Towards Transferable Unrestricted Adversarial Examples with Minimum Changes

Fangcheng Liu  
Peking University  
equation@stu.pku.edu.cn

Chao Zhang  
Peking University  
c.zhang@pku.edu.cn

Hongyang Zhang  
University of Waterloo  
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 \( \ell_{\infty} \)-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 \( \ell_{\infty} \)-norm and unrestricted threat models. We propose a principled method for the partition of training and validation models by encouraging intra-group diversity while penalizing extra-group similarity. 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 Challenge: 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.

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]. 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 \( \ell_{\infty} \)-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 6). On the other hand, unrestricted 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 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 \( S_{\text{total}} \)) between transferability and imperceptibility. As shown in Fig. 4, the transfer success rate of our method GA-DTMIFGSME surpasses the baseline DTMIFGSME (see Eq. (1)) by up to 16% in absolute value under the same average perturbation reward (see Eq. (5)). Besides, our method GA-DTMIFSA is able to generate semantic-preserving yet transferable unrestricted adversarial examples under the unrestricted threat model (see Eq. (3), Figs. 1, 6, 10, and 11).
We propose a Geometry-Aware (GA) framework, where...

Under unrestricted setting, we propose a transfer-based...

B. 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 $\ell_\infty$-norm setting, our GA 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-DTM-FGSM to the CVPR’21 Security AI Challenger [24], 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 [25] with transfer-based $\ell_\infty$-norm attacks to generate semantic-preserving yet transferable adversarial examples (see Figs. 1 and 10). Moreover, the crafted adversarial examples transfer well to adversarially robust models (see Table. III, Figs. 6 and 11).

II. RELATED WORK

A. $\ell_p$-norm Adversarial Examples

Existing gradient-based white-box attacks either search for adversarial examples in a fixed $\ell_p$-norm ball [26], [27], or optimize the perturbation for each image independently to get a minimum-norm solution such as DeepFool [28], CW [29], and fast adaptive boundary attack [30]. However, white-box assumption usually does not hold in real-world scenarios. In query-based black-box setting, attackers utilize output logits [31], [32] or predicted label [33], [34] of the target model to generate adversarial examples. But these attacks typically suffer from high query complexity, making it easy to be detected [35]. Transfer-based black-box attacks [36]–[39] 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 $\ell_p$-norm distance is not an ideal perceptual similarity metric [40], [41], which oversimplifies the diversity of real-world perturbations. Unrestricted adversarial examples have received significant attention in recent years [42]. Most of the current unrestricted attacks aim to generate imperceptible adversarial examples under white-box setting, such as geometric transformations [19], [43], [44] and distance metrics beyond $\ell_p$ norm [20], [21]. Color-based attacks [45]–[50] 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 [51]–[54] search for adversarial embeddings in...
the latent space. Style transfer [50], [55] is inherently an
unrestricted attack as it preserves the semantic of the content
image. However, constructing transferable unrestricted adver-
sarial examples is still less explored. In this work, we will fill
this gap by combining the white-box feature space attack [25]
with transfer-based \( \ell_\infty \)-norm attacks to generate semantic-
preserving yet transferable unrestricted adversarial examples.

C. Adversarial Defenses

There have been long-standing arms races between de-
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 [26], and TRADES [56] with a new robust loss based on
the trade-off between robustness and accuracy. Geometry-
aware instance-reweighted adversarial training [57], which
is proven falling into gradient masking [58], 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) [59], aiming to improve robustness
against perturbations that were unseen during training. Unlike
empirical defenses, Certified defenses [60–63] could provide
robustness guarantee under a certain \( \ell_p \)-norm budget.

III. PRELIMINARIES

Notation. A deep neural network classifier can be described
as a function \( f(x; \theta) : \mathcal{X} \rightarrow \mathbb{R}^C \), parameterized by weights
\( \theta \), which maps a vector \( x \) in \( \mathcal{X} \) to its output logits. Given an
input \( x \) of class \( y \in \{1, 2, \ldots, C\} \), the predicted label of
\( f(x; \theta) \) is \( \hat{y} = \arg \max_j f_j(x; \theta) \), where \( f_j(x; \theta) \) represents
the \( j \)-th entry of \( f(x; \theta) \). We use \( L(f(x; \theta), y) \) to represent the cross-entropy loss and denote the \( \varepsilon \)-neighborhood
of \( x \) by \( B(x, \varepsilon) := \{ x' \in \mathcal{X} : D(x, x') \leq \varepsilon \} \), where \( D \)
is a distance metric that describes the change between the
adversarial example \( x' \) and the nature example \( x \). We denote
the black-box test model by \( g \), and split the set of accessible
source models \( \Theta = \{ \theta_1, \theta_2, \ldots, \theta_g \} \) into the set of training
models \( f \) and the set of validation models \( h \).

A. Transfer-based \( \ell_\infty \)-norm Attacks

Existing transfer-based attacks typically search for adver-
sarial examples in a fixed-radius \( \ell_p \)-norm ball, i.e.,
\( D(x, x') = \| x' - x \|_p \leq \varepsilon \). Various methods were pro-
posed 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 \( \ell_\infty \)-norm baseline
DTMI-FGSM by combining all these techniques, i.e.,
\[
\begin{align}
    m_{t+1} &= \gamma \cdot m_t + W \ast \nabla_{x_t} L(f(T(x_t, p); \theta), y) \\
    &\quad + \| W \ast \nabla_{x_t} L(f(T(x_t, p); \theta), y) \|_1, \\
    x_{t+1} &= \Pi_{B(x, \varepsilon)} (x_t + \alpha \cdot \text{sign}(m_{t+1})),
\end{align}
\]

where \( m_0 = 0, W \) is a pre-defined kernel with a convolu-
cation operation \( * \), \( \alpha \) is the step size, \( \Pi \) is the projection operator,
and \( \gamma \) is the decay factor for the momentum term. \( T(x_t, p) \)
represents the input transformation on \( x_t \) 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 [64] in style transfer, Xu et al. [25]
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) [25]
with transfer-based \( \ell_\infty \)-norm attacks [16], [17]. Given an
encoder \( \phi \), we extract the style features of input \( x \) as channel-
wise mean \( \mu(\phi(x)) \in \mathbb{R}^C \) and channel-wise standard devia-
tion \( \sigma(\phi(x)) \in \mathbb{R}^C \). Specifically,
\[
\begin{align}
    \mu_c &= \frac{1}{HW} \sum_{h=1}^{H} \sum_{w=1}^{W} \phi_c(x)_{hw}, \\
    \sigma_c &= \left( \frac{1}{HW} \sum_{h=1}^{H} \sum_{w=1}^{W} (\phi_c(x)_{hw} - \mu_c)^2 \right)^{1/2},
\end{align}
\]

where \( \phi(x) \in \mathbb{R}^{C \times H \times W} \) represents the latent embedding. Xu
et al. [25] adds adversarial perturbations on \( \mu \) and \( \sigma \) before
projecting \( \phi(x) \) back to the input space \( \mathcal{X} \) with a pre-trained
1 decoder \( \phi^{-1} \), namely,
\[
\hat{\phi}(x) = e^{\tau^x} \cdot (\phi(x) - \mu) + e^{\tau^\sigma} \cdot \mu,
\]
\[
x' = \phi^{-1} (\hat{\phi}(x)), \quad \| \tau^\sigma \|_\infty \leq \ln \varepsilon, \| \tau^\mu \|_\infty \leq \ln \varepsilon,
\]

where \( \phi(x) \) enlarges or shrinks the mean \( \mu \) and the standard
deviation \( \sigma \) of the embedding \( \phi(x) \) by a factor of \( e^{\tau^\mu} \) and
\( e^{\tau^\sigma} \), respectively. In this way, the distance metric
\( D(x, x') = \max (\| \tau^\sigma \|_\infty, \| \tau^\mu \|_\infty) \leq \varepsilon \). In order to preserve the semantic
of the unrestricted adversarial example \( x' \), a content loss was
added during the attacking process, i.e.,
\[
\min_{\tau^\sigma, \tau^\mu} L(x', y) = \lambda \cdot \text{top-5} (f(x'; \theta), y) + \| \phi(x') - \hat{\phi}(x) \|_2,
\]

where \( \lambda \) balance the trade-off between adversarial and the
content loss. Following Xu et al. [25], 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
\( \ell_\infty \)-norm attack on parameters \( \tau^\mu \) and \( \tau^\sigma \). Moreover, the
same techniques in Sec. III-A such as input diversity \( T(x', p) \) (only
for the margin loss) and momentum-based method can be
integrated to improve transferability, i.e.,
\[
\begin{align}
    m_{t+1} &= \gamma \cdot m_t + \nabla_{x_t} L(f(T(x_t, p); \theta), y) \\
    &\quad + \| \nabla_{x_t} L(f(T(x_t, p); \theta), y) \|_1, \\
    x_{t+1} &= \Pi_{B(x, \varepsilon)} (x_t + \alpha \cdot \text{sign}(m_{t+1})),
\end{align}
\]

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

1 We use the official pre-trained shallow decoder: https://github.com/
quilingxu/FeatureSpaceAttack.
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 [65]. Sharif et al. [66] found that $\ell_p$-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 $D$ as our main evaluation metrics. Consider a dataset $S = \{(x_i, y_i)\}_{i=1}^N$ and the corresponding adversarial examples $S_{adv} = \{(x'_i, y_i)\}_{i=1}^N$ that are crafted on the training model $f$. Let $N_0 = \sum_{i=1}^N \mathbb{1}\{\hat{g}(x'_i) \neq y_i\}$ be the number of misclassified adversarial examples on the test model $g$. We define the average total score as:

$$S_{total} = \frac{1}{N} \sum_{i=1}^N \mathbb{1}\{\hat{g}(x'_i) \neq y_i\} \cdot F_{\text{reward}}(D(x_i, x'_i))$$
$$= \frac{N_0}{N} \cdot \frac{1}{N_0} \sum_{i=1}^{N_0} \mathbb{1}\{\hat{g}(x'_i) \neq y_i\} \cdot F_{\text{reward}}(D(x_i, x'_i))$$
$$= \frac{N_0}{N} \cdot S_{\text{APR}},$$

(5)

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

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 $D$, i.e.,

$$\min_{x'} D(x, x'), \quad \text{s.t.} \quad \hat{g}(x') \neq y.$$  

(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 (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 have 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 dynamic radii. To approximately solve problem (6), we first divide the attack in the ball $B(x, \varepsilon)$ into $K$ sub-procedures. In the $k$-th sub-procedure, we re-run a fixed-radius attack algorithm $A$ such

Algorithm 1 Geometry-Aware Framework

Require:
Benign input $x$ with label $y$; training models $f$; validation model $h$; number of sub-procedures $K$; maximum perturbation size $\varepsilon$ and threshold $\eta$; attack algorithm $A$;

Ensure:
Transfer-based unrestricted adversarial example $x'$ with approximately minimum change;

1: $x_0 = x$;
2: for $k = 1, 2, \ldots, K$ do
3: $x_k = A(x, x_{k-1}, f, \frac{\varepsilon}{K})$; ▷ fixed budget
4: $\text{conf} \leftarrow \frac{\text{exp}(h(x_i, y_i))}{\sum_i \text{exp}(h(x_i, y_i))}$;
5: if $\text{conf} < \eta$ then
6: Return $x_K$; ▷ early-stopping in Eq. (7)
7: end if
8: end for
9: Return $x_K$;
as DMI-FSA (see Eq. (4)) under the perturbation budget
\[ \epsilon_k = \frac{k}{K} \times \epsilon, \quad k = 1, 2, \cdots, K. \]

Each sub-procedure starts from the solution of 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 true class on the validation model is smaller than a threshold \( \eta \), i.e.,
\[ P \left( \hat{h}(x) = y \right) = \frac{\exp(h_y(x; \theta))}{\sum_j \exp(h_j(x; \theta))} < \eta. \tag{7} \]

Our GA framework is summarized in Algorithm 1 and illustrated in Fig. 2. Note that the output of GA framework is related to the choice of the training model \( f \) and the validation model \( h \). Thus, it is important to figure out which partition of the source models performs better.

We split \( n \) pre-trained models \( \Phi = \{ \phi_1, \phi_2, \cdots, \phi_n \} \) into \( k \) training models and \( n - k \) validation models. Instead of traversing all possible partitions to select the optimal split by querying the test model \( g \) by \( C_n^k \) times, we propose a query-free approach that only utilizes the information of transferability between the pre-trained models (see Fig. 3). Let \( w_{ij} \) \( (w_{ij} \neq w_{ji}) \) be the transfer success rate (\( \frac{N_{ij}}{N} \) in Eq. (5)) from the source model \( \phi_i \) to the target model \( \phi_j \) under a fixed-radius attack (e.g., DMI-FGSM). Denote the binary partition function as \( \tilde{g} \). The training set and the validation set can be formulated as \( T = \{ i \mid \tilde{g}(\phi_i) = 0 \} \) and \( V = \{ j \mid \tilde{g}(\phi_j) = 1 \} \), respectively. We define the partition loss \( \ell_G \) as:
\[ \ell_G = \frac{1}{k} \sum_{i \in T} \ell_G(\phi_i) + \frac{1}{n-k} \sum_{j \in V} \ell_G(\phi_j), \]
\[ \ell_G(\phi_i) = \frac{1}{k-1} \sum_{j \neq i, j \in T} w_{ji} + \frac{1}{n-k} \sum_{t \in V} w_{it}, \tag{8} \]
\[ \ell_G(\phi_j) = \frac{1}{n-k-1} \sum_{i \neq j, i \in T} w_{ij} + \frac{1}{k} \sum_{t \in V} w_{jt}. \]

For both the training set loss \( \ell_G(\phi_i) \) and the validation set loss \( \ell_G(\phi_j) \): 1) Minimizing the first formula on the right of Eq. (8) encourages intra-group diversity. To make the decision boundary of the ensemble model \( f \) or \( h \) more general and effective, we minimize the transfer success rate between any two pre-trained models inside the group. 2) Minimizing the second formula on the right of Eq. (8) is penalizing extra-group similarity. If \( f = h \), the early-stopping mechanism in Eq. (7) will be triggered too early, leading to small adversarial perturbations for all inputs. Reducing the transferability from the training model \( f \) to the validation model \( h \) might improve generalization of the adversarial examples to unknown models. Empirically, we find that the proposed partition loss \( \ell_G \) negatively correlates with the average total score \( S_{total} \).

V. EXPERIMENTS

A. Experimental Setup

Datasets & Networks. Similar to Xie et al. [17], we randomly select 1,000 images from ILSVRC 2012 validation set [67], 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 [68]. More details about the networks are summarised in Table I. The transferability between these models under \( \ell_{\infty} \)-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.

---

**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 |
|----------|-------|------------|----------------|------------|
| Normal   | 0     | ViT-S/16   | 76.01%         | 99.8%      |
|          | 2     | Swin-B/patch4-window7 | 84.23% | 99.4% |
|          | 4     | ResNeXt50-32x4d-ssl | 78.90% | 99.7% |
|          | 6     | Inception-v3 | 76.94% | 100.0% |
| Ensemble | 8     | Ens-adv-Inception-v3 | 76.49% | 100.0% |
|          | 9     | Ens-adv-Inception-ResNet-v2 | 78.98% | 99.9% |
Implementation Details. Given the maximum perturbation size $\varepsilon$ and number of sub-procedures $K$ (5 as default) in our geometry-aware framework, we set the step size $\alpha = \frac{1.55\varepsilon}{K+1}$ in the $k$-th sub-procedure, where the number of iterations $T$ is set to 10 in the $\ell_\infty$ setting and 50 in the unrestricted setting. $\varepsilon$ is set to 20 in the $\ell_\infty$-norm setting and 3.5 in the unrestricted setting. When running a fixed-radius baseline at perturbation budget $\varepsilon_k$, we set the number of iteration as $\frac{T}{7}(1+\frac{k-1}{2})$ with step size $\alpha$ 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 $F_{\text{reward}}(\varepsilon_0)$ 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. [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 $5 \times 5$.

### 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 1, respectively. The test model is Inception-ResNet-v2. The optimal threshold $\eta$ (see Eq. (7)) is searched from a finite set ranging from 0.001 to 0.9 by querying the test model $g$ to achieve the best average total score $S_{\text{total}}$. For each $\eta$, we execute our method and compute the average perturbation reward $S_{\text{APR}}$ (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 $\ell_\infty$-norm setting, we combine our Geometry-Aware (GA) framework with DI-FGSM, DTI-FGSM, DMI-FGSM, and Admix-DTI-FGSM [38] (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-DTI-MI-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 $\eta$ (leading to different $S_{\text{APR}}$). Fixing $S_{\text{APR}}$ as 0.115, our approach GA-DTI-FGSM surpasses the baseline DMI-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 $S_{\text{total}}$ across various threat models, especially in the $\ell_\infty$-norm setting where both the transfer success rate and $S_{\text{APR}}$ 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 $S_{\text{total}}$. TSR: Transfer Success Rate.

![Contour of average total score $S_{\text{total}}$ (higher is better). Fixing transfer success rate as 80%, our approach GA-DTI-FGSM surpasses the baseline DMI-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 $S_{\text{total}}$ across various threat models, especially in the $\ell_\infty$-norm setting where both the transfer success rate and $S_{\text{APR}}$ are improved.](attachment:image.png)

### C. Case Study: CVPR’21 Security AI Challenger

In the CVPR’21 Security AI Challenger: Unrestricted Adversarial Attacks on ImageNet [24], contestants were asked to submit adversarial examples without any access to the defense models. The dataset is a subset of ILSVRC 2012 validation set [67], 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 $S_q = 0$, otherwise $S_q = 1$. The image quality $S_q$ (equivalent to our reward function

\[ S_q = \frac{1}{N} \sum_{i=1}^{N} I_q(i) \]

where $N$ is the number of images in the submission, and $I_q(i)$ is a binary indicator of whether the $i$-th image is considered to be of high quality.
Implemention Details. For the ReColor attack [46], 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 [25], 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 [73].

Experimental Results. From Table. III, we can conclude the following observations: a) GA-Admix-DTI transfers better than GA-DTMIFGSM. b) $\ell_\infty$-norm transfer attack can hardly break $\ell_\infty$-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. 6). c) DeepAugment [73], which utilizes semantic-preserving augmentations during training, exhibits non-trivial robustness against GA-DMI-FSA attack. d) Efficientnet-l2-ns [74] 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 close to 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 Resnext101-DenoiseAll [71] in Figs. 6, where our method GA-DMI-FSA is able to generate semantic-preserving yet transferable unrestricted adversarial examples. For more visualization results, please see Fig. 11 in Sec. VII-A.
Fig. 6: Visualization of transfer attack results on Resnext101-DenoiseAll [71]. 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 $\ell_\infty$-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-DTMI-FGSM). Besides, the other two unrestricted attacks suffer from weaker transferability when compared to our method GA-DMI-FSA, which adjusts the images’ color and texture that ImageNet-trained CNNs might be biased to [79].

E. Ablation Studies and Discussions

The optimal $\eta$ depends on the train-valid splitting. As declared in the implementation details in Sec. V-B, the threshold $\eta$ is searched from a finite set ranging from 0.001 to 0.9. We now investigate how the hyper-parameter $\eta$ will affect the average total score $S_{total}$. From Fig. 7, we observe that the optimal $\eta^*$ varies across different splittings and can be larger if the transferability from the training model $f$ to the test model $g$ is higher enough. However, as shown in Table IV, the improvement of our GA framework upon fixed-radius baseline (DTMI-FGSM) is stable and independent of the splitting.

TABLE IV: Comparison of the improvement upon fixed-radius baseline under different train-valid splittings. The test model $g$ is model 7 in Table I and we select three (two of them are highlighted in Fig. 8a) representative splittings according to the partition loss $\ell_G$. For each setting, we repeat the experiment three times and report the mean and the standard deviation (in the parenthesis).

| $f$   | $h$   | $\ell_G$ | Method     | $S_{avg}$ ($\%$) | $S_{total}$ ($\%$) |
|-------|-------|----------|------------|------------------|-------------------|
| (4, 5, 6) | (1, 2, 3) | 2.22     | DTMI       | 0.1080 (0.001)   | 8.66% (0.16%)     |
|        |        |          | GA-DTMI    | 0.1456 (6e-5)    | 11.6% (0.21%)     |
| (2, 3, 5) | (1, 4, 6) | 2.56     | DTMI       | 0.1130 (0.006)   | 7.32% (0.07%)     |
|        |        |          | GA-DTMI    | 0.1417 (0.001)   | 9.48% (0.03%)     |
| (1, 2, 3) | (4, 5, 6) | 2.70     | DTMI       | 0.1273 (0.007)   | 6.41% (0.07%)     |
|        |        |          | GA-DTMI    | 0.1291 (0.006)   | 8.23% (0.05%)     |

Fig. 7: Comparison between various train-valid splittings. The index in the legend corresponds to the model index in Table I. As the threshold $\eta$ 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 $\eta^*$ that yields the maximum $S_{total}$ is dependent on the partition (right).

The effectiveness of $\ell_G$ under different numbers of pre-trained models and different kinds of pre-trained models.

To investigate the robustness of the proposed partition loss $\ell_G$ under different settings, we carefully design controlled experiments in Figs. 8 and 9. There are total $C_n^k$ kinds of partitions when selecting $k$ training models from $n$ pre-trained models. Given $n$ and $k$, we run our method GA-DTMI-FGSM for all the train-valid splittings and obtain a scatter plot with $C_n^k$ points. We observe a strong negative correlation between
and the average total score $S_{\text{total}}$. For example, the average Pearson correlation coefficient over the six scatterplots in Figs. 8 and 9 is around -0.82. Moreover, the negative correlation is significant and consistent across different numbers of pre-trained models (see Fig. 8) and different kinds of pre-trained models (see Fig. 9).

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 $\ell_\infty$-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 $\ell_\infty$-norm attacks to generate semantic-preserving yet transferable unrestricted adversarial examples.

Acknowledgements. Fangcheng Liu and Chao Zhang are supported by the National Nature Science Foundation of China under Grant 62071013 and 61671027, and National Key R&D Program of China under Grant 2018AAA0100300. Hongyang Zhang is supported by NSERC Discovery Grant RGPIN-2022-03215, DGECR-2022-00357.

Fig. 8: The effectiveness of the partition loss $\ell_G$ under different numbers of pre-trained models. The test model $g$ is model 7 in Table I. We conduct bayesian ridge regression and plot the mean of the predictive distribution as dashed lines. The Bayesian Credible Intervals range from mean - standard deviation (of the predictive distribution) to mean + standard deviation.

Fig. 9: The effectiveness of the partition loss $\ell_G$ under different kinds of pre-trained models. The test model $g$ is model 7 in Table I. We simulate different kinds of pre-trained models by removing one model from the fixed set such that $(n, k) = (6, 3)$.

REFERENCES

[1] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in CVPR, 2016, pp. 770–778.
[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.
[3] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, “Intriguing properties of neural networks,” in ICLR, 2014.
[4] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” in ICLR, 2015.
[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.
[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.
[7] Y. Bai, J. Mei, A. Yuille, and C. Xie, “Are transformers more robust than cnns?” in NeurIPS, 2021.
[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.
[9] O. M. Parkhi, A. Vedaldi, and A. Zisserman, “Deep face recognition,” in BMVC, BMVA Press, September 2015, pp. 41.1–41.12.
[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.
[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.

[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.

[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 security, 2017, pp. 506–519.

[14] Y. Liu, X. Chen, C. Liu, and D. Song, “Delving into transferable adversarial examples and black-box attacks,” in ICLR, 2017.

[15] F. Tramer, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel, “Ensemble adversarial training: Attacks and defenses,” in ICLR, 2018.

[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.

[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.

[18] Y. Dong, T. Pang, H. Su, and J. Zhu, “Evasion defenses to transferable adversarial examples by translation-invariant attacks,” in CVPR, 2019, pp. 4312–4321.

[19] C. Xiao, J.-Y. Zhu, B. Li, W. He, M. Liu, and D. Song, “Spatially transformed adversarial examples,” in ICLR, 2018.

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

[21] C. Laidlaw, S. Singla, and S. Feizi, “Perceptual adversarial robustness: Defense against unseen threat models,” in ICLR, 2021.

[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.

[23] Z. Katsir and Y. Elovici, “Who’s afraid of adversarial transferability?” arXiv preprint arXiv:2105.00433, 2021.

[24] Y. Chen, X. Mao, Y. He, H. Xue, C. Li, Y. Dong, Q.-A. Fu, X. Yang, W. Xiang, T. Pang et al., “Unrestricted adversarial attacks on ImageNet competition,” arXiv preprint arXiv:2110.09901, 2021.

[25] Q. Xu, G. Tao, S. Cheng, and X. Zhang, “Towards feature space adversarial attack by style perturbation,” AAAI, vol. 35, no. 12, pp. 10523–10531, May 2021.

[26] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, “Towards deep learning models resistant to adversarial attacks,” in ICLR, 2018.

[27] 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.

[28] 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.

[29] 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.

[30] F. Croce and M. Hein, “Minimally distorted adversarial examples with a fast adaptive boundary attack,” in ICML. PMLR, 2020, pp. 2196–2205.

[31] 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.

[32] 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.

[33] W. Brendel, J. Rauber, and M. Bethge, “Decision-based adversarial attacks: Reliable attacks against black-box machine learning models,” in ICLR, 2018.

[34] 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.

[35] D. Willmott, A. K. Sahu, F. Sheikhholeslami, F. Condessa, and Z. Kolter, “You only query once: Effective black box adversarial attacks with minimal repeated queries,” arXiv preprint arXiv:2102.00029, 2021.
smoothed classifiers,” in NeurIPS. Red Hook, NY, USA: Curran Associates Inc., 2019.

[62] 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.

[63] K. Leino, Z. Wang, and M. Fredrikson, “Globally-robust neural networks,” in ICML, 2021.

[64] X. Huang and S. Belongie, “Arbitrary style transfer in real-time with adaptive instance normalization,” in ICCV, 2017, pp. 1501–1510.

[65] 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.

[66] 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.

[67] 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.

[68] R. Wightman, “Pytorch image models,” https://github.com/rwightman/pytorch-image-models, 2019.

[69] E. Wong, L. Rice, and J. Z. Kolter, “Fast is better than free: Revisiting adversarial training,” in ICLR, 2020.

[70] A. Shafahi, M. Najibi, M. A. Ghiassi, 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.

[71] 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.

[72] 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.

[73] 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.

[74] 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.

[75] Z. Liu, Y. Lin, Y. Can, H. Hu, Y. Wei, Z. Zhang, S. Lin, and F. Guo, “Swin transformer: Hierarchical vision transformer using shifted windows,” arXiv preprint arXiv:2103.14030, 2021.

[76] F. Tramer and D. Boneh, “Adversarial training and robustness for multiple perturbations,” in NeurIPS, vol. 32. Curran Associates, Inc., 2019.

[77] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar, “Do ImageNet classifiers generalize to ImageNet?” in ICML. PMLR, 2019, pp. 5389–5400.

[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.

[79] 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.
VII. APPENDIX

A. More Visualization Results

Fig. 10: Adversarial examples of GA-DMI-FSA, which are misclassified by all the models in Table. III. Top: benign examples $x$. Middle: unrestricted adversarial examples $x'$. Bottom: normalized adversarial perturbations $x' - x$.

Fig. 11: Visualization of transfer attack on Resnext101-DenoiseAll [71].