is available at https://github.com/Equationliu/GA-Attack.

1 Introduction

Though deep neural networks have exhibited impressive performance in various fields (He et al., 2016; Dosovitskiy et al., 2021), they are vulnerable to adversarial examples (Szegedy et al., 2014; Goodfellow et al., 2015; Shao et al., 2021; Bhojanapalli et al., 2021; Bai et al., 2021), where test inputs that have been modified slightly strategically cause misclassification. Adversarial examples have posed serious threats to various security-critical applications, such as autonomous driving (Bojarski et al., 2016) and face recognition (Parkhi et al., 2015). Most positive results on adversarial attacks have focused on white-box settings (Athalye et al., 2018; Tramer et al., 2020), where the attacker can get full access to the defense models. However, the problem becomes more challenging when it comes to the black-box setting, where the attacker has no information about the model architecture, hyper-parameters, and even the outputs of the black-box model. In this setting, adversarial examples

*Corresponding author

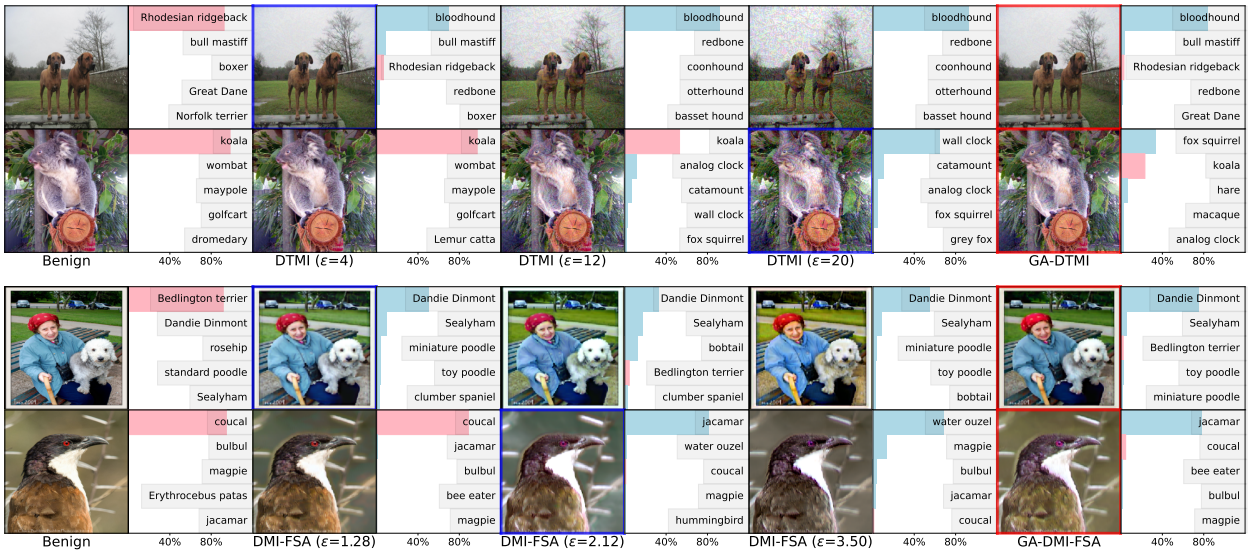

Figure 1: **Comparison between our method and two baselines under both the ℓ_∞-norm (top) and unrestricted (bottom) threat models using various perturbation radii**. In the even columns, we present the top-5 confidence bars of the target model for the images in the left. The ground-truth label is marked by pink and other labels are marked by blue. In each row, the misclassified adversarial example with minimum perturbation radius is highlighted by a blue bounding box, indicating that *the perturbation budgets required for distinct images are different*. Note that the "human-imperceptible" constraint is violated when the ℓ_∞-norm perturbation radius is too large the fifth and seventh columns in the top sub-figure. However, our Geometry-Aware (GA) framework generates transferable unrestricted adversarial examples (highlighted by red bounding boxes) with lower budgets and smaller changes when compared to the benign images.

are typically generated via *transfer-based* methods (Szegedy et al., 2014; Papernot et al., 2016, 2017), e.g., attacking an ensemble of accessible source models and hoping that the same adversarial examples are able to fool the unknown target/test model (Liu et al., 2017; Tramer et al., 2018).

Despite a large amount of work on transfer-based attacks, many fundamental questions remain unresolved. For example, existing transfer-based attacks (Dong et al., 2018; Xie et al., 2019b; Dong et al., 2019) that search for adversarial examples in a fixed-radius ℓ_∞-norm ball often require a high perturbation budget to reach a satisfactory transfer success rate. However, such perturbations might be perceptible to humans (see Figs. 1 and 7). On the other hand, unrestricted adversarial attacks that aim to generate minimum human-imperceptible perturbations (Xiao et al., 2018; Wong et al., 2019; Laidlaw et al., 2021) suffer from weaker transferability to the target model. This is in part due to the difference between the decision boundaries of the source and target models. Given the trade-off between transferability and imperceptibility, one of the long-standing questions is generating transferable adversarial examples by minimum changes of natural examples.

1.1 Our methodology and results

In this work, we propose a novel geometry-aware framework, into which fixed-budget methods can be integrated to generate transferable unrestricted adversarial examples with minimum changes.

Our intuition is that the smallest perturbation budgets w.r.t. distinct images should be different (see Fig. 1) and should depend on their geometrical relationship with the decision boundary of the target model (see Fig. 2). Unfortunately, finding transferable minimum-budget adversarial perturbations is an intractable optimization problem (see Eq (4.1)) as the target model is unknown. We approximately solve this problem by discretizing the continuous space of perturbation radius into a finite set and choosing the minimum perturbation budget that is able to fool the test model. The main challenge here is to evaluate whether a given perturbation can transfer well to the unknown target model (Cheng et al., 2019b; Katzir and Elovici, 2021).

To overcome this challenge, we split all accessible white-box source models into training and validation sets, where adversarial perturbations are crafted only on the training set. We use the validation set to select the smallest perturbation radius for each input that suffices to fool the validation model with a certain confidence level through an *early-stopping* mechanism. When the training (or validation) set consists of multiple models, we use their average ensemble (Liu et al., 2017). Experimentally, our method yields a significant performance boost on the trade-off between transferability and imperceptibility. As shown in Fig. 4, the transfer success rate of our method GA-DTMI-FGSM surpasses the baseline DTMI-FGSM (see Eq (3.1)) by up to 16% in absolute value under the same average perturbation reward (see Eq (3.4)). Besides, our method GA-DMI-FSA is able to generate semantic-preserving yet transferable adversarial examples under the unrestricted threat model (see Eq (3.2), Figs. 1, 6, 7 and 9).

1.2 Summary of our contributions

- We propose a Geometry-Aware (GA) framework, where fixed-budget attacking methods can be integrated, to generate transferable unrestricted adversarial examples with approximately minimum changes. To the best of our knowledge, we are the first to explore transfer-based black-box attacks with adaptive perturbation budgets.

- Under ℓ_∞-norm setting, our geometry-aware framework improves the imperceptibility of the crafted adversarial examples by a large margin without the decrease of transfer success rate (see Fig. 4). By applying our method GA-DTMI-FGSM to the *CVPR'21 Security AI Challenger: Unrestricted Adversarial Attacks on ImageNet* (Chen et al., 2021), we ranked 1st place out of 1,559 teams and surpassed the runner-up submissions by 4.59% and 23.91% in terms of final score and average image quality level, respectively.

- Under unrestricted setting, we propose a transfer-based unrestricted attack (see Eq (3.2)) by combining the *white-box* feature space attack (Xu et al., 2021) with the transfer-based ℓ_∞-norm attacks to generate semantic-preserving yet transferable adversarial examples (see Figs. 1 and 6). Moreover, the crafted adversarial examples can even transfer to adversarially trained robust models (see Tab. 3, Figs. 7 and 9).

2 Related Work

ℓ_p-norm adversarial examples. Existing gradient-based white-box attacks either search for adversarial examples in a fixed ℓ_p-norm ball (Madry et al., 2018; Kurakin et al., 2019), or optimize

the perturbation for each image independently to get a *minimum-norm* solution such as DeepFool (Moosavi-Dezfooli et al., 2016), CW attack (Carlini and Wagner, 2017), and fast adaptive boundary attack (Croce and Hein, 2020). However, white-box assumption usually does not hold in real-world scenarios. In query-based black-box threat model, attackers utilize output logits (Chen et al., 2017; Andriushchenko et al., 2020) or predicted label (Brendel et al., 2018; Cheng et al., 2019a) of the target model to generate adversarial examples. But these query-based attacks typically suffer from high query complexity, making it easy to be detected (Willmott et al., 2021). Transfer-based black-box attacks can pose serious threats in practice as they need no information about the defense models. Dong et al. (2018) boosted transferability by integrating momentum into gradient-based methods. Liu et al. (2017) found that attacking a group of substituted source models simultaneously can improve transferability. Besides, transferability benefits from input transformations such as input diversity (Xie et al., 2019b) and translation-invariant method (Dong et al., 2019).

Unrestricted adversarial examples. The ℓ_p-norm distance is not an ideal perceptual similarity metric (Johnson et al., 2016; Isola et al., 2017), which oversimplifies the diversity of real-world perturbations. Unrestricted adversarial examples have received significant attention in recent years (Brown et al., 2018). Most of the current unrestricted attacks aim to generate imperceptible adversarial examples under *white-box* setting, such as geometric transformations (Xiao et al., 2018; Alaifari et al., 2019; Engstrom et al., 2019) and distance metrics beyond ℓ_p norm (Wong et al., 2019; Laidlaw et al., 2021). Color-based attacks (Hosseini and Poovendran, 2018; Laidlaw and Feizi, 2019; Zhao et al., 2020a,b; Shamsabadi et al., 2020; Bhattad et al., 2020) were also proposed to generate large but imperceptible perturbations, however, the modified color can sometimes be unnatural. Instead of optimizing in the input space directly, generative approaches (Song et al., 2018; Gowal et al., 2020; Qiu et al., 2020; Wong and Kolter, 2021) search for adversarial embeddings in the latent space. Style transfer (Prabhu et al., 2018; Bhattad et al., 2020) is inherently an unrestricted attack as it preserves the semantic of the content image. However, constructing transferable unrestricted adversarial examples under black-box settings is still less explored. In this work, we will fill this gap by combining the white-box feature space attack (Xu et al., 2021) with the transfer-based ℓ_∞-norm attacks to generate semantic-preserving yet transferable adversarial examples, which is directly an extension of our previous technical report (Liu et al., 2021a) that mainly focus on ℓ_∞-norm setting.

Adversarial defenses. There have been long-standing arms races between defenders and attackers. *Adversarial training* (Goodfellow et al., 2015) is one of the most promising defense methods. Many variants of adversarial training framework were proposed, e.g., ensemble adversarial training (Liu et al., 2017) for transfer-based attacks, PGD-based adversarial training (Madry et al., 2018), and TRADES (Zhang et al., 2019) with a new robust loss based on the trade-off between robustness and accuracy. Geometry-aware instance-reweighted adversarial training (Zhang et al., 2021), which is proven falling into gradient masking (Hitaj et al., 2021), shares similar insights with us that the importance of distinct inputs in adversarial training should be different. Laidlaw et al. (2021) integrate adversarial training with Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al., 2018), aiming to improve robustness against perturbations that were unseen during training. Unlike empirical defenses, Certified defenses (Cohen et al., 2019; Salman et al., 2019; Zhang et al., 2020; Leino et al., 2021) could provide robustness guarantee under a certain ℓ_p-norm budget.

3 Preliminaries

Notation. A deep neural network classifier can be described as a function $f(\boldsymbol{x};\boldsymbol{\theta}) : \mathcal{X} \to \mathbb{R}^C$, parameterized by weights $\boldsymbol{\theta}$, which maps a vector $\boldsymbol{x} \in \mathcal{X}$ to its output logits. Given an input $\boldsymbol{x}$ of class $y \in \{1, 2, \cdots, C\}$, the predicted label of $f(\boldsymbol{x};\boldsymbol{\theta})$ is $\widehat{f}(\boldsymbol{x}) := \arg\max_j f_j(\boldsymbol{x};\boldsymbol{\theta})$, where $f_j(\boldsymbol{x};\boldsymbol{\theta})$ represents the j-th entry of $f(\boldsymbol{x};\boldsymbol{\theta})$. We use $L\left(f(\boldsymbol{x};\boldsymbol{\theta}), y\right)$ to represent the cross-entropy loss and denote the ε-neighborhood of $\boldsymbol{x}$ by $\mathbb{B}(\boldsymbol{x}, \varepsilon) := \{\boldsymbol{x}' \in \mathcal{X} : \mathcal{D}(\boldsymbol{x}, \boldsymbol{x}') \leq \varepsilon\}$, where $\mathcal{D}$ is a distance metric that describes the change between the adversarial example $\boldsymbol{x}'$ and the nature example $\boldsymbol{x}$. We denote the black-box test model by g, and split the set of accessible source models into the set of training models f and the set of validation models h.

3.1 Transfer-based ℓ_∞-norm attacks

Existing transfer-based attacks typically search for adversarial examples in a fixed-radius ℓ_p-norm ball, i.e., $\mathcal{D}(\boldsymbol{x}, \boldsymbol{x}') = \|\boldsymbol{x}' - \boldsymbol{x}\|_p \leq \varepsilon$. Various methods were proposed to boost transferability of the generated adversarial examples, such as input Diversity Iterative Fast Gradient Sign Method (DI-FGSM) (Xie et al., 2019b), Momentum-based Iterative (MI-FGSM) method (Dong et al., 2018) and Translation-invariant Iterative (TI-FGSM) method (Dong et al., 2019). We formulate a strong ℓ_∞-norm baseline DTMI-FGSM by combining all these techniques, i.e.,

$$\begin{aligned} \boldsymbol{m}_{t+1} &= \gamma \cdot \boldsymbol{m}_t + \frac{\mathbf{W} * \nabla_{\boldsymbol{x}_t} L\left(f\left(T(\boldsymbol{x}_t, p);\boldsymbol{\theta}\right), y\right)}{\|\mathbf{W} * \nabla_{\boldsymbol{x}_t} L\left(f\left(T(\boldsymbol{x}_t, p);\boldsymbol{\theta}\right), y\right)\|_1}, \\ \boldsymbol{x}_{t+1} &= \Pi_{\mathbb{B}(\boldsymbol{x},\varepsilon)}\left(\boldsymbol{x}_t + \alpha \cdot \mathtt{sign}(\boldsymbol{m}_{t+1})\right), \end{aligned} \tag{3.1}$$

where $\boldsymbol{m}_0 = \mathbf{0}$, $\mathbf{W}$ is a pre-defined kernel with a convolution operation $*$, α is the step size, Π is the projection operator, and γ is the decay factor for the momentum term. $T(\boldsymbol{x}_t, p)$ represents the input transformation on $\boldsymbol{x}_t$ with probability p. When $\gamma = 0$, DTMI-FGSM attack degenerates to the DTI-FGSM attack. When $p = 0$, DTMI-FGSM attack degenerates to the DMI-FGSM attack.

3.2 Transfer-based unrestricted attack

Inspired by prior work (Huang and Belongie, 2017) in style transfer, Xu et al. (2021) tries to find stylized adversarial examples by assuming that the image pairs from the same class share consistent content and differ mainly in their styles. Here we propose to generate semantic-preserving yet transferable unrestricted adversarial examples by combining the Feature Space Attack (FSA) (Xu et al., 2021) with transfer-based ℓ_∞-norm attacks (Dong et al., 2018; Xie et al., 2019b). Given an encoder ϕ, we extract the style features of input $\boldsymbol{x}$ as channel-wise mean $\boldsymbol{\mu}\left(\phi(\boldsymbol{x})\right) \in \mathbb{R}^C$ and channel-wise standard deviation $\boldsymbol{\sigma}\left(\phi(\boldsymbol{x})\right) \in \mathbb{R}^C$, Specifically,

$$\mu_c = \frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W}\phi_c(\boldsymbol{x})_{hw}, \qquad \sigma_c = \sqrt{\frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W}\left(\phi_c(\boldsymbol{x})_{hw} - \mu_c\right)^2},$$

where $\phi(\boldsymbol{x}) \in \mathbb{R}^{C\times H\times W}$ represents the latent embedding. Xu et al. (2021) adds adversarial perturbations on $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ before projecting $\phi(\boldsymbol{x})$ back to the input space $\mathcal{X}$ with a pre-trained[1]

[1]We use the official pre-trained shallowest decoder: https://github.com/qiulingxu/FeatureSpaceAttack.

decoder ϕ^{-1}, namely,

$$\begin{aligned}\widetilde{\phi}(\boldsymbol{x}) &= \left(e^{\boldsymbol{\tau}^{\sigma}} \cdot (\phi(\boldsymbol{x}) - \boldsymbol{\mu}) + e^{\boldsymbol{\tau}^{\mu}} \cdot \boldsymbol{\mu},\right) \\ \boldsymbol{x}' &= \phi^{-1}\left(\widetilde{\phi}(\boldsymbol{x})\right), \quad \|\boldsymbol{\tau}^{\mu}\|_{\infty} \leq \ln \varepsilon, \|\boldsymbol{\tau}^{\sigma}\|_{\infty} \leq \ln \varepsilon,\end{aligned} \tag{3.2}$$

where $\widetilde{\phi}(\boldsymbol{x})$ enlarges or shrinks the mean $\boldsymbol{\mu}$ and the standard deviation $\boldsymbol{\sigma}$ of the embedding $\phi(\boldsymbol{x})$ by a factor of $e^{\boldsymbol{\tau}^{\mu}}$ and $e^{\boldsymbol{\tau}^{\sigma}}$, respectively. In this way, the distance metric $\mathcal{D}(\boldsymbol{x}, \boldsymbol{x}') = \max\left(e^{\|\boldsymbol{\tau}^{\mu}\|_{\infty}}, e^{\|\boldsymbol{\tau}^{\sigma}\|_{\infty}}\right) \leq \varepsilon$. In order to preserve the semantic of the unrestricted adversarial example $\boldsymbol{x}'$, an additional content loss was added during the attacking process, i.e.,

$$\min_{\boldsymbol{\tau}^{\mu}, \boldsymbol{\tau}^{\sigma}} \mathcal{L}(\boldsymbol{x}', y) = \lambda \cdot \mathcal{L}_{\text{top-5}}\left(f(\boldsymbol{x}'; \boldsymbol{\theta}), y\right) + \|\phi(\boldsymbol{x}') - \widetilde{\phi}(\boldsymbol{x})\|_2,$$

where λ balance the trade-off between adversarial loss and the content loss. Following Xu et al. (2021), we set $\lambda = 128$ and use the top-5 margin loss for adversarial attack. With all above, the unrestricted attack (see Eq (3.2)) can be solved by conventional ℓ_{∞}-norm attack on parameters $\boldsymbol{\tau}^{\mu}$ and $\boldsymbol{\tau}^{\sigma}$. Moreover, the same techniques in Sec. 3.1 such as input diversity $T(\boldsymbol{x}', p)$ (only for the margin loss) and momentum-based method can be integrated to improve transferability, i.e.,

$$\begin{aligned}\boldsymbol{m}_{t+1} &= \gamma \cdot \boldsymbol{m}_t + \frac{\nabla_{\boldsymbol{\tau}_t} \mathcal{L}\left(T(\boldsymbol{x}', p), y\right)}{\|\nabla_{\boldsymbol{\tau}_t} \mathcal{L}\left(T(\boldsymbol{x}', p), y\right)\|_1}, \\ \boldsymbol{\tau}_{t+1} &= \Pi_{\mathbb{B}(\boldsymbol{x}, \varepsilon)}\left(\boldsymbol{\tau}_t - \alpha \cdot \texttt{sign}(\boldsymbol{m}_{t+1})\right),\end{aligned} \tag{3.3}$$

where $\boldsymbol{\tau}$ represents $\boldsymbol{\tau}^{\mu}$ or $\boldsymbol{\tau}^{\sigma}$. When $\gamma = 0$, the DMI-FSA attack degenerates to the DI-FSA attack.

3.3 Evaluation metric for transfer-based attack

Different from transferability, the imperceptibility of adversarial examples is hard to evaluate due to the lack of precise quantization of human perception (Wang et al., 2004). Sharif et al. (2018) found that the ℓ_p-norm distance is not an ideal perceptual similarity metric and suggest setting adaptive perturbation budget for every sample to ensure that the attacks' output would be imperceptible. Therefore, we choose transfer success rate and the perturbation budget under certain distance metric $\mathcal{D}$ as our main evaluation metrics. Consider a dataset $\widehat{S} = \{(\boldsymbol{x_i}, y_i)\}_{i=1}^n$ and the corresponding adversarial examples $\widehat{S}_{\text{adv}} = \{(\boldsymbol{x_i'}, y_i)\}_{i=1}^n$ that are crafted on the training model f. Let $n_0 = \sum_{i=1}^n \mathbb{1}\{\widehat{g}(\boldsymbol{x}_i') \neq y_i\}$ be the number of misclassified adversarial examples on the test model g. We define the average total score as follows:

$$\begin{aligned}S_{total} &= \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{\widehat{g}(\boldsymbol{x}_i') \neq y_i\} \cdot \mathcal{F}_{\text{reward}}\left(\mathcal{D}(\boldsymbol{x}_i, \boldsymbol{x}_i')\right) \\ &= \frac{n_0}{n} \cdot \frac{1}{n_0} \sum_{i=1}^{n} \mathbb{1}\{\widehat{g}(\boldsymbol{x}_i') \neq y_i\} \cdot \mathcal{F}_{\text{reward}}\left(\mathcal{D}(\boldsymbol{x}_i, \boldsymbol{x}_i')\right) \\ &\overset{\text{def}}{=} \frac{n_0}{n} \cdot S_{APR},\end{aligned} \tag{3.4}$$

where S_{APR} is the Average Perturbation Reward of adversarial examples that are misclassified by test model g and the reward function $\mathcal{F}_{\text{reward}}$ is a decreasing function w.r.t. the metric $\mathcal{D}(\boldsymbol{x}, \boldsymbol{x}')$.

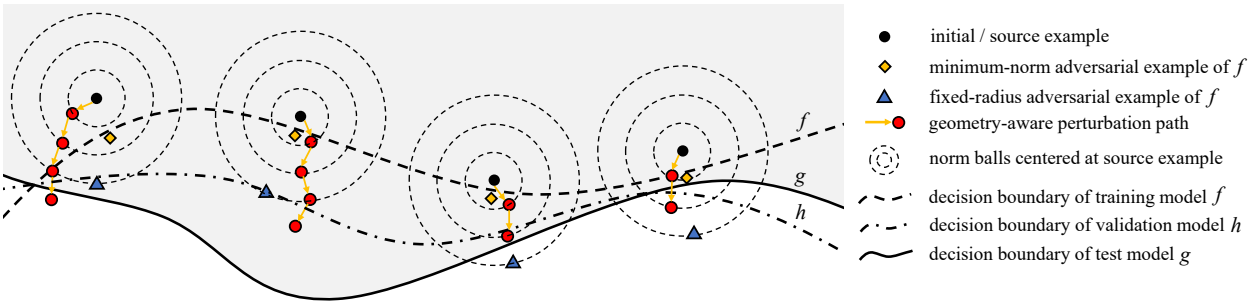

Figure 2: **Our geometry-aware framework**. Existing fixed-budget methods typically overlook the importance of geometrical distances from inputs to the decision boundary of the test model g. In contrast, our geometry-aware framework aims to find geometry-aware minimal-change perturbation via a validation model h. The goal of the validation model is to prevent an attack algorithm overfitting f by forcing the solution to cross the decision boundary of h with a certain margin. Our framework consists of multiple sub-procedures with adaptive perturbation budgets. In each sub-procedure, we start from the solution of the last sub-procedure (the red solid points) and re-run the attack algorithm on the training model f. The procedure stops if the output probability of the true class on the validation model h is smaller than a certain threshold η.

4 Methodology: Geometry-Aware Framework

Eq (3.4) factorizes the average total score as the product of transfer success rate and average perturbation reward, which motivates us to find the adversarial example with minimum changes under metric $\mathcal{D}$, i.e.,[2]

$$\min_{\boldsymbol{x}'} \ \mathcal{D}(\boldsymbol{x}, \boldsymbol{x}') \qquad \text{s.t. } \widehat{g}(\boldsymbol{x}') \neq y. \tag{4.1}$$

However, direct optimization of problem (4.1) is intractable, in part due to the lack of information about test model g. We approximately solve this problem by discretizing the continuous space of perturbation radius into a discrete set and choosing the minimum perturbation budget such that the attack is able to fool the test model g. However, the challenge is that it is typically difficult to decide whether a given perturbation radius can also fool the test model (Cheng et al., 2019b; Katzir and Elovici, 2021). This problem is also known as model selection problem, and a classic approach to tackle this problem is to use a validation model h to select the right perturbation radius. More specifically, we split all accessible source models into training model set and validation model set. With validation model h, we are able to generate transferable adversarial examples with smaller changes. To approximately solve problem (4.1), we first divide the attack in the ball $\mathbb{B}(\boldsymbol{x}, \varepsilon)$ into K sub-procedures. In the k-th sub-procedure, we re-run a fixed-radius attack algorithm $\mathcal{A}$ such as DMI-FSA (see Eq (3.3)) under the perturbation budget

$$\varepsilon_k = \frac{k}{K} \times \varepsilon, \ k = 1, 2, \cdots, K.$$

We start each sub-procedure from the solution of the last sub-procedure to accelerate the convergence. To obtain a minimum-radius solution, we perform an early-stopping mechanism at the end of each

[2] In contrast to existing white-box minimum-norm attack (Carlini and Wagner, 2017), g is an unknown model here.

sub-procedure if the probability of the true class on the validation model h is smaller than a certain threshold η, i.e.,

$$\mathbb{P}\left(\widehat{h}(\boldsymbol{x}) = y\right) = \frac{\exp\left(h_y(\boldsymbol{x};\boldsymbol{\theta})\right)}{\sum_j \exp\left(h_j(\boldsymbol{x};\boldsymbol{\theta})\right)} < \eta. \tag{4.2}$$

Our Geometry-Aware (GA) framework is summarized in Algorithm 1 and illustrated in Fig. 2.

Algorithm 1 Geometry-Aware Framework

Require:

Benign input $\boldsymbol{x}$ with label y; training models f; validation model h; number of sub-procedures K; maximum perturbation size ε and threshold η; attack algorithm $\mathcal{A}$;

Ensure:

Transfer-based unrestricted adversarial example $\boldsymbol{x}'$ with approximately minimum change;

$\boldsymbol{x}_0 = \boldsymbol{x}$;
for $k = 1, 2, \cdots, K$ **do**
  $\boldsymbol{x}_k = \mathcal{A}(\boldsymbol{x}, \boldsymbol{x}_{k-1}, f, \frac{k\varepsilon}{K})$; ▷ `fixed-budget attack`
  conf $\leftarrow \frac{\exp(h_y(\boldsymbol{x}_k;\boldsymbol{\theta}))}{\sum_j \exp(h_j(\boldsymbol{x}_k;\boldsymbol{\theta}))}$;
  if conf $< \eta$ **then**
    Return $\boldsymbol{x}_k$; ▷ `early-stopping in Eq` (4.2)
  end if
end for
Return $\boldsymbol{x}_K$;

5 Experiments

5.1 Experimental setup

Datasets & Networks. Similar to Xie et al. (2019b), we randomly select 1,000 images from ILSVRC 2012 validation set (Deng et al., 2009), which are almost correctly classified by all the attacking models. All these images are resized to $229 \times 229 \times 3$ beforehand. We consider eight normally trained models and two ensemble adversarially trained models (Tramer et al., 2018). The weights of all these models are publicly available at Wightman (2019). More details about the networks are summarised in Tab. 1. The transferability between these models under ℓ_∞-norm setting is summarised in Fig. 3. It is much easier for the generated adversarial examples to transfer from vision transformers to convolutional neural networks (CNNs), which is consistent with the empirical observation in Shao et al. (2021). Surprisingly, the robustness of naturally trained vision transformers under transfer attack is even on par with two ensemble adversarially trained CNNs.

Implementation details. Given the maximum perturbation size ε and number of sub-procedures K (5 as default) in our geometry-aware framework, we set the step size $\alpha = \frac{1.25\times\varepsilon_k}{T}$ in the k-th sub-procedure, where the number of iterations T is set to 10 in the ℓ_∞ setting and 50 in the

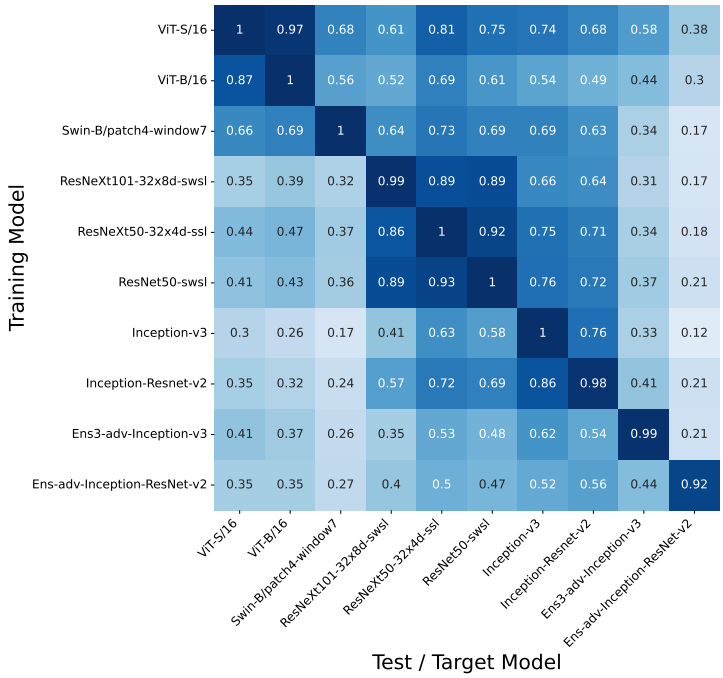

Figure 3: **Transferability between different networks under DTMI-FGSM attack**. The rows stand for source models and the columns stand for target models. Adversarial examples transfer well between models with similar architectures.

Table 1: **An overview of all considered networks for generating adversarial examples**. Top-1_{ImageNet} represents the accuracy on the ILSVRC 2012 validation set while Top-1_{1000} represents the accuracy on the randomly selected 1000 images.

Training	Index	Model Name	Top-1_{ImageNet}	Top-1_{1000}	Index	Model Name	Top-1_{ImageNet}	Top-1_{1000}
Normal	0	ViT-S/16	76.01%	99.8%	1	ViT-B/16	81.08%	99.1%
	2	Swin-B/patch4-window7	**84.23%**	99.4%	3	ResNeXt101-32x8d-swsl	83.62%	99.9%
	4	ResNeXt50-32x4d-ssl	78.90%	99.7%	5	ResNet50-swsl	79.97%	99.5%
	6	Inception-v3	76.94%	100.0%	7	Inception-ResNet-v2	79.85%	99.9%
Ensemble	8	Ens3-adv-Inception-v3	76.49%	100.0 %	9	Ens-adv-Inception-ResNet-v2	78.98%	99.9%

unrestricted setting. ε is set to 20 in the ℓ_∞-norm setting and 3.5 in the unrestricted setting. When running a fixed-radius baseline at perturbation budget ε_k, we set the number of iterations $N = \frac{T}{2}\left(1 + \frac{K\varepsilon_k}{\varepsilon}\right)$ with step size α to keep the same total perturbation budget (the sum of step size across all iterations) as our geometry-aware framework for fair comparison. The reward function $\mathcal{F}_{\text{reward}}(\varepsilon_0)$ is set to $1/\varepsilon_0$ as smaller perturbation radius exhibits significantly higher image quality. For the momentum term, we set the decay factor $\mu = 1$ as in Dong et al. (2018). For DI-FGSM (Xie et al., 2019b), we set the transformation probability to $p = 0.7$. The input is first randomly resized to be an $r \times r \times 3$ image with $r \in [(1-\gamma)s, (1+\gamma)s]$, and then padded to size $(1+\gamma)s \times (1+\gamma)s \times 3$. The transformed input is then resized to $s \times s \times 3$ for different input size s of various models, i.e., 224, 299 and 384. We set $\gamma = 0.1$ as default. For TI-FGSM (Dong et al., 2019), we use Gaussian kernel with kernel size 5×5.

5.2 Balancing transfer success rate and average perturbation reward

Implementation details. Benefiting from the adaptive choice of perturbation budgets, our geometry-aware framework can generate transferable unrestricted adversarial examples with smaller changes. In this experiment, the training model f and validation model h are an ensemble of models $\{2,3,5\}$ and $\{1,4,6\}$ in Tab. 1, respectively. The test model is Inception-ResNet-v2. The optimal threshold η (see Eq (4.2)) is searched from a finite set ranging from 0.001 to 0.9 by querying[3] the test model g to achieve the best average total score S_{total}. For each η, we execute our method and compute the average perturbation reward S_{APR} (the x-axis of each red point in Fig. 4). Then the corresponding fixed-radius baseline is run at the same x-axis. We conduct experiments on two threat models. For the ℓ_∞-norm setting, we combine our Geometry-Aware (GA) framework with DI-FGSM, DTI-FGSM, and DTMI-FGSM, named GA-DI-FGSM, GA-DTI-FGSM, and GA-DTMI-FGSM, respectively; For the unrestricted setting, we combine our GA framework with DMI-FSA and DI-FSA, named GA-DMI-FSA and GA-DI-FSA, respectively.

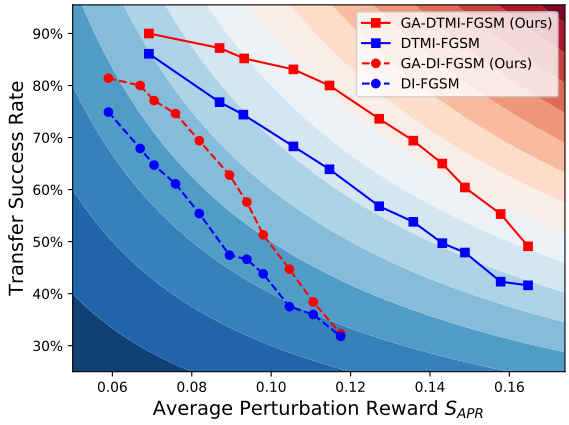

Figure 4: **Contour of average total score S_{total} (see Eq (3.4)).** Fixing transfer success rate as 80%, our approach GA-DTMI-FGSM surpasses the baseline DTMI-FGSM (see Eq (3.1)) by up to 43.35% in terms of average perturbation reward.

Table 2: **Comparison of our method with baselines**. We report the results when both our approach and baselines achieve highest average total score S_{total}. TSR: Transfer Success Rate.

Method	TSR (↑)	S_{APR} (↑)	S_{total} (↑)
DI-FGSM	61.1%	0.0759	4.64%
GA-DI-FGSM (ours)	**69.4%**	**0.0819**	**5.68%**
DTI-FGSM	57.3%	0.1101	5.68%
GA-DTI-FGSM (ours)	**67.9%**	**0.1176**	**7.98%**
DTMI-FGSM	63.9%	0.1147	7.33%
GA-DTMI-FGSM (ours)	**69.4%**	**0.1358**	**9.42%**
DI-FSA	48.3%	0.5328	25.73%
GA-DI-FSA (ours)	**50.4%**	**0.5541**	**27.93%**
DMI-FSA	51.3%	**0.5616**	28.81%
GA-DMI-FSA (ours)	**58.3%**	0.5355	**31.32%**

Experimental results We present the contour plot of average total score in Fig. 4, where the improvement of our method upon baselines depends on the choice of hyper-parameter η (leading to different S_{APR}). Fixing S_{APR} as 0.115, our approach GA-DTMI-FGSM surpasses the baseline DTMI-FGSM by up to 16.1% in terms of transfer success rate; As shown in Tab. 2, our approach yields a significant performance boost on the average total score S_{total} across various threat models, especially in the ℓ_∞-norm setting where both the transfer success rate and S_{APR} are improved.

[3]In contrast to conventional query-based attacks that need the logits or predicted label on the target model, we query whether an adversarial example transfers to the target model successfully. Besides, we have strong prior information on η which depends on the similarity between f, h and g, making the query complexity rather limited.

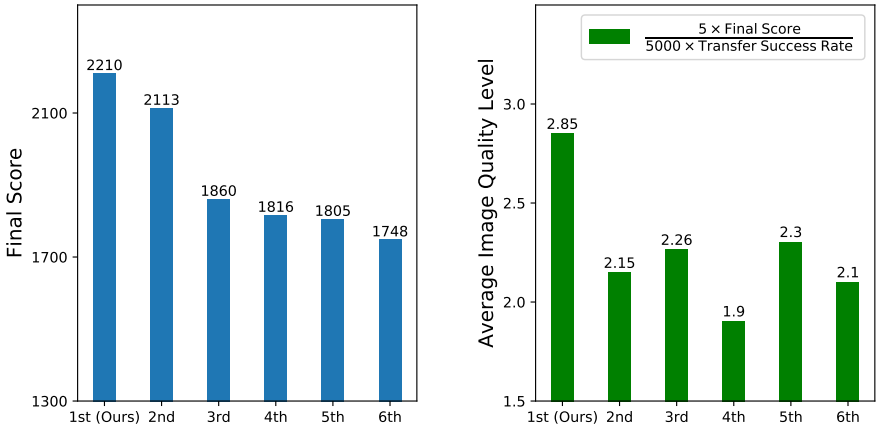

Figure 5: **Top-6 results** (out of 1,559 teams) in the CVPR'21 Security AI Challenger: Unrestricted Adversarial Attacks on ImageNet. The final scores were manually scored by multiple human referees.

5.3 Case study: CVPR'21 Unrestricted Adversarial Attacks on ImageNet

In the *CVPR'21 Security AI Challenger: Unrestricted Adversarial Attacks on ImageNet* (Chen et al., 2021), contestants were asked to submit adversarial examples without any access to the defense models. The dataset is a subset of ILSVRC 2012 validation set (Deng et al., 2009), which consists of 5,000 images with 5 images per class. The final score of each submission was *manually* scored from two aspects: 1) image semantic and 2) quality. If the semantic of the submitted image changes, then $S_s = 0$, otherwise $S_s = 1$. The image quality S_q (equivalent to our reward function $\mathcal{F}_{\text{reward}}$) was quantified with five levels $S_q \in \{1, 2, 3, 4, 5\}$ by multiple human referees. The final score is given by $\sum_i \mathbb{1}\{\widehat{g}(\boldsymbol{x}'_i) \neq y_i\} \times S_s(\boldsymbol{x}'_i) \times \frac{S_q(\|\boldsymbol{x}'_i - \boldsymbol{x}_i\|)}{5}$.

We apply our method GA-DTMI-FGSM ($\eta = 0.01$) to the competition, where our entry ranked 1st place out of 1,559 teams. We report the final score and average image quality level (equivalent to our average perturbation reward) in Fig. 5. It shows that our method outperforms other approaches by a large margin. In particular, we surpass the runner-up submissions by 4.59% and 23.91% in terms of final score and average image quality level, respectively.

5.4 Transferable unrestricted adversarial examples beyond ℓ_p norm

Most of current defenses can be easily broken by unseen attacks in a white-box manner. Adversarial training against multiple ℓ_p-norm attacks (Tramer and Boneh, 2019) solved this issue partially, however, at the cost of robustness against single ℓ_p-norm attack. Laidlaw et al. (2021) integrated adversarial training with Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al., 2018), aiming to improve robustness against perturbations that were unseen during training. However, the proposed attack (Laidlaw et al., 2021), similar to other unrestricted attacks (Laidlaw and Feizi, 2019; Engstrom et al., 2019), suffers from weaker transferability to the target model. In practice, attackers typically have no information about the defense models (thus transfer attack is required) and the defenders do not have the ground truth to make pixel-level comparison (perturbation can be large as long as the generated adversarial examples are semantic-preserving). Therefore, we propose to benchmark classification models on ImageNet under *transfer-based unrestricted attacks*.

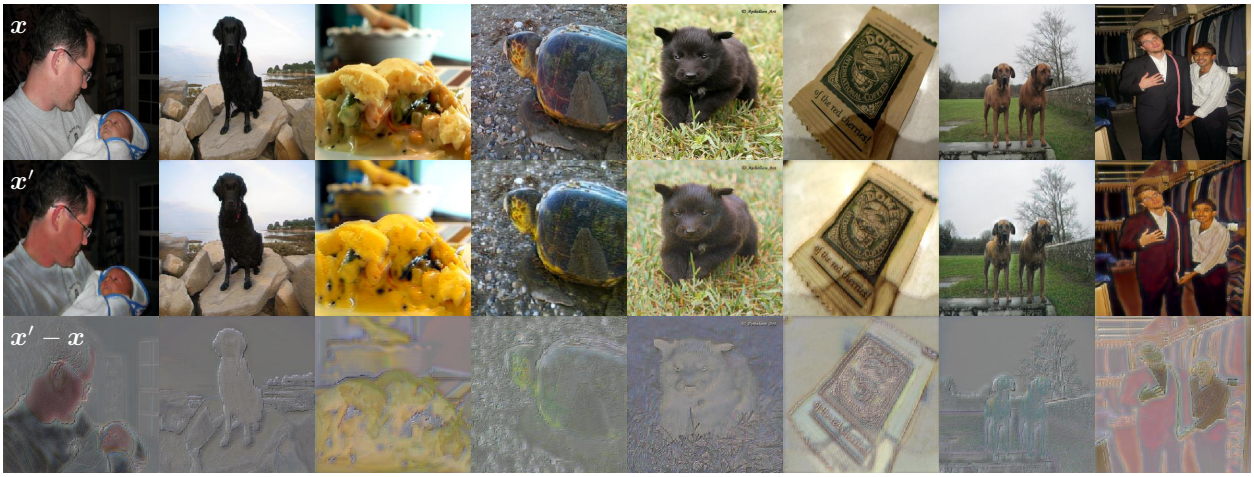

Figure 6: **Adversarial examples of our method GA-DMI-FSA**, which are misclassified by *all* the models in Tab. 3. Top: benign examples $\boldsymbol{x}$. Middle: unrestricted adversarial examples $\boldsymbol{x}'$. Bottom: normalized adversarial perturbations $\boldsymbol{x}' - \boldsymbol{x}$.

Table 3: **Benchmarking classification on Imagenet under transfer-based unrestricted attacks**. PGD^*_{40} indicates the white-box PGD attack (Madry et al., 2018) with 40 iterations ($\varepsilon = 4/255$). We denote the adversarial examples generated by GA-DTMI-FGSM and GA-DMI-FSA in Table 2 as $\text{GA}_{\text{DTMI-FGSM}}$ and $\text{GA}_{\text{DMI-FSA}}$, respectively. FSA represents the feature space attack (Xu et al., 2021). ReColor represents the color-based attack (Laidlaw and Feizi, 2019).

Defenses	Clean	PGD^*_{40}	ReColor	FSA	$\text{GA}_{\text{DTMI-FGSM}}$	$\text{GA}_{\text{DMI-FSA}}$
Inception-ResNet-v2$_{\text{Ens-adv}}$ (Tramer et al., 2018)	99.9%	10.0%	93.2%	90.0%	87.1%	43.5%
Fast_{AT} (Wong et al., 2020)	66.9%	35.7%	62.8%	59.3%	64.7%	**28.5%**
Free_{AT} (Shafahi et al., 2019)	77.3%	40.6%	71.6%	68.4%	74.0%	35.1%
Resnet152-Base (Xie et al., 2019a)	67.6%	39.0%	64.1%	61.2%	65.1%	37.4%
Resnext101-DenoiseAll (Xie et al., 2019a)	80.3%	**52.2%**	77.0%	73.8%	78.7%	47.8%
Resnet152-Denoise (Xie et al., 2019a)	72.2%	41.8%	68.2%	65.2%	70.7%	40.3%
RVT-Tiny (Mao et al., 2021)	96.7%	0.0%	78.9%	81.3%	44.5%	33.6%
DeepAugment+AugMix (Hendrycks et al., 2021)	96.1%	0.0%	82.8%	89.3%	58.9%	63.2%
Efficientnet-l2-ns (Xie et al., 2020)	99.2%	0.0%	95.7%	97.0%	81.3%	**86.0%**
Swin-L/patch4-window-12 (Liu et al., 2021b)	99.1%	0.0%	88.4%	90.7%	66.6%	61.8%

Implementation details For the ReColor attack (Laidlaw and Feizi, 2019), we set $\varepsilon = 1.0$ and iterations $T = 100$ which achieves 89.7% attack success rate on training model f (the same as Sec. 5.2) and 9.1% transfer success rate on the test model Inception-ResNet-v2. For FSA attack (Xu et al., 2021), we set $\varepsilon = 3.0$ and $T = 500$ which achieves 54.2%[4] attack success rate and 12.4% transfer success rate on the same training and test models. Besides six adversarially trained and two high-performance classification models, we select two state-of-the-art models on the ImageNet-R dataset (Hendrycks et al., 2021). Note that it's hard to compare the robustness across different attacks (columns in Tab. 3) due to the lack of a consistent perceptual metric.

[4]Note that our method GA-DMI-FSA achieves 95.5% attack success rate and 58.3% transfer success rate, indicating that the input diversity and momentum techniques in Eq (3.3) boost both the attacking ability and transferability.

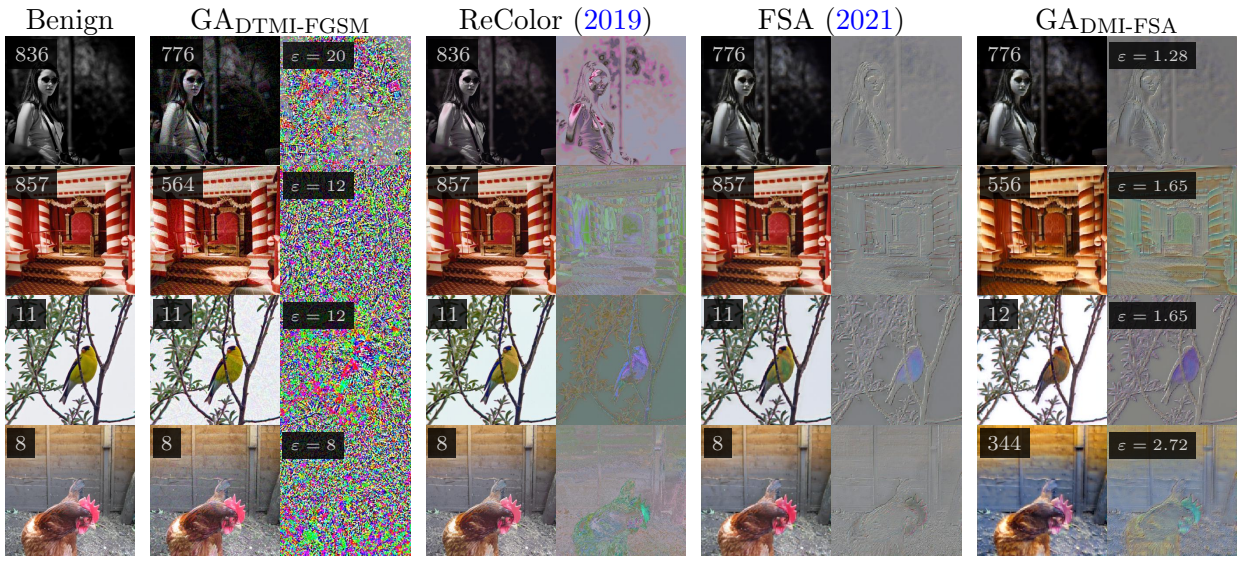

Figure 7: **Visualization of transfer attack results on Resnext101-DenoiseAll** (Xie et al., 2019a). For each image, we print its predicted label on model Resnext101-DenoiseAll in the upper left corner. For each transfer-based adversarial example, we present the perturbation on its right. For each perturbation crafted via our geometry-aware framework, we print its perturbation budget in the upper left corner. Although the transfer-based ℓ_∞-norm attack GA-DTMI-FGSM is able to fool the defense test model to a certain extent, the generated perturbations can be "human-perceptible" (the first and third rows of GA$_{\text{DTMI-FGSM}}$). Besides, the other two unrestricted attacks suffer from weaker transferability when compared to our method GA$_{\text{DMI-FSA}}$, which adjusts the images' color and texture that ImageNet-trained CNNs might be biased to (Geirhos et al., 2018).

Experimental results From Tab. 3, we can conclude the following three observations: 1) ℓ_∞-norm based transfer attack (GA-DTMI-FGSM) can hardly break ℓ_∞-norm adversarially trained models while the unrestricted attack (GA-DMI-FSA) reduces the accuracy of these models by a large margin (see also in Fig. 7). 2) DeepAugment (Hendrycks et al., 2021), which utilizes semantic-preserving augmentations during training, exhibits non-trivial robustness against GA-DMI-FSA attack. 3) Efficientnet-l2-ns (Xie et al., 2020) performs well under all the transfer-based attacks and enjoys 86% accuracy against GA-DMI-FSA attack, showing that the distribution of our generated adversarial examples is not too far from the natural examples'. Note that Efficientnet-l2-ns is also the best-performing model on the ImageNet-V2 dataset (Recht et al., 2019; Taori et al., 2020).

We visualize part of the transfer attack results on adversarially trained model Resnext101-DenoiseAll (Xie et al., 2019a) in Figs. 7 and 9, where our method GA-DMI-FSA is able to generate semantic-preserving yet transferable adversarial examples.

5.5 Ablation studies

The optimal η depends on the combination of f, h, g. Another natural question is how to choose the training model f and validation model h. As illustrated in Fig. 2, the effectiveness of our geometry-aware framework relies on the relationship between the decision boundaries of training

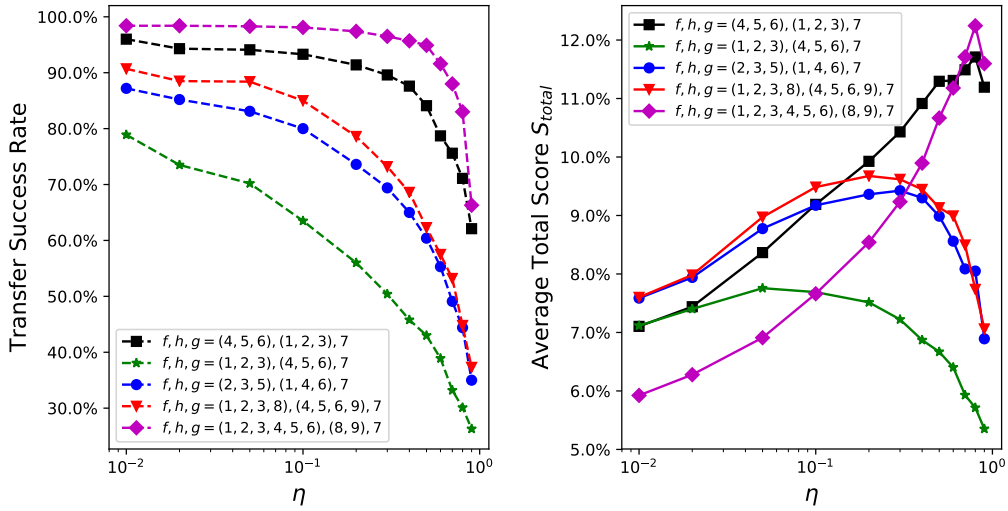

Figure 8: **Comparison between various kinds of partition for training, validation, and test models**. The index in the legend corresponds to the model index in Tab. 1. As the threshold η for early-stopping increases, the generated adversarial examples have higher confidence (probability of the true class) on the validation model h, leading to a lower transfer success rate (**left**). Besides, the optimal η that yields the maximum average total score is dependent on the partition (**right**).

model f, validation model h, and test model g. Thus we compared various kinds of partition of training, validation, and test models in Fig. 8, from which we can conclude the following two observations: 1) The role of training model f and validation model h is not *symmetrical*. The transfer success rate changed a lot when we exchanged the role of models $\{4, 5, 6\}$ and $\{1, 2, 3\}$ for training and validation. 2) Ensembling more models yields better performance. When additional models were added into the training and validation sets, the average total score improved significantly.

6 Conclusion

In this work, we propose a geometry-aware framework, where fixed-radius methods can be integrated to generate transferable unrestricted adversarial examples with minimum changes. Under ℓ_∞-norm setting, our framework improves the imperceptibility of the crafted adversarial examples by a large margin without the decrease of transfer success rate. Besides, we propose a transfer-based unrestricted attack by combining the white-box feature space attack with transfer-based ℓ_∞-norm attacks to generate semantic-preserving yet transferable adversarial examples.

References

Alaifari, R., Alberti, G. S. and Gauksson, T. (2019). ADef: an iterative algorithm to construct adversarial deformations. In *ICLR*.

Andriushchenko, M., Croce, F., Flammarion, N. and Hein, M. (2020). Square attack: A query-efficient black-box adversarial attack via random search. In *ECCV*.

Athalye, A., Carlini, N. and Wagner, D. (2018). Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *ICML*, vol. 1.

Bai, Y., Mei, J., Yuille, A. and Xie, C. (2021). Are transformers more robust than cnns? In *NeurIPS*.

Bhattad, A., Chong, M. J., Liang, K., Li, B. and Forsyth, D. A. (2020). Unrestricted adversarial examples via semantic manipulation. In *ICLR*.

Bhojanapalli, S., Chakrabarti, A., Glasner, D., Li, D., Unterthiner, T. and Veit, A. (2021). Understanding robustness of transformers for image classification. In *ICCV*.

Bojarski, M., Del Testa, D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., Jackel, L. D., Monfort, M., Muller, U., Zhang, J. et al. (2016). End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316* .

Brendel, W., Rauber, J. and Bethge, M. (2018). Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. In *ICLR*.

Brown, T. B., Carlini, N., Zhang, C., Olsson, C., Christiano, P. and Goodfellow, I. (2018). Unrestricted adversarial examples. *arXiv preprint arXiv:1809.08352* .

Carlini, N. and Wagner, D. (2017). Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*. IEEE.

Chen, P.-Y., Zhang, H., Sharma, Y., Yi, J. and Hsieh, C.-J. (2017). Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM workshop on artificial intelligence and security*.

Chen, Y., Mao, X., He, Y., Xue, H., Li, C., Dong, Y., Fu, Q.-A., Yang, X., Xiang, W., Pang, T. et al. (2021). Unrestricted adversarial attacks on imagenet competition. *arXiv preprint arXiv:2110.09903* .

Cheng, M., Le, T., Chen, P.-Y., Zhang, H., Yi, J. and Hsieh, C.-J. (2019a). Query-efficient hard-label black-box attack: An optimization-based approach. In *ICLR*.

Cheng, S., Dong, Y., Pang, T., Su, H. and Zhu, J. (2019b). Improving black-box adversarial attacks with a transfer-based prior. In *NeurIPS*, vol. 32.

Cohen, J., Rosenfeld, E. and Kolter, Z. (2019). Certified adversarial robustness via randomized smoothing. In *ICML*. PMLR.

Croce, F. and Hein, M. (2020). Minimally distorted adversarial examples with a fast adaptive boundary attack. In *ICML*. PMLR.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *CVPR*. IEEE.

Dong, Y., Liao, F., Pang, T., Su, H., Zhu, J., Hu, X. and Li, J. (2018). Boosting adversarial attacks with momentum. In *CVPR*.

Dong, Y., Pang, T., Su, H. and Zhu, J. (2019). Evading defenses to transferable adversarial examples by translation-invariant attacks. In *CVPR*.

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*.

Engstrom, L., Tran, B., Tsipras, D., Schmidt, L. and Madry, A. (2019). Exploring the landscape of spatial robustness. In *ICML*. PMLR.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A. and Brendel, W. (2018). Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231* .

Goodfellow, I. J., Shlens, J. and Szegedy, C. (2015). Explaining and harnessing adversarial examples. In *ICLR*.

Gowal, S., Qin, C., Huang, P.-S., Cemgil, T., Dvijotham, K., Mann, T. and Kohli, P. (2020). Achieving robustness in the wild via adversarial mixing with disentangled representations. In *CVPR*.

He, K., Zhang, X., Ren, S. and Sun, J. (2016). Deep residual learning for image recognition. In *CVPR*.

Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., Song, D., Steinhardt, J. and Gilmer, J. (2021). The many faces of robustness: A critical analysis of out-of-distribution generalization. *ICCV* .

Hitaj, D., Pagnotta, G., Masi, I. and Mancini, L. V. (2021). Evaluating the robustness of geometry-aware instance-reweighted adversarial training. *arXiv preprint arXiv:2103.01914* .

Hosseini, H. and Poovendran, R. (2018). Semantic adversarial examples. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*.

Huang, X. and Belongie, S. (2017). Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*.

Isola, P., Zhu, J.-Y., Zhou, T. and Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *CVPR*.

Johnson, J., Alahi, A. and Fei-Fei, L. (2016). Perceptual losses for real-time style transfer and super-resolution. In *ECCV*. Springer.

Katzir, Z. and Elovici, Y. (2021). Who's afraid of adversarial transferability? *arXiv preprint arXiv:2105.00433* .

Kurakin, A., Goodfellow, I. J. and Bengio, S. (2019). Adversarial examples in the physical world. *ICLR 2017 - Workshop Track Proceedings* 1–14.

Laidlaw, C. and Feizi, S. (2019). Functional adversarial attacks. In *NeurIPS*.

Laidlaw, C., Singla, S. and Feizi, S. (2021). Perceptual adversarial robustness: Defense against unseen threat models. In *ICLR*.

Leino, K., Wang, Z. and Fredrikson, M. (2021). Globally-robust neural networks. In *ICML*.

Liu, F., Zhang, C. and Zhang, H. (2021a). Towards transferable adversarial perturbations with minimum norm. In *ICML 2021 Workshop on Adversarial Machine Learning*.

Liu, Y., Chen, X., Liu, C. and Song, D. (2017). Delving into transferable adversarial examples and black-box attacks. In *ICLR*.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S. and Guo, B. (2021b). Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030* .

Madry, A., Makelov, A., Schmidt, L., Tsipras, D. and Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. In *ICLR*.

Mao, X., Qi, G., Chen, Y., Li, X., Ye, S., He, Y. and Xue, H. (2021). Rethinking the design principles of robust vision transformer. *arXiv preprint arXiv:2105.07926* .

Moosavi-Dezfooli, S.-M., Fawzi, A. and Frossard, P. (2016). Deepfool: a simple and accurate method to fool deep neural networks. In *CVPR*.

Papernot, N., McDaniel, P. and Goodfellow, I. (2016). Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv preprint arXiv:1605.07277* .

Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B. and Swami, A. (2017). Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security*.

Parkhi, O. M., Vedaldi, A. and Zisserman, A. (2015). Deep face recognition. In *BMVC*. BMVA Press.

Prabhu, V. U., Whaley, J. and Francisco, S. (2018). Art-attack! on style transfers with textures, label categories and adversarial examples.

Qiu, H., Xiao, C., Yang, L., Yan, X., Lee, H. and Li, B. (2020). Semanticadv: Generating adversarial examples via attribute-conditioned image editing. In *ECCV*. Springer.

Recht, B., Roelofs, R., Schmidt, L. and Shankar, V. (2019). Do imagenet classifiers generalize to imagenet? In *ICML*. PMLR.

Salman, H., Yang, G., Li, J., Zhang, P., Zhang, H., Razenshteyn, I. and Bubeck, S. (2019). Provably robust deep learning via adversarially trained smoothed classifiers. In *NeurIPS*. Curran Associates Inc., Red Hook, NY, USA.

Shafahi, A., Najibi, M., Ghiasi, M. A., Xu, Z., Dickerson, J., Studer, C., Davis, L. S., Taylor, G. and Goldstein, T. (2019). Adversarial training for free! In *NeurIPS*, vol. 32. Curran Associates, Inc.

Shamsabadi, A. S., Sanchez-Matilla, R. and Cavallaro, A. (2020). Colorfool: Semantic adversarial colorization. In *CVPR*.

Shao, R., Shi, Z., Yi, J., Chen, P.-Y. and Hsieh, C.-J. (2021). On the adversarial robustness of visual transformers. *arXiv preprint arXiv:2103.15670* .

Sharif, M., Bauer, L. and Reiter, M. K. (2018). On the suitability of lp-norms for creating and preventing adversarial examples. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*.

Song, Y., Shu, R., Kushman, N. and Ermon, S. (2018). Constructing unrestricted adversarial examples with generative models. In *NeurIPS*, vol. 31. Curran Associates, Inc.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. and Fergus, R. (2014). Intriguing properties of neural networks. In *ICLR*.

Taori, R., Dave, A., Shankar, V., Carlini, N., Recht, B. and Schmidt, L. (2020). Measuring robustness to natural distribution shifts in image classification. In *NeurIPS*.

Tramer, F. and Boneh, D. (2019). Adversarial training and robustness for multiple perturbations. In *NeurIPS*, vol. 32. Curran Associates, Inc.

Tramer, F., Carlini, N., Brendel, W. and Madry, A. (2020). On adaptive attacks to adversarial example defenses. In *NeurIPS*, vol. 33. Curran Associates, Inc.

Tramer, F., Kurakin, A., Papernot, N., Goodfellow, I., Boneh, D. and McDaniel, P. (2018). Ensemble adversarial training: Attacks and defenses. In *ICLR*.

Wang, Z., Bovik, A. C., Sheikh, H. R. and Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* **13** 600–612.

Wightman, R. (2019). Pytorch image models. https://github.com/rwightman/pytorch-image-models.

Willmott, D., Sahu, A. K., Sheikholeslami, F., Condessa, F. and Kolter, Z. (2021). You only query once: Effective black box adversarial attacks with minimal repeated queries. *arXiv preprint arXiv:2102.00029* .

Wong, E. and Kolter, J. Z. (2021). Learning perturbation sets for robust machine learning. In *ICLR*.

Wong, E., Rice, L. and Kolter, J. Z. (2020). Fast is better than free: Revisiting adversarial training. In *ICLR*.

Wong, E., Schmidt, F. and Kolter, Z. (2019). Wasserstein adversarial examples via projected sinkhorn iterations. In *ICML*. PMLR.

Xiao, C., Zhu, J.-Y., Li, B., He, W., Liu, M. and Song, D. (2018). Spatially transformed adversarial examples. In *ICLR*.

Xie, C., Wu, Y., Maaten, L. v. d., Yuille, A. L. and He, K. (2019a). Feature denoising for improving adversarial robustness. In *CVPR*.

Xie, C., Zhang, Z., Zhou, Y., Bai, S., Wang, J., Ren, Z. and Yuille, A. (2019b). Improving transferability of adversarial examples with input diversity. In *CVPR*.

Xie, Q., Luong, M.-T., Hovy, E. and Le, Q. V. (2020). Self-training with noisy student improves imagenet classification. In *CVPR*.

Xu, Q., Tao, G., Cheng, S. and Zhang, X. (2021). Towards feature space adversarial attack by style perturbation. *AAAI* **35** 10523–10531.

Zhang, H., Chen, H., Xiao, C., Gowal, S., Stanforth, R., Li, B., Boning, D. and Hsieh, C.-J. (2020). Towards stable and efficient training of verifiably robust neural networks. In *ICLR*.

Zhang, H., Yu, Y., Jiao, J., Xing, E. P., Ghaoui, L. E. and Jordan, M. I. (2019). Theoretically principled trade-off between robustness and accuracy. In *ICML*.

Zhang, J., Zhu, J., Niu, G., Han, B., Sugiyama, M. and Kankanhalli, M. (2021). Geometry-aware instance-reweighted adversarial training. In *ICLR*.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E. and Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*.

Zhao, Z., Liu, Z. and Larson, M. (2020a). Towards large yet imperceptible adversarial image perturbations with perceptual color distance. In *CVPR*.

Zhao, Z., Liu, Z. and Larson, M. A. (2020b). Adversarial color enhancement: Generating unrestricted adversarial images by optimizing a color filter. In *BMVC*.

A Appendix

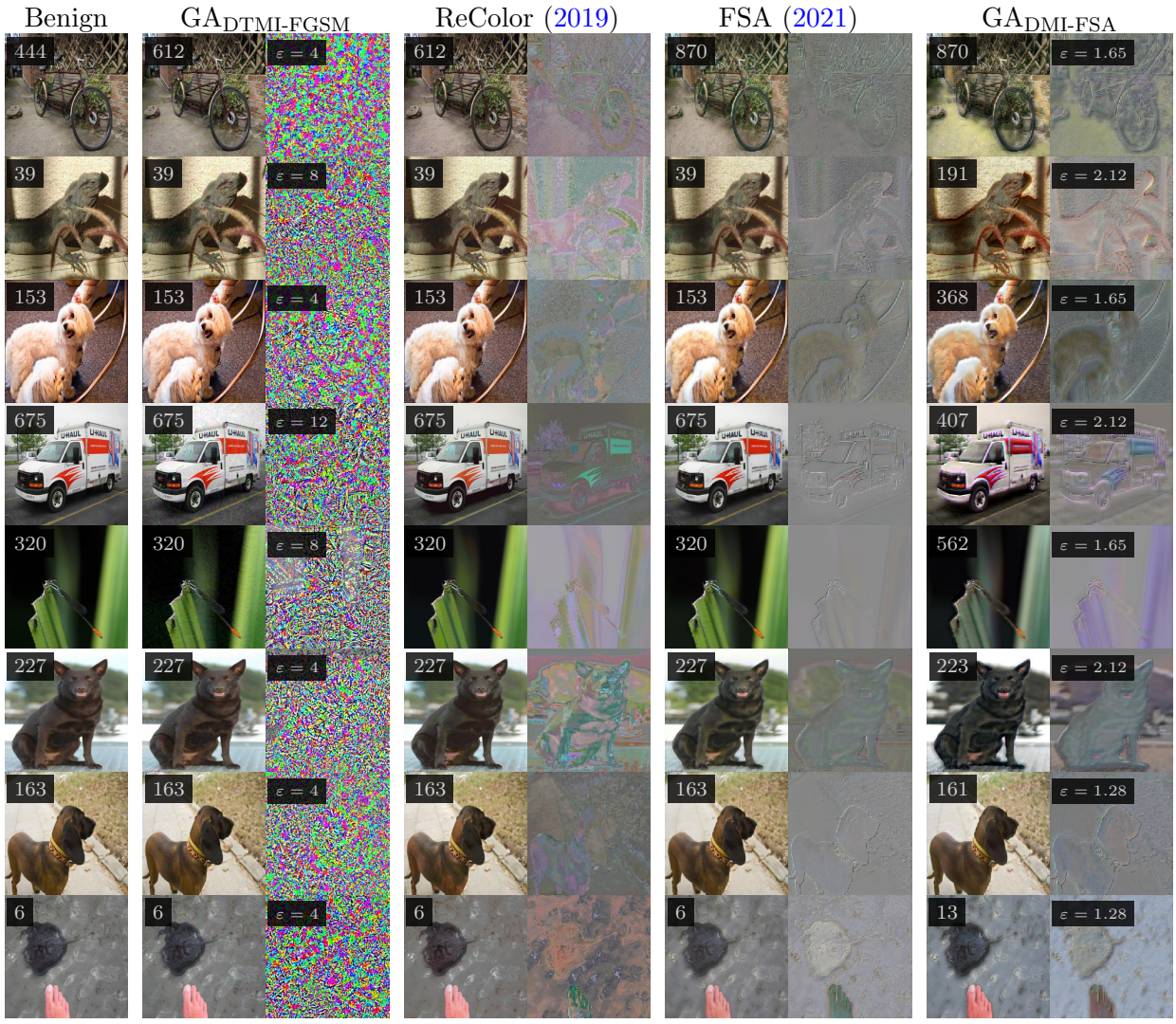

Figure 9: **Visualization of transfer attack on Resnext101-DenoiseAll** (Xie et al., 2019a).