Towards Transferable Unrestricted Adversarial Examples with Minimum Changes

Anonymous submission

***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 best 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 adversarial competition of a 2021 conference, 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.**

***Index Terms*—Transfer-based Black-box Attack, Unrestricted Adversarial Examples**

I. Introduction

Though deep neural networks have exhibited impressive performance in various fields [1], [2], they are vulnerable to adversarial examples [3]–[7], 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 [8] and face recognition [9]. Most positive results on adversarial attacks have focused on white-box settings [10], [11], 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 are typically generated via *transfer-based* methods [3], [12], [13], 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 [14], [15].

Despite a large amount of work on transfer-based attacks, many fundamental questions remain unresolved. For example, existing transfer-based attacks [16]–[18] 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 [19]–[21] 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.

A. 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. (6)) 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 [22], [23].

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 [14]. Experimentally, our method yields a significant performance boost on the trade-off (leading to higher S_{total} in Table. II) 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. (1)) by up to 16% in absolute value under the same average perturbation reward (see Eq. (5)). Besides, our method GA-DMI-FSA is able to generate semantic-preserving yet transferable unrestricted adversarial examples under the unrestricted threat model (see Eq. (3), Figs. 1, 6, 7 and 9).

B. Summary of Our Contributions

- We propose a Geometry-Aware (GA) framework, where fixed-budget attacking methods can be integrated, to

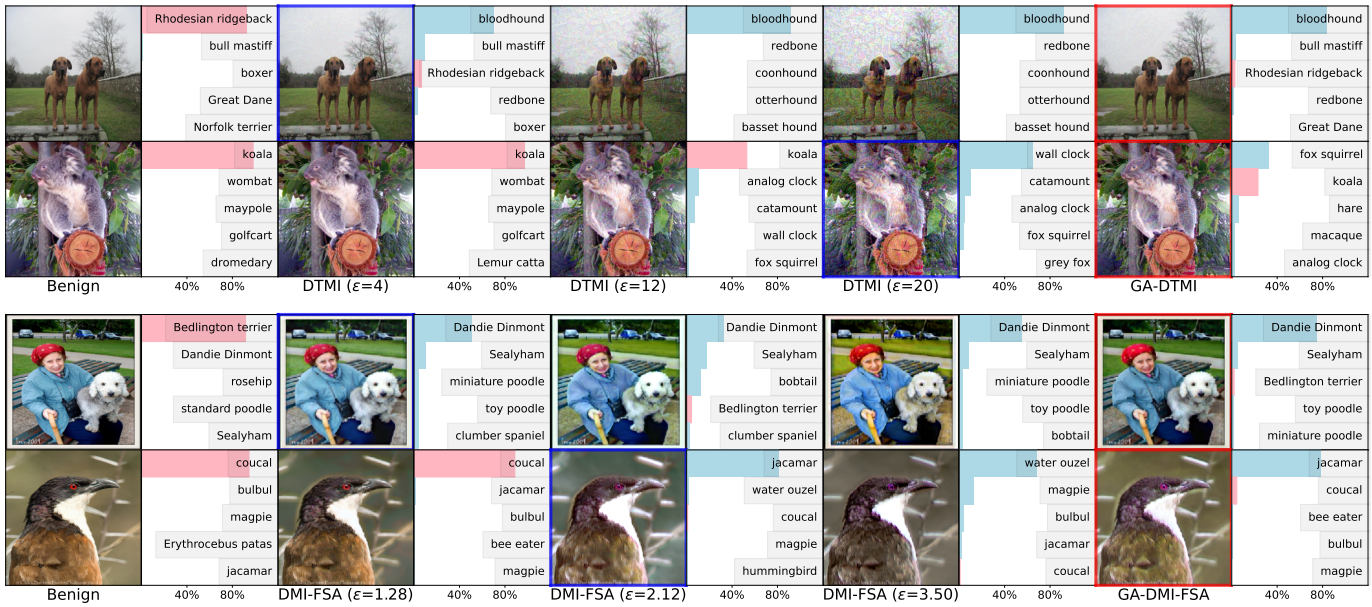

Fig. 1: **Comparison between our method and two baselines under both the ℓ_∞-norm (top) and unrestricted (bottom) threat models using various perturbation radii**. In the even columns, we present the top-5 confidence bars of the target model for the images in the left. The ground-truth label is marked 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. However, our GA framework generates transferable unrestricted adversarial examples (highlighted by red bounding boxes) with lower budgets and smaller changes when compared to the benign images.

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 adversarial competition of a 2021 conference, 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)) by combining the *white-box* feature space attack [24] 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 transfer well to adversarially trained robust models (see Table. III, Figs. 7 and 9).

II. Related Work

A. ℓ_p-norm Adversarial Examples

Existing gradient-based white-box attacks either search for adversarial examples in a fixed ℓ_p-norm ball [25], [26], or optimize the perturbation for each image independently to get a *minimum-norm* solution such as DeepFool [27], CW [28], and fast adaptive boundary attack [29]. However, white-box assumption usually does not hold in real-world scenarios. In query-based black-box setting, attackers utilize output logits [30], [31] or predicted label [32], [33] of the target model to generate adversarial examples. But these attacks typically suffer from high query complexity, making it easy to be detected [34]. (Transfer-based black-box attacks [35]–[38] can pose serious threats in practice as they need no information about the defense models.) Dong *et al.* [16] boosted transferability by integrating momentum into gradient-based methods. Liu *et al.* [14] found that attacking a group of substituted source models simultaneously can improve transferability. Besides, transferability benefits from input transformations such as input diversity [17] and translation-invariant method [18].

B. Unrestricted Adversarial Examples

The ℓ_p-norm distance is not an ideal perceptual similarity metric [39], [40], which oversimplifies the diversity of real-world perturbations. Unrestricted adversarial examples have received significant attention in recent years [41]. Most of the current unrestricted attacks aim to generate imperceptible adversarial examples under *white-box* setting, such as geometric transformations [19], [42], [43] and distance metrics beyond ℓ_p norm [20], [21]. Color-based attacks [44]–[49] were also proposed to generate large but imperceptible perturbations, however, the modified color can sometimes be unnatural. Instead of optimizing in the input space, generative approaches [50]–[53] search for adversarial embeddings in

the latent space. Style transfer [49], [54] is inherently an unrestricted attack as it preserves the semantic of the content image. However, constructing transferable unrestricted adversarial examples is still less explored. In this work, we will fill this gap by combining the white-box feature space attack [24] with transfer-based ℓ_∞-norm attacks to generate semantic-preserving yet transferable unrestricted adversarial examples.

C. Adversarial Defenses

There have been long-standing arms races between defenders and attackers. *Adversarial training* [4] is one of the most promising defense methods. Many variants of adversarial training framework were proposed, e.g., ensemble adversarial training [14] for transfer-based attacks, PGD-based adversarial training [25], and TRADES [55] with a new robust loss based on the trade-off between robustness and accuracy. Geometry-aware instance-reweighted adversarial training [56], which is proven falling into gradient masking [57], shares similar insights with us that the importance of distinct inputs in adversarial training should be different. Laidlaw *et al.* [21] integrate adversarial training with Learned Perceptual Image Patch Similarity (LPIPS) [58], aiming to improve robustness against perturbations that were unseen during training. Unlike empirical defenses, Certified defenses [59]–[62] could provide robustness guarantee under a certain ℓ_p-norm budget.

III. Preliminaries

Notation. A deep neural network classifier can be described as a function $f(\boldsymbol{x};\boldsymbol{\theta}):\mathcal{X}\to\mathbb{R}^{\mathcal{C}}$, 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 $y\in\{1,2,\cdots,\mathcal{C}\}$, the predicted label of $f(\boldsymbol{x};\boldsymbol{\theta})$ is $\hat{f}(\boldsymbol{x}):=\arg\max_j f_j(\boldsymbol{x};\boldsymbol{\theta})$, 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 $\mathbb{B}(\boldsymbol{x},\varepsilon):=\{\boldsymbol{x}'\in\mathcal{X}:\mathcal{D}(\boldsymbol{x},\boldsymbol{x}')\le\varepsilon\}$, 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.

A. Transfer-based ℓ_∞-norm Attacks

Existing transfer-based attacks typically search for adversarial examples in a fixed-radius ℓ_p-norm ball, i.e., $\mathcal{D}(\boldsymbol{x},\boldsymbol{x}')=\|\boldsymbol{x}'-\boldsymbol{x}\|_p\le\varepsilon$. Various methods were proposed to boost transferability of the generated adversarial examples, such as input Diversity Iterative Fast Gradient Sign Method (DI-FGSM) [17], Momentum-based Iterative (MI-FGSM) method [16] and Translation-invariant Iterative (TI-FGSM) method [18]. We formulate a strong ℓ_∞-norm baseline DTMI-FGSM by combining all these techniques, i.e.,

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

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.

B. Transfer-based Unrestricted Attack

Inspired by prior work [63] in style transfer, Xu *et al.* [24] 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) [24] with transfer-based ℓ_∞-norm attacks [16], [17]. 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,

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

where $\phi(\boldsymbol{x})\in\mathbb{R}^{C\times H\times W}$ represents the latent embedding. Xu *et al.* [24] 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] decoder ϕ^{-1}, namely,

$$\begin{aligned}\tilde{\phi}(\boldsymbol{x})&=\left(e^{\boldsymbol{\tau}^\sigma}\cdot\left(\phi(\boldsymbol{x})-\boldsymbol{\mu}\right)+e^{\boldsymbol{\tau}^\mu}\cdot\boldsymbol{\mu},\right)\\ \boldsymbol{x}'&=\phi^{-1}\left(\tilde{\phi}(\boldsymbol{x})\right),\quad\|\boldsymbol{\tau}^\mu\|_\infty\le\ln\varepsilon,\|\boldsymbol{\tau}^\sigma\|_\infty\le\ln\varepsilon,\end{aligned}\tag{3}$$

where $\tilde{\phi}(\boldsymbol{x})$ 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 $\mathcal{D}(\boldsymbol{x},\boldsymbol{x}')=\max\left(e^{\|\boldsymbol{\tau}^\mu\|_\infty},e^{\|\boldsymbol{\tau}^\sigma\|_\infty}\right)\le\varepsilon$. In order to preserve the semantic of the unrestricted adversarial example $\boldsymbol{x}'$, a content loss was added during the attacking process, i.e.,

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

where λ balance the trade-off between adversarial and the content loss. Following Xu *et al.* [24], we set $\lambda=128$ and use the top-5 margin loss for adversarial attack. With all above, the unrestricted attack (see Eq. (3)) can be solved by conventional ℓ_∞-norm attack on parameters $\boldsymbol{\tau}^\mu$ and $\boldsymbol{\tau}^\sigma$. Moreover, the same techniques in Sec. III-A 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.,

$$\begin{aligned}\boldsymbol{m}_{t+1}&=\gamma\cdot\boldsymbol{m}_t+\frac{\nabla_{\boldsymbol{\tau}_t}\mathcal{L}\left(T(\boldsymbol{x}',p),y\right)}{\|\nabla_{\boldsymbol{\tau}_t}\mathcal{L}\left(T(\boldsymbol{x}',p),y\right)\|_1},\\ \boldsymbol{\tau}_{t+1}&=\Pi_{\mathbb{B}(\boldsymbol{x},\varepsilon)}\left(\boldsymbol{\tau}_t-\alpha\cdot\mathtt{sign}(\boldsymbol{m}_{t+1})\right).\end{aligned}\tag{4}$$

When $\gamma=0$, the DMI-FSA attack degenerates to the DI-FSA attack.

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

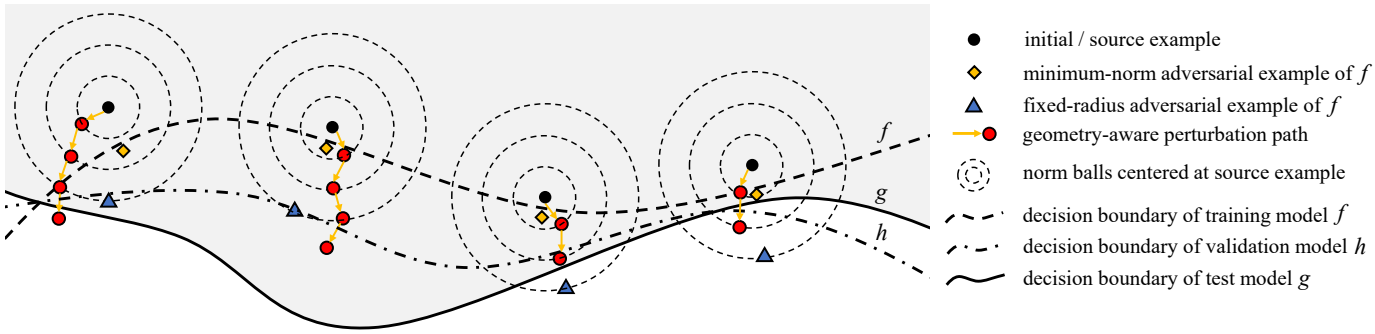

Fig. 2: **Our geometry-aware framework**. Existing fixed-budget methods typically overlook the importance of geometrical distances from inputs to the decision boundary of the test model g. In contrast, our geometry-aware framework aims to find geometry-aware minimal-change perturbation via a validation 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.

C. Evaluation Metric for Transfer-based Attack

The imperceptibility of adversarial examples is hard to evaluate due to the lack of precise quantization of human perception [64]. Sharif *et al.* [65] found that 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 distance metric $\mathcal{D}$ as our main evaluation metrics. Consider a dataset $\hat{S} = \{(\boldsymbol{x_i}, y_i)\}_{i=1}^n$ and the corresponding adversarial examples $\hat{S}_{\text{adv}} = \{(\boldsymbol{x'_i}, y_i)\}_{i=1}^n$ that are crafted on the training model f. Let $n_0 = \sum_{i=1}^n \mathbb{1}\{\hat{g}(\boldsymbol{x}'_i) \neq y_i\}$ be the number of misclassified adversarial examples on the test model g. We define the average total score as:

$$\begin{aligned} S_{total} &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{\hat{g}(\boldsymbol{x}'_i) \neq y_i\} \cdot \mathcal{F}_{\text{reward}}\left(\mathcal{D}(\boldsymbol{x}_i, \boldsymbol{x}'_i)\right) \\ &= \frac{n_0}{n} \cdot \frac{1}{n_0}\sum_{i=1}^{n} \mathbb{1}\{\hat{g}(\boldsymbol{x}'_i) \neq y_i\} \cdot \mathcal{F}_{\text{reward}}\left(\mathcal{D}(\boldsymbol{x}_i, \boldsymbol{x}'_i)\right) \\ &\stackrel{\text{def}}{=} \frac{n_0}{n} \cdot S_{APR}, \end{aligned} \tag{5}$$

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. metric $\mathcal{D}(\boldsymbol{x}, \boldsymbol{x}')$.

IV. Methodology: Geometry-Aware Framework

Eq. (5) 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.,

$$\min_{\boldsymbol{x}'} \ \mathcal{D}(\boldsymbol{x}, \boldsymbol{x}') \qquad \text{s.t.} \ \ \hat{g}(\boldsymbol{x}') \neq y. \tag{6}$$

However, direct optimization of problem (6) 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 [22], [23]. (This problem is also known as model selection problem (we view the source models as training data, then the generated adversarial perturbation is the so-called *selected model* or optimized parameters), and a classic approach to tackle this problem is to split a validation set.) 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 (6), 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 $\mathcal{A}$

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`
  $\text{conf} \leftarrow \frac{\exp(h_y(\boldsymbol{x}_k;\boldsymbol{\theta}))}{\sum_j \exp(h_j(\boldsymbol{x}_k;\boldsymbol{\theta}))}$;
  if $\text{conf} < \eta$ **then**
    Return $\boldsymbol{x}_k$; ▷ `early-stopping in Eq. (7)`
  end if
end for
Return $\boldsymbol{x}_K$;

TABLE I: **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%

such as DMI-FSA (see Eq. (4)) 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 sub-procedure if the probability of the true class on the validation model h is smaller than a threshold η, i.e.,

$$\mathbb{P}\left(\hat{h}(\boldsymbol{x}) = y\right) = \frac{\exp\left(h_y(\boldsymbol{x};\boldsymbol{\theta})\right)}{\sum_j \exp\left(h_j(\boldsymbol{x};\boldsymbol{\theta})\right)} < \eta. \tag{7}$$

Our GA framework is summarized in Algorithm 1 and illustrated in Fig. 2.

V. Experiments

A. Experimental Setup

Datasets & Networks. Similar to Xie *et al.* [17], we randomly select 1,000 images from ILSVRC 2012 validation set [66], 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 [15]. The weights of all these models are publicly available [67]. More details about the networks are summarised in Table. I. 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.* [5]. 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 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,

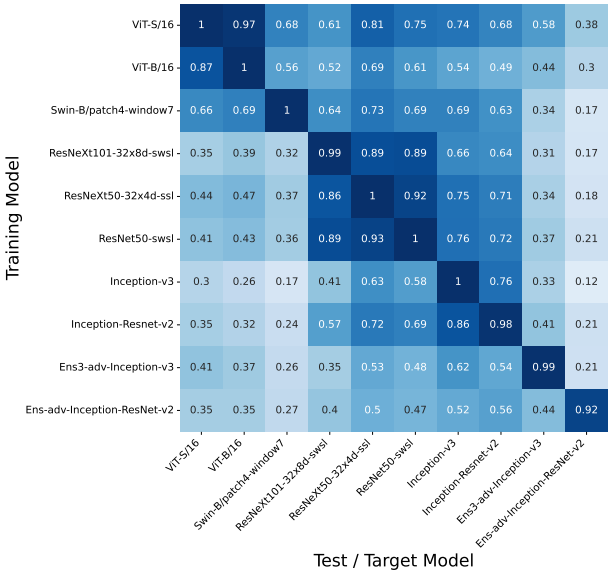

Fig. 3: **Transferability between networks under DTMI-FGSM attack**. The rows stand for source models and the columns stand for target models. Adversarial examples transfer well between models with similar architectures.

we set the decay factor $\mu = 1$ as in Dong *et al.* [16]. For DI-FGSM [17], 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 [18], we use Gaussian kernel with kernel size 5×5.

B. Balancing Transfer Success Rate and 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 Table. I, respectively. The test model is Inception-ResNet-v2. The optimal threshold η (see Eq (7)) is searched from a finite set ranging from 0.001 to 0.9 by querying[2] the

[2]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 prior information on η which depends on the similarity between f, h and g, making the query complexity rather limited.

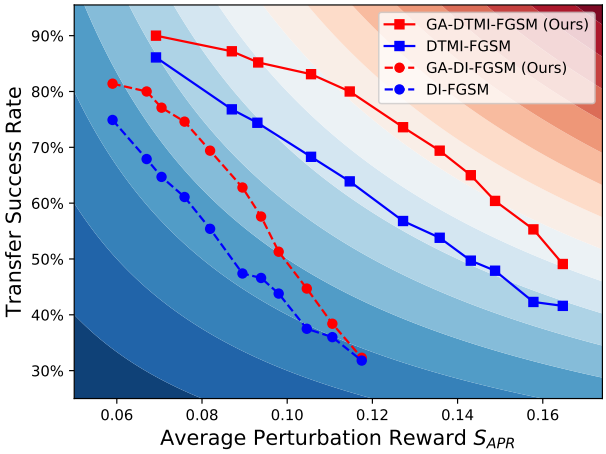

Fig. 4: **Contour of average total score** S_{total}. Fixing transfer success rate as 80%, our approach GA-DTMI-FGSM surpasses the baseline DTMI-FGSM (see Eq. (1)) by up to 43.35% in terms of average perturbation reward.

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, DTMI-FGSM, and Admix-DTI-FGSM [37] (limited by the memory of a single NVIDIA RTX 3090, we set the number of admixed images $m_1 = 3$ and the number of randomly sampled images from other categories $m_2 = 2$), named GA-DI-FGSM, GA-DTI-FGSM, GA-DTMI-FGSM, and GA-Admix-DTI-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.

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 Table. II, 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.

TABLE II: **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 [17]	61.1%	0.0759	4.64%
GA-DI-FGSM	**69.4%**	**0.0819**	**5.68%**
DTI-FGSM [18]	57.3%	0.1101	5.68%
GA-DTI-FGSM	**67.9%**	**0.1176**	**7.98%**
DTMI-FGSM	63.9%	0.1147	7.33%
GA-DTMI-FGSM	**69.4%**	**0.1358**	**9.42%**
(Admix-DTI-FGSM [37])	68.1%	0.1248	8.50%
(GA-Admix-DTI-FGSM)	**82.5%**	**0.1299**	**10.72%**
DI-FSA	48.3%	0.5328	25.73%
GA-DI-FSA	**50.4%**	**0.5541**	**27.93%**
DMI-FSA	51.3%	**0.5616**	28.81%
GA-DMI-FSA	**58.3%**	0.5355	**31.32%**

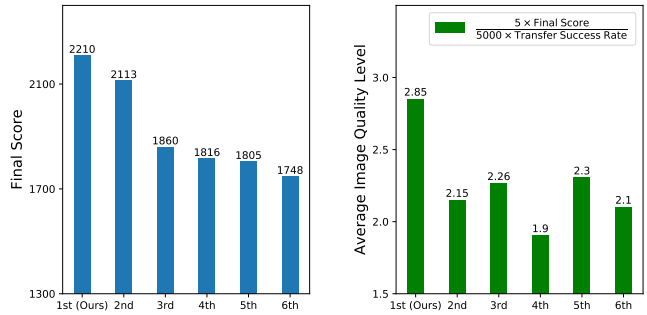

Fig. 5: **Top-6 results** in the adversarial competition of a 2021 conference. The final scores were manually scored by multiple human referees.

C. Case Study: Adversarial Competition of a 2021 Conference

In the adversarial competition of a 2021 conference, contestants were asked to submit adversarial examples without any access to the defense models. The dataset is a subset of ILSVRC 2012 validation set [66], 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 (judged by human referees), 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 $\sum_i \mathbb{1}\{\hat{g}(\boldsymbol{x}'_i) \neq y_i\} \times S_s(\boldsymbol{x}'_i) \times \frac{S_q(\|\boldsymbol{x}'_i - \boldsymbol{x}_i\|)}{5}$.

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.

D. Transferable Unrestricted Adversarial Examples

Most of current defenses can be easily broken by unseen attacks in a white-box manner. Adversarial training against multiple ℓ_p-norm attacks [75] solved this issue partially, however, at the cost of robustness against single ℓ_p-norm attack. Laidlaw *et al.* [21] integrated adversarial training with Learned Perceptual Image Patch Similarity (LPIPS) [58], aiming to improve robustness against perturbations that were unseen during training. However, the proposed attack [21], similar to other unrestricted attacks [43], [45], suffers from weaker transferability to the target model. In practice, attackers

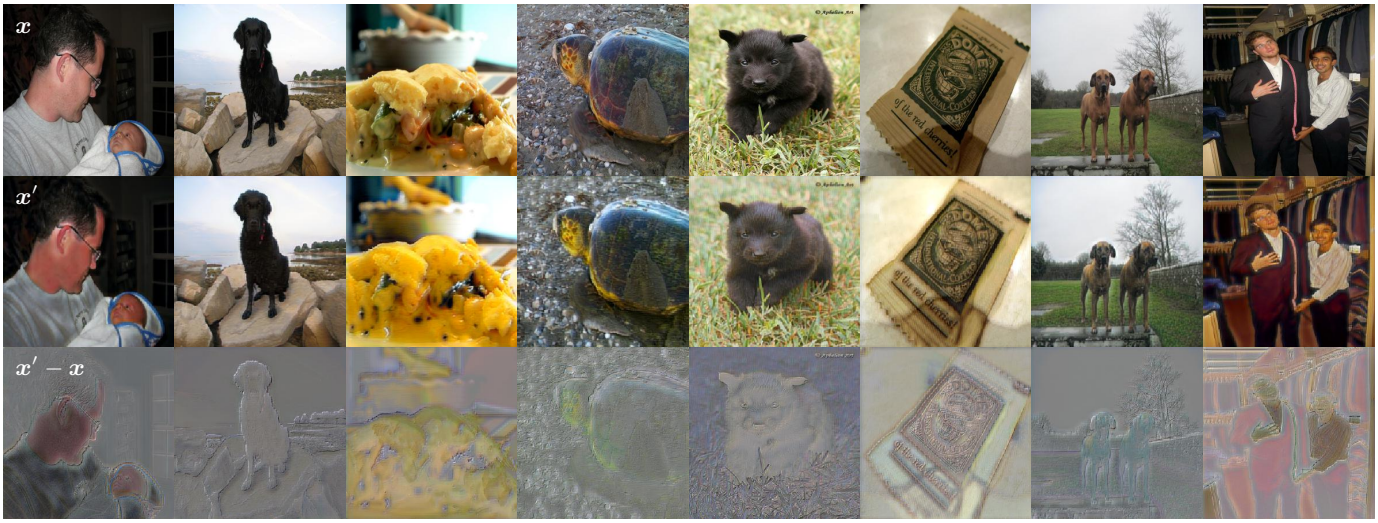

Fig. 6: **Adversarial examples of GA-DMI-FSA**, which are misclassified by *all* the models in Table. III. Top: benign examples $\boldsymbol{x}$. Middle: unrestricted adversarial examples $\boldsymbol{x}'$. Bottom: normalized adversarial perturbations $\boldsymbol{x}' - \boldsymbol{x}$.

TABLE III: **Benchmarking classification on Imagenet under transfer-based unrestricted attacks**. PGD^*_{40} indicates the PGD attack with 40 steps ($\varepsilon = \frac{4}{255}$). (We denote the adversarial examples crafted by GA-DTMI-FGSM, GA-Admix-DTI-FGSM, and GA-DMI-FSA in Table. II as $\text{GA}_{\text{DTMI-FGSM}}$, $\text{GA}_{\text{Admix-DTI}}$, and $\text{GA}_{\text{DMI-FSA}}$, respectively. **Bold** and underline indicate the lowest and second lowest in each row.)

Defenses	Clean	PGD^*_{40}	ReColor	FSA	$\text{GA}_{\text{DTMI-FGSM}}$	($\text{GA}_{\text{Admix-DTI}}$)	$\text{GA}_{\text{DMI-FSA}}$
Inception-ResNet-v2$_{\text{Ens-adv}}$ [15]	99.9%	**10.0%**	93.2%	90.0%	87.1%	78.1%	43.5%
Fast$_{\text{AT}}$ [68]	66.9%	35.7%	62.8%	59.3%	64.7%	64.1%	**28.5%**
Free$_{\text{AT}}$ [69]	77.3%	40.6%	71.6%	68.4%	74.0%	73.0%	**35.1%**
Resnet152-Base [70]	67.6%	39.0%	64.1%	61.2%	65.1%	65.7%	**37.4%**
Resnext101-DenoiseAll [70]	80.3%	52.2%	77.0%	73.8%	78.7%	78.1%	**47.8%**
Resnet152-Denoise [70]	72.2%	41.8%	68.2%	65.2%	70.7%	69.9%	**40.3%**
RVT-Tiny [71]	96.7%	**0.0%**	78.9%	81.3%	44.5%	26.3%	33.6%
DeepAugment+AugMix [72]	96.1%	**0.0%**	82.8%	89.3%	58.9%	41.4%	63.2%
Efficientnet-l2-ns [73]	99.5%	**0.0%**	95.7%	97.0%	81.3%	78.6%	86.0%
Swin-L/patch4-window-12 [74]	99.1%	**0.0%**	88.4%	90.7%	66.6%	58.7%	61.8%

typically have no information about the defense models 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*.

Implementation Details. For the ReColor attack [45], we set $\varepsilon = 1.0$ and iterations $T = 100$ which achieves 89.7% attack success rate on training model f (the same as Sec. V-B) and 9.1% transfer success rate on the test model Inception-ResNet-v2. For FSA attack [24], we set $\varepsilon = 3.0$ and $T = 500$ which achieves 54.2% attack success rate and 12.4% transfer success rate on the same training and test models (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. (4) boost both the attacking ability and transferability.). Besides six adversarially trained and two high-performance classification models, we select two state-of-the-art models on the ImageNet-R dataset [72].

Experimental Results. From Table. III, we can conclude the following observations: (a) $\text{GA}_{\text{Admix-DTI}}$ transfers better than $\text{GA}_{\text{DTMI-FGSM}}$.) b) ℓ_∞-norm transfer attack 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). c) DeepAugment [72], which utilizes semantic-preserving augmentations during training, exhibits non-trivial robustness against GA-DMI-FSA attack. d) Efficientnet-l2-ns [73] 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 [77], [78]. We visualize part of the transfer attack results on adversarially trained model Resnext101-DenoiseAll [70] in Figs. 7, where our method GA-DMI-FSA is able to generate semantic-preserving yet transferable unrestricted adversarial examples. (For more visualization results, please see Fig. 9

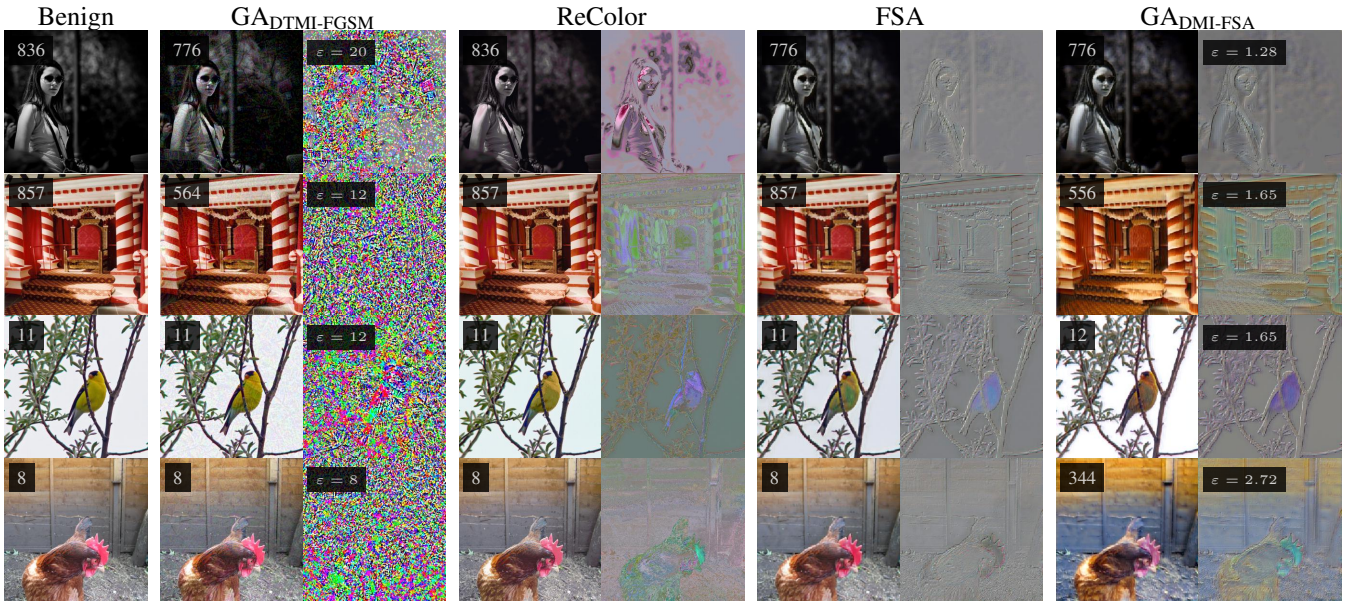

Fig. 7: **Visualization of transfer attack results on Resnext101-DenoiseAll** [70]. For each image, we print its predicted label on model Resnext101-DenoiseAll in the upper left corner. For each transfer-based adversarial example, we present the perturbation on its right. For each perturbation crafted 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 [76].

in Sec. VII-A.)

E. (Ablation Studies and Discussions)

The optimal η depends on the combination of f, h, g. As illustrated in Fig. 2, the effectiveness of our geometry-aware framework relies on the relationship between the decision boundaries of f, h, and 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: a) 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. b) Ensembling more models yields better performance. When adding models 8 and 9 into $f = (2, 3, 5)$ and $g = (1, 4, 6)$ respectively, there is a remarkable improvement in the average total score.

(**Train-valid splitting via diversity control.** Fig.8 shows that there is a large variance of the performance depending on the train-valid split, which motivates us to develop a partition strategy that performs better than random splitting. Actually, we can traverse all possible partitions to select the best split which achieves the highest average total score S_{total}. However, this approach is time-consuming and is limited by the maximum number of queries to the test model. Here we propose a query-free approach that only utilizes the information of transferability (see Fig.3) between the pre-trained models. Let $\phi_1, \phi_2, \cdots, \phi_n$ represent the pre-trained models and w_{ij}

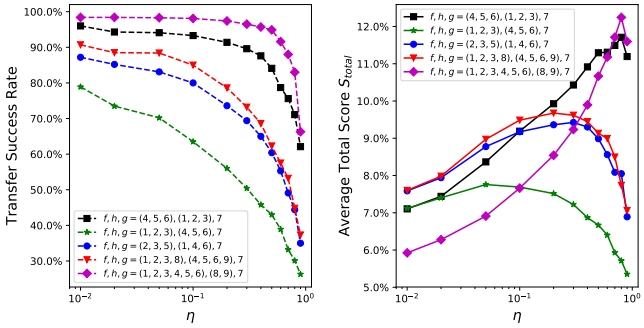

Fig. 8: **Comparison between various kinds of partition for training, validation, and test models**. The index in the legend corresponds to the model index in Table. I. As the threshold η for early-stopping increases, the generated adversarial examples have higher confidence (probability of the true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

($w_{ij} \neq w_{ji}$) be the transfer success rate from the source model ϕ_i to the target model ϕ_j under a fixed-radius attack (e.g., DTMI-FGSM). Denote the binary partition function as $\mathcal{G}(\phi)$, the training set and the validation set can be formulated as $\mathcal{T} = \{i \mid \mathcal{G}(\phi_i) = 0\}$ and $\mathcal{V} = \{j \mid \mathcal{G}(\phi_j) = 1\}$, respectively. We partition the n pre-trained models into k training models and $n-k$ validation models (i.e., $|\mathcal{T}| = k, |\mathcal{V}| = n-k$). Given

TABLE IV: (Comparison between the optimal partition derived from diversity control and random splitting. The test model g for the top group and the bottom group is model 7 in Table I and Efficientnet-l2-ns [73], respectively. In both groups, the proposed diversity score $\ell_{\mathcal{G}}$ negatively correlates with the average total score S_{total}.)

f	h	$\ell_{\mathcal{G}}$	TSR (↑)	S_{APR} (↑)	S_{total} (↑)
$(1,2,3)$	$(4,5,6)$	2.877	67%	0.1239	8.3%
$(1,2,6)$	$(3,4,5)$	2.756	75.4%	0.1336	10.07%
$(2,4,5)$	$(1,3,6)$	2.727	74.1%	0.1324	9.81%
$(4,5,6)$	$(1,2,3)$	2.577 ($\mathcal{G}^*$)	79.2%	0.1457	11.54%
$(0,3,5,6,9)$	$(1,2,4,7,8)$	2.596	25.5%	0.114	2.91%
$(2,3,4,5,7)$	$(0,1,6,8,9)$	2.437	45.3%	0.087	3.94%
$(3,4,5,7,8)$	$(0,1,2,6,9)$	2.318	23.8%	0.1353	3.22%
$(3,4,5,6,7)$	$(0,1,2,8,9)$	1.991 ($\mathcal{G}^*$)	27.5%	0.118	3.25%

any partition function $\mathcal{G}$, we define the **diversity score** $\ell_{\mathcal{G}}$ as:

$$\begin{aligned} \ell_{\mathcal{G}} &= \frac{1}{k}\sum_{i\in\mathcal{T}} \ell_{\mathcal{G}}(\phi_i) + \frac{1}{n-k}\sum_{j\in\mathcal{V}} \ell_{\mathcal{G}}(\phi_j), \\ \ell_{\mathcal{G}}(\phi_i) &= \frac{1}{k-1}\sum_{t\neq i, t\in\mathcal{T}} w_{it} + \max_{t\in\mathcal{V}} w_{it}, \\ \ell_{\mathcal{G}}(\phi_j) &= \frac{1}{n-k-1}\sum_{t\neq j, t\in\mathcal{V}} w_{jt} + \max_{t\in\mathcal{T}} w_{tj}. \end{aligned} \tag{8}$$

For both the training set loss $\ell_{\mathcal{G}}(\phi_i)$ and the validation set loss $\ell_{\mathcal{G}}(\phi_j)$: 1) The first term encourages inner-group diversity. 2) The second term tries to minimize the maximum transfer success rate from the training model to the validation model (the training model f and the validation model h cannot be too similar). Finally, we initiate the train-valid split with the optimal partition $\mathcal{G}^* = \arg\min_{\mathcal{G}} \ell_{\mathcal{G}}$, which does not require any query to the test model. $\mathcal{G}^*$ can be efficiently searched by traversing all the C_n^k kinds of combinations.)

(Most importantly, we find that the diversity score $\ell_{\mathcal{G}}$ correlates well with the final performance. As shown in Table IV, the splits that have the lowest diversity score $\ell_{\mathcal{G}}$ outperform many partitions with higher $\ell_{\mathcal{G}}$.)

VI. 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 could improve 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 unrestricted adversarial examples.

References

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778. 1

[2] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2021. 1

[3] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *ICLR*, 2014. 1

[4] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *ICLR*, 2015. 1, 3

[5] R. Shao, Z. Shi, J. Yi, P.-Y. Chen, and C.-J. Hsieh, "On the adversarial robustness of visual transformers," *arXiv preprint arXiv:2103.15670*, 2021. 1, 5

[6] S. Bhojanapalli, A. Chakrabarti, D. Glasner, D. Li, T. Unterthiner, and A. Veit, "Understanding robustness of transformers for image classification," in *ICCV*, October 2021, pp. 10 231–10 241. 1

[7] Y. Bai, J. Mei, A. Yuille, and C. Xie, "Are transformers more robust than cnns?" in *NeurIPS*, 2021. 1

[8] M. Bojarski, D. Del Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang *et al.*, "End to end learning for self-driving cars," *arXiv preprint arXiv:1604.07316*, 2016. 1

[9] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *BMVC*. BMVA Press, September 2015, pp. 41.1–41.12. 1

[10] A. Athalye, N. Carlini, and D. Wagner, "Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples," in *ICML*, vol. 1, 2018, pp. 436–448. 1

[11] F. Tramer, N. Carlini, W. Brendel, and A. Madry, "On adaptive attacks to adversarial example defenses," in *NeurIPS*, vol. 33. Curran Associates, Inc., 2020, pp. 1633–1645. 1

[12] N. Papernot, P. McDaniel, and I. Goodfellow, "Transferability in machine learning: from phenomena to black-box attacks using adversarial samples," *arXiv preprint arXiv:1605.07277*, 2016. 1

[13] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against machine learning," in *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, 2017, pp. 506–519. 1

[14] Y. Liu, X. Chen, C. Liu, and D. Song, "Delving into transferable adversarial examples and black-box attacks," in *ICLR*, 2017. 1, 2, 3

[15] F. Tramer, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel, "Ensemble adversarial training: Attacks and defenses," in *ICLR*, 2018. 1, 5, 7

[16] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li, "Boosting adversarial attacks with momentum," in *CVPR*, 2018, pp. 9185–9193. 1, 2, 3, 5

[17] C. Xie, Z. Zhang, Y. Zhou, S. Bai, J. Wang, Z. Ren, and A. Yuille, "Improving transferability of adversarial examples with input diversity," in *CVPR*, 2019. 1, 2, 3, 5, 6

[18] Y. Dong, T. Pang, H. Su, and J. Zhu, "Evading defenses to transferable adversarial examples by translation-invariant attacks," in *CVPR*, 2019, pp. 4312–4321. 1, 2, 3, 5, 6

[19] C. Xiao, J.-Y. Zhu, B. Li, W. He, M. Liu, and D. Song, "Spatially transformed adversarial examples," in *ICLR*, 2018. [Online]. Available: https://openreview.net/forum?id=HyydRMZC- 1, 2

[20] E. Wong, F. Schmidt, and Z. Kolter, "Wasserstein adversarial examples via projected sinkhorn iterations," in *ICML*. PMLR, 2019, pp. 6808–6817. 1, 2

[21] C. Laidlaw, S. Singla, and S. Feizi, "Perceptual adversarial robustness: Defense against unseen threat models," in *ICLR*, 2021. 1, 2, 3, 6

[22] S. Cheng, Y. Dong, T. Pang, H. Su, and J. Zhu, "Improving black-box adversarial attacks with a transfer-based prior," in *NeurIPS*, vol. 32, 2019. 1, 4

[23] Z. Katzir and Y. Elovici, "Who's afraid of adversarial transferability?" *arXiv preprint arXiv:2105.00433*, 2021. 1, 4

[24] Q. Xu, G. Tao, S. Cheng, and X. Zhang, "Towards feature space adversarial attack by style perturbation," *AAAI*, vol. 35, no. 12, pp. 10 523–10 531, May 2021. 2, 3, 7

[25] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *ICLR*, 2018. 2, 3

[26] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," *ICLR 2017 - Workshop Track Proceedings*, no. c, pp. 1–14, 2019. 2

[27] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "Deepfool: a simple and accurate method to fool deep neural networks," in *CVPR*, 2016, pp. 2574–2582. 2

[28] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *2017 ieee symposium on security and privacy (sp)*. IEEE, 2017, pp. 39–57. 2

[29] F. Croce and M. Hein, "Minimally distorted adversarial examples with a fast adaptive boundary attack," in *ICML*. PMLR, 2020, pp. 2196–2205. 2

[30] P.-Y. Chen, H. Zhang, Y. Sharma, J. Yi, and C.-J. Hsieh, "Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models," in *Proceedings of the 10th ACM workshop on artificial intelligence and security*, 2017, pp. 15–26. 2

[31] M. Andriushchenko, F. Croce, N. Flammarion, and M. Hein, "Square attack: A query-efficient black-box adversarial attack via random search," in *ECCV*, 2020, pp. 484–501. [Online]. Available: https://doi.org/10.1007/978-3-030-58592-1_29 2

[32] W. Brendel, J. Rauber, and M. Bethge, "Decision-based adversarial attacks: Reliable attacks against black-box machine learning models," in *ICLR*, 2018. 2

[33] M. Cheng, T. Le, P.-Y. Chen, H. Zhang, J. Yi, and C.-J. Hsieh, "Query-efficient hard-label black-box attack: An optimization-based approach," in *ICLR*, 2019. 2

[34] D. Willmott, A. K. Sahu, F. Sheikholeslami, F. Condessa, and Z. Kolter, "You only query once: Effective black box adversarial attacks with minimal repeated queries," *arXiv preprint arXiv:2102.00029*, 2021. 2

[35] J. Lin, C. Song, K. He, L. Wang, and J. E. Hopcroft, "Nesterov accelerated gradient and scale invariance for adversarial attacks," in *International Conference on Learning Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=SJlHwkBYDH 2

[36] D. Wu, Y. Wang, S.-T. Xia, J. Bailey, and X. Ma, "Skip connections matter: On the transferability of adversarial examples generated with resnets," in *International Conference on Learning Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=BJlRs34Fvr 2

[37] X. Wang, X. He, J. Wang, and K. He, "Admix: Enhancing the transferability of adversarial attacks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 16 158–16 167. 2, 6

[38] X. Wang, J. Ren, S. Lin, X. Zhu, Y. Wang, and Q. Zhang, "A unified approach to interpreting and boosting adversarial transferability," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=X76iqnUbBjz 2

[39] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *ECCV*. Springer, 2016, pp. 694–711. 2

[40] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *CVPR*, 2017, pp. 1125–1134. 2

[41] T. B. Brown, N. Carlini, C. Zhang, C. Olsson, P. Christiano, and I. Goodfellow, "Unrestricted adversarial examples," *arXiv preprint arXiv:1809.08352*, 2018. 2

[42] R. Alaifari, G. S. Alberti, and T. Gauksson, "ADef: an iterative algorithm to construct adversarial deformations," in *ICLR*, 2019. [Online]. Available: https://openreview.net/forum?id=Hk4dFjR5K7 2

[43] L. Engstrom, B. Tran, D. Tsipras, L. Schmidt, and A. Madry, "Exploring the landscape of spatial robustness," in *ICML*. PMLR, 2019, pp. 1802–1811. 2, 6

[44] H. Hosseini and R. Poovendran, "Semantic adversarial examples," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 1614–1619. 2

[45] C. Laidlaw and S. Feizi, "Functional adversarial attacks," in *NeurIPS*, 2019. 2, 6, 7

[46] Z. Zhao, Z. Liu, and M. Larson, "Towards large yet imperceptible adversarial image perturbations with perceptual color distance," in *CVPR*, 2020, pp. 1039–1048. 2

[47] Z. Zhao, Z. Liu, and M. A. Larson, "Adversarial color enhancement: Generating unrestricted adversarial images by optimizing a color filter," in *BMVC*, 2020. [Online]. Available: https://www.bmvc2020-conference.com/assets/papers/0099.pdf 2

[48] A. S. Shamsabadi, R. Sanchez-Matilla, and A. Cavallaro, "Colorfool: Semantic adversarial colorization," in *CVPR*, 2020, pp. 1151–1160. 2

[49] A. Bhattad, M. J. Chong, K. Liang, B. Li, and D. A. Forsyth, "Unrestricted adversarial examples via semantic manipulation," in *ICLR*, 2020. 2, 3

[50] Y. Song, R. Shu, N. Kushman, and S. Ermon, "Constructing unrestricted adversarial examples with generative models," in *NeurIPS*, vol. 31. Curran Associates, Inc., 2018. 2

[51] S. Gowal, C. Qin, P.-S. Huang, T. Cemgil, K. Dvijotham, T. Mann, and P. Kohli, "Achieving robustness in the wild via adversarial mixing with disentangled representations," in *CVPR*, 2020, pp. 1211–1220. 2

[52] H. Qiu, C. Xiao, L. Yang, X. Yan, H. Lee, and B. Li, "Semanticadv: Generating adversarial examples via attribute-conditioned image editing," in *ECCV*. Springer, 2020, pp. 19–37. 2

[53] E. Wong and J. Z. Kolter, "Learning perturbation sets for robust machine learning," in *ICLR*, 2021. 2

[54] V. U. Prabhu, J. Whaley, and S. Francisco, "Art-attack! on style transfers with textures, label categories and adversarial examples," 2018. 3

[55] H. Zhang, Y. Yu, J. Jiao, E. P. Xing, L. E. Ghaoui, and M. I. Jordan, "Theoretically principled trade-off between robustness and accuracy," in *ICML*, 2019. 3

[56] J. Zhang, J. Zhu, G. Niu, B. Han, M. Sugiyama, and M. Kankanhalli, "Geometry-aware instance-reweighted adversarial training," in *ICLR*, 2021. 3

[57] D. Hitaj, G. Pagnotta, I. Masi, and L. V. Mancini, "Evaluating the robustness of geometry-aware instance-reweighted adversarial training," *arXiv preprint arXiv:2103.01914*, 2021. 3

[58] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *CVPR*, 2018, pp. 586–595. 3, 6

[59] J. Cohen, E. Rosenfeld, and Z. Kolter, "Certified adversarial robustness via randomized smoothing," in *ICML*. PMLR, 2019, pp. 1310–1320. 3

[60] H. Salman, G. Yang, J. Li, P. Zhang, H. Zhang, I. Razenshteyn, and S. Bubeck, "Provably robust deep learning via adversarially trained smoothed classifiers," in *NeurIPS*. Red Hook, NY, USA: Curran Associates Inc., 2019. 3

[61] H. Zhang, H. Chen, C. Xiao, S. Gowal, R. Stanforth, B. Li, D. Boning, and C.-J. Hsieh, "Towards stable and efficient training of verifiably robust neural networks," in *ICLR*, 2020. [Online]. Available: https://openreview.net/forum?id=Skxuk1rFwB 3

[62] K. Leino, Z. Wang, and M. Fredrikson, "Globally-robust neural networks," in *ICML*, 2021. 3

[63] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *ICCV*, 2017, pp. 1501–1510. 3

[64] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004. 4

[65] M. Sharif, L. Bauer, and M. K. Reiter, "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*, 2018, pp. 1605–1613. 4

[66] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*. IEEE, 2009, pp. 248–255. 5, 6

[67] R. Wightman, "Pytorch image models," https://github.com/rwightman/pytorch-image-models, 2019. 5

[68] E. Wong, L. Rice, and J. Z. Kolter, "Fast is better than free: Revisiting adversarial training," in *ICLR*, 2020. [Online]. Available: https://openreview.net/forum?id=BJx040EFvH 7

[69] A. Shafahi, M. Najibi, M. A. Ghiasi, Z. Xu, J. Dickerson, C. Studer, L. S. Davis, G. Taylor, and T. Goldstein, "Adversarial training for free!" in *NeurIPS*, vol. 32. Curran Associates, Inc., 2019. [Online]. Available: https://proceedings.neurips.cc/paper/2019/file/7503cfacd12053d309b6bed5c89de212-Paper.pdf 7

[70] C. Xie, Y. Wu, L. v. d. Maaten, A. L. Yuille, and K. He, "Feature denoising for improving adversarial robustness," in *CVPR*, 2019, pp. 501–509. 7, 8, 12

[71] X. Mao, G. Qi, Y. Chen, X. Li, S. Ye, Y. He, and H. Xue, "Rethinking the design principles of robust vision transformer," *arXiv preprint arXiv:2105.07926*, 2021. 7

[72] D. Hendrycks, S. Basart, N. Mu, S. Kadavath, F. Wang, E. Dorundo, R. Desai, T. Zhu, S. Parajuli, M. Guo, D. Song, J. Steinhardt, and J. Gilmer, "The many faces of robustness: A critical analysis of out-of-distribution generalization," *ICCV*, 2021. 7

[73] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le, "Self-training with noisy student improves imagenet classification," in *CVPR*, 2020, pp. 10 687–10 698. 7, 9

[74] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," *arXiv preprint arXiv:2103.14030*, 2021. 7

[75] F. Tramer and D. Boneh, "Adversarial training and robustness for multiple perturbations," in *NeurIPS*, vol. 32. Curran Associates, Inc., 2019. 6

[76] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel, "Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness," *arXiv preprint arXiv:1811.12231*, 2018. 8

[77] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar, "Do imagenet classifiers generalize to imagenet?" in *ICML*. PMLR, 2019, pp. 5389–5400. 7

[78] R. Taori, A. Dave, V. Shankar, N. Carlini, B. Recht, and L. Schmidt, "Measuring robustness to natural distribution shifts in image classification," in *NeurIPS*, 2020. 7

[79] S. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "Deepfool: A simple and accurate method to fool deep neural networks," in *CVPR*. IEEE Computer Society, 2016, pp. 2574–2582. [Online]. Available: https://doi.org/10.1109/CVPR.2016.282 13

VII. Appendix

A. More Visualization Results

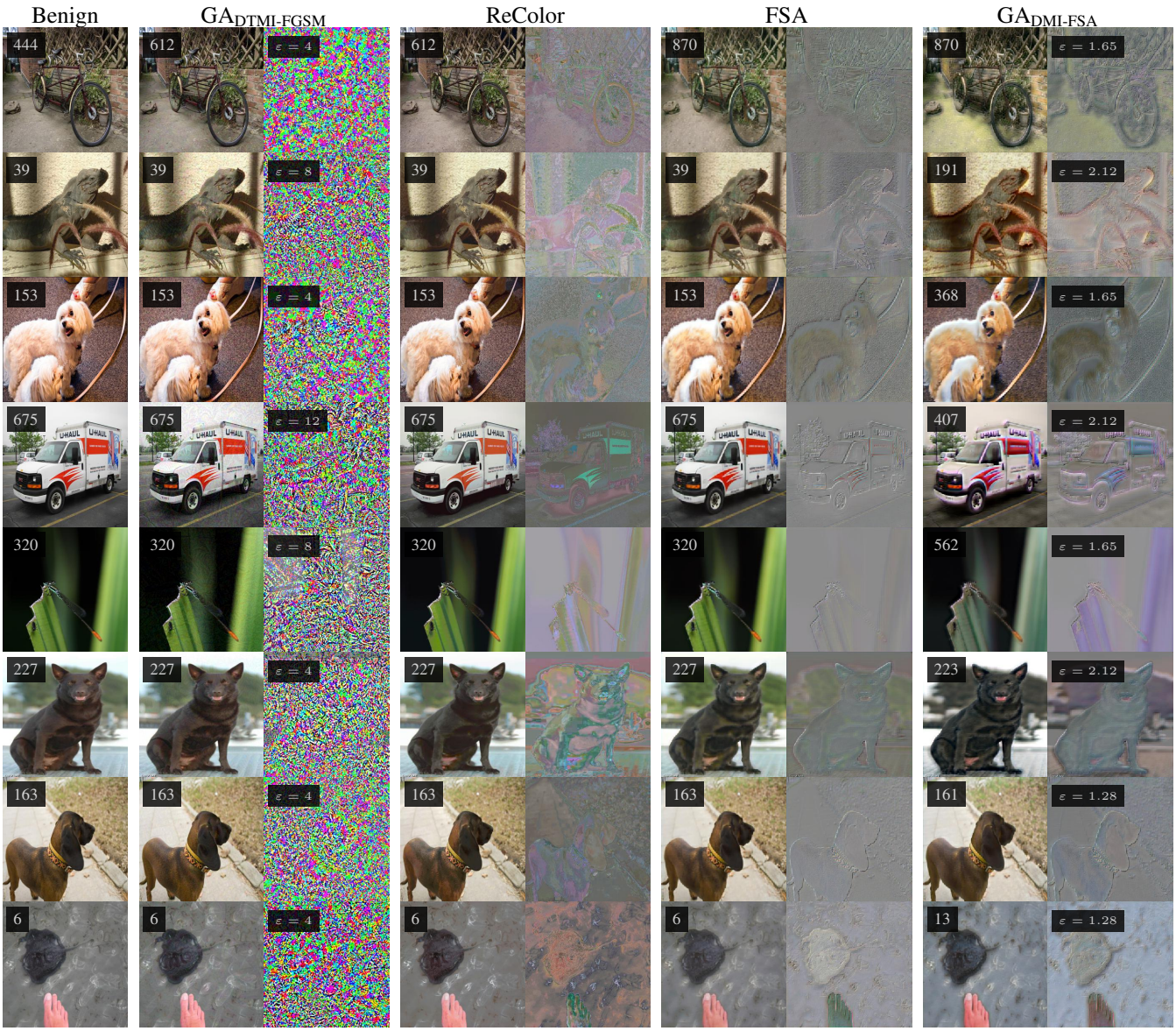

Fig. 9: **Visualization of transfer attack on Resnext101-DenoiseAll** [70].

B. Summary of Changes from Prior Submission

We thank all the reviewers for their detailed and constructive comments, which greatly helped us improve the manuscript. We have comprehensively addressed these comments in the revision, and we believe the manuscript has been notably improved. We feel encouraged that our geometry-aware (GA) framework is acknowledged by the reviewers:

- *"**The idea behind GA sounds interesting.**"*
- *"**The paper is well-written. The ideas are new, well-motivated and well performing in practice.**"*

Below are our responses to each reviewer's questions:

1) Reviewer #1:

- *"There is no insight and/or analysis into why and how the current realization of GA framework is (always) capable of finding the minimum change adversarial example with a fixed set of K sub-procedures. Apart from transferability performance to different target models (like Fig. 4) and some qualitative figures (1 and 6), there is no analysis in this regard."* [Re: Finding the optimal minimum-norm adversarial perturbation is intractable for deep networks even in white-box attacks. For example, DeepFool [79] demonstrates its effectiveness by showing that their perturbations have smaller ℓ_∞-norms compared to FGSM when achieving a comparable attack success rate (similar to our Fig. 4). Our framework GA *approximately* solves Eq. (6) by discretizing the continuous perturbation radius into K candidates and choosing the minimum candidate that suffices to fool the validation model with a certain margin. Following we provide two insights why GA could achieve a good solution to Eq. (6): 1) K is capable of controlling the trade-off between S_{APR} and efficiency. 2) The *early-stopping* mechanism on *validation* model h could make it possible to generate smaller perturbations without a sharp decrease in transfer success rate (TSR). To verify the above two insights, we add three groups of experiments in Table. **??**.]
- *" In the proposed framework how to partition the available white-box source models into training and validation sets is quite unclear. There are no clear guidelines. Furthermore, the ablation studies in Fig. 8 doesn't explain why certain combination splits work better than others."* [Re: we provide two guidelines in Sec. V-E and add detailed description about Fig. 8.]
- *"The GA framework seems quite sensitive to the eta parameter and is also dependent on the specific partition of source models into training and validations sets (Fig. 8)."* [Re: It's not GA sensitive to η, but the optimal η^* is dependent on the relationship between f, h and g, which might be inevitable in transfer-based black-box attacks. In practice, η is firstly sampled from a prior distribution and then finetuned by the feedback of TSR. Fortunately, from Fig. 8, we find it effective to set $\eta \in (0.02, 0.2)$.]

2) Reviewer #2:

- *"Parts of the notation could be clearer to improve readability of the presented ideas. Examples for this is the use of lower-case letters for both single models and sets of models. In addition, while the implementation choices are well documented, making the code base publicly available would allow for much easier reproduction of the results."*
- *"**The ideas presented in the paper are new, well-motivated and well performing in practice, and they can be easily used with other attacks.** The authors show extensive experimental evidence regarding the performance of their method. Parts where the paper could benefit is more care to finest details in the notation and making the codebase available for easier reproduction."*
- *"After reading the rebuttal and other reviewers comments, I agree that more experiments and fairer comparisons to baselines are needed. Resubmitting it to the next venue would be beneficial for the paper's contributions and strengths."* [Re: For the potential fairness concern, we add detailed explanations in Sec. **??**.]

3) Reviewer #3:

- *"Lack of SOTA methods of transfer-based attack in the related work, e.g., SIM, SGM, Admix, IR."* [Re: As also mentioned in the strengths@Reviewer #2, GA allows for easy use of incorporation of arbitrary fixed-budget transfer-based attacks.]
- *"Is the C in L180 (the number of classes) and the C in L230 (the number of channels in the feature) the same meaning? If not, please revise the notation with another alphabet."* [Re: They do not have the same meaning, we have changed the notation of the number of classes to $\mathcal{C}$.]
- *"The concern about the fairness in the experiments. (1) The proposed GA methods utilize both training models f and validation model h when generating the adversarial examples. In the baseline method of comparison, whether the ensemble model of both f and h is used? (2) In L414 the author says that they choose the threshold eta by querying the test model g. I suppose that test model should be a black-box model and totally can't be obtained by attackers. I think this behavior of tuning hyperparameters on the test set is unacceptable."* [Re: (1) The ensemble model in baseline uses f only but we add detailed explanations in Sec. **??**. (2) We provide three points of view that querying the test model g for TSR is acceptable in our designed experiments in Sec. V-B: 1) We query TSR on the test model to select the optimal η that exhibits the best performance. For a fair comparison, the baseline method also queries the test model multiple times at the corresponding x-axis to find the best fixed x that performs well. 2) We only query the TSR as feedback[3], i.e., the logits and the predict label of single query are still not accessible. However, conventional query-based attacks may require thousands of queries on logits or predict label for each data. 3) We might DO NOT need (or only several times) this query procedure in practice by following the guidelines proposed in Sec. V-E.]

[3]In practice, we typically know if the crafted adversarial example works.