AGS: Affordable and Generalizable Substitute Training for Transferable Adversarial Attack

Ruikui Wang[1,2], **Yuanfang Guo**[1,2*], **Yunhong Wang**[2]

[1]State Key Laboratory of Software Development Environment, Beihang University, China
[2]School of Computer Science and Engineering, Beihang University, China
{rkwang, andyguo, yhwang}@buaa.edu.cn

Abstract

In practical black-box attack scenarios, most of the existing transfer-based attacks employ pretrained models (e.g. ResNet50) as the substitute models. Unfortunately, these substitute models are not always appropriate for transfer-based attacks. Firstly, these models are usually trained on a large-scale annotated dataset, which is extremely expensive and time-consuming to construct. Secondly, the primary goal of these models is to perform a specific task, such as image classification, which is not developed for adversarial attacks. To tackle the above issues, i.e., high cost and over-fitting on task-specific models, we propose an **A**ffordable and **G**eneralizable **S**ubstitute (AGS) training framework tailored for transfer-based adversarial attack. Specifically, we train the substitute model from scratch by our proposed adversary-centric constrastive learning. This proposed learning mechanism introduces another sample with slight adversarial perturbations as an additional positive view of the input image, and then encourages the adversarial view and two benign views to interact comprehensively with each other. To further boost the generalizability of the substitute model, we propose adversarial invariant learning to maintain the representations of the adversarial example invariants under augmentations with various strengths. Our AGS model can be trained solely with unlabeled and out-of domain data and avoid overfitting to any task-specific models, because of its inherently self-supervised nature. Extensive experiments demonstrate that our AGS achieves comparable or superior performance compared to substitute models pretrained on the complete ImageNet training set, when executing attacks across a diverse range of target models, including ViTs, robustly trained models, object detection and segmentation models. Our source codes are available at https://github.com/lwmming/AGS.

Introduction

In recent years, a wide range of computer vision tasks (He et al. 2016b; Ren et al. 2015) have been revolutionized by Deep Neural Networks (DNNs). Despite the impressive capabilities of DNNs, they are exceedingly vulnerable to adversarial examples, i.e., the imperceptible perturbation injected on the input can easily alter the model's decision. This phenomenon raises security concerns in current deployed DNN-based applications, such as face recognition application (Vakhshiteh, Ramachandra, and Nickabadi 2020), autonomous driving (Kumar et al. 2021), etc. Consequently, research on adversarial example is necessary and imperious.

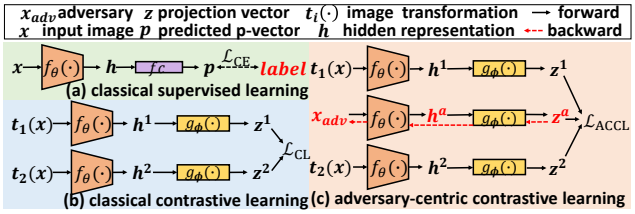


Figure 1: Conceptual functioning pipelines of different substitute training schemes, whose backgrounds are best viewed in distinctive colors.

Actually, the key property that allows the adversarial examples to be practically applied is the transferability (Szegedy et al. 2014), referring to adversarial examples generated on one model can successfully deceive other models even with different architectures. Nowadays, a growing number of methods (Dong et al. 2018; Xie et al. 2019; Huang et al. 2019; Lin et al. 2020; Wu et al. 2020; Wang et al. 2021a; Zhang et al. 2022a; Yang et al. 2022; Zhang et al. 2022b; Huang and Kong 2022) have been proposed to boost the transferability of adversarial examples. As shown in Fig. 1(a), the substitute models employed in these methods are typically pretrained on a large-scale annotated dataset, e.g., ImageNet (Russakovsky et al. 2015). Unfortunately, constructing such datasets costs significant human efforts and time, which is often prohibitive to afford. Meanwhile, the employed substitute models are usually optimized for specific tasks. For instance, ResNet50 (He et al. 2016a), a commonly used substitute model, is optimized with the cross-entropy loss to primarily fulfill image classification tasks. Consequently, the adversarial examples crafted on ResNet50 tend to overfit to the task-specific models.

The aforementioned issues tend to limit the practical applicability of the transfer-based black-box attacks. Recently, several literatures (Li, Guo, and Chen 2020; Ban and Dong 2022; Sun et al. 2022; Li et al. 2023; Malik et al. 2022) have sought to overcome these limitations by harnessing self-supervised models (Chen et al. 2020; Chen and He 2021)

*Corresponding author.

as substitute models. Notably, Li *et al.* (Li, Guo, and Chen 2020) pioneer to train an auto-encoder model as the substitute model. Then, this work is further improved by (Malik et al. 2022), where an auto-encoder framework is trained with a min-max objective. Subsequently, contrastive learning has been utilized for training the substitute model (Ban and Dong 2022; Sun et al. 2022; Li et al. 2023). However, as shown in Fig. 1(b), most of these methods straightforwardly adopt contrastive learning without specifically considering interactions between benign and adversarial examples, resulting in unsatisfactory adversarial transferability.

In this paper, to tackle the limitations and issues above, we present the affordable and generalizable substitute training framework (AGS) for transfer-based adversarial attack. In general, AGS explicitly establishes connection between contrastive learning and adversarial transferability. Specifically, as shown in Fig. 1(c), on the basis of classical contrastive learning, we introduce another sample with slight adversarial perturbations as an additional type of positive view of the input image. This strategy enriches both the set of positive and negative views, and fosters comprehensive interactions between the adversarial and benign examples. Consequently, both the instance discrimination between the clean and adversarial examples and adversarial robustness of the substitute model are enhanced, i.e., adversarial examples with better transferability can be generated. Besides, inspired by the proven efficacy of invariant learning in enhancing model generalization across various downstream tasks (Foster, Pukdee, and Rainforth 2021), we further improve AGS with an elaborated adversarial invariant learning regularization. This regularization aims to further boost the generalizability of the trained substitute model, to allow the crafted adversarial examples to effectively mislead models across diverse tasks. In general, our AGS is trained only with the unlabeled and out-of-domain data, which can easily be collected and makes our substitute model training affordable. Meanwhile, guided by the self-supervised signals, our AGS avoids overfitting to any task-specific models.

Our main contributions are briefly summarized as follows: **1)** We propose an affordable and generalizable substitute training framework that enables the effectiveness of transfer-based attacks solely with unlabeled and out-of-domain data; **2)** We develop two optimization schemes, adversarial-centric contrastive learning and adversarial invariant learning tailored for substitute training; **3)** Comprehensive attack experiments on the target models with different architectures from various vision tasks demonstrate that our AGS achieves excellent adversarial transferability.

Related Work

Transfer-based attack typically fools the black-box target models, i.e., the details of the target model is unknown, via adversarial examples, which are crafted with a local substitute model. According to the application scenarios and the amount of resources the attacker can access, existing transfer-based attack methods can be briefly classified into three categories, i.e., cross-architecture, cross-domain and cross-paradigm transfer-based methods.

Cross-Architecture Transfer-based Attack. Cross-architecture transfer-based attack methods assume that the substitute and target models are trained in the same data domain. Under this setting, gradient-based methods (Dong et al. 2018; Lin et al. 2020; Wang and He 2021) achieve high transferability by advanced optimization algorithm, such as momentum and variance tuning, which leads to stronger yet more stable gradients. Meanwhile, intermediate level attacks (Huang et al. 2019; Inkawhich et al. 2020; Wu et al. 2020; Wang et al. 2021b, 2022; Zhang et al. 2022a) improve adversarial transferability by perturbing the features from intermediate layers. In addition, the adversarial transferability can also benefit from the input augmentation strategies (Xie et al. 2019; Lin et al. 2020; Wang et al. 2021a; Byun et al. 2022). Lastly, some methods also boost adversarial transferability by refining the substitute model, such as (Wu et al. 2019; Gubri et al. 2022). More recently, some methods (Qin et al. 2022; Gubri et al. 2022; Wu et al. 2018) study the adversarial transferability from the perspective of flatness in the input space and weight space.

Cross-Domain Transfer-based Attack. For the cross-domain attacks, the training data for the substitute model and target models is from different domains, which is more practical than the cross-architecture setting. Some typical methods include CDA (Naseer et al. 2019), TTP (Naseer et al. 2021), LTP (Salzmann et al. 2021), BIA (Zhang et al. 2022b) and GAMA (Aich et al. 2022).

Cross-Paradigm Transfer-based Attack. For the cross-paradigm attacks, on the basis of cross-domain, further require that the training data of substitute models is unlabeled. Therefore, the knowledge of these three types of attacks gradually decreases, and the attack task gradually becomes more practical yet challenging. To tackle this setting, recently, a number of studies explore to generate the adversarial examples based on the models trained by self-supervised learning. Since the substitute and target models are trained by different learning paradigms, we name this setting as the cross-paradigm transfer-based attack. Among this type of methods, Li *et al.* (Li, Guo, and Chen 2020) train multiple classical auto-encoder models as substitute models. However, their most competitive scheme, 'Prototypical', is not purely label-free, i.e., it still needs partial annotations. Although Ban *et al.* (Ban and Dong 2022), Sun *et al.* (Sun et al. 2022) and Li *et al.* (Li et al. 2023) explore to incorporate classical contrastive learning into substitute training, they have not considered the interaction of the benign and adversarial examples, which leads to unsatisfactory adversarial transferability. Malik *et al.* (Malik et al. 2022) use min-max objective to train effective substitute models. However, its framework is based on auto-encoder, which generally performs weaker than contrastive learning in terms of self-supervised representation. In this paper, to tackle cross-paradigm setting, we aim to develop more advanced self-supervised based substitute training methods for better adversarial transferability.

Methodology

Our goal is to train a dedicated substitute model tailored for transferable adversarial attacks. Intuitively, this substi-

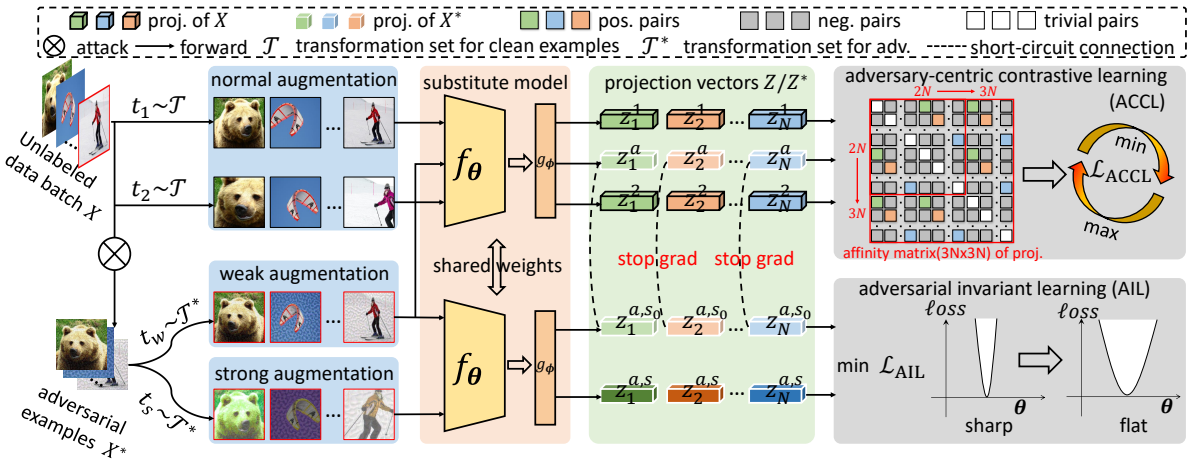


Figure 2: Illustration of our proposed AGS (For projection vectors, the ones from different instances are marked in distinctive colors. The ones in the same color with different brightness denote the projection vectors with different strengths of augmentations. In ACCL, the size of the affinity matrix of projection vectors is expanded from 2N×2N to 3N×3N, due to the participation of the adversarial view. In AIL, the representations of the adversarial examples under augmentations with various strengths examples are encouraged to be invariant).

tute model is desired to possess two merits: 1) Affordability, indicating that its training corpus is easily attainable without the need for extensive annotations; 2) Generalizability, signifying that adversarial examples crafted using this substitute model can seamlessly transfer and deceive different unknown architectures across diverse vision tasks.

Preliminaries

Given N unlabeled images $\boldsymbol{X} = \{\boldsymbol{x_1}, \boldsymbol{x_2}, ..., \boldsymbol{x_N}\}$ drawn from data distribution $\mathcal{X}$, we aim to learn a substitute model $f_{\boldsymbol{\theta}}$ parameterized by $\boldsymbol{\theta}$. To achieve the goal mentioned above, the baseline scheme for substitute training is classical contrastive learning, in which the agreement between positive views is maximized and the discrepancy between the negative views is enlarged. Taking image $\boldsymbol{x_i}$ as an example, we firstly apply two transformations $t_1 \sim \mathcal{T}$ and $t_2 \sim \mathcal{T}$ to it, where $\mathcal{T}$ denotes the set of transformations for the benign examples, and obtain two positive views $t_1(\boldsymbol{x_i}), t_2(\boldsymbol{x_i})$. Then, we feed them into $f_{\boldsymbol{\theta}}$ with the projection head $g_{\boldsymbol{\phi}}$ parameterized by $\boldsymbol{\phi}$, and obtain the corresponding projection vectors, i.e., $\boldsymbol{z_i^j} = g_{\boldsymbol{\phi}} \circ f_{\boldsymbol{\theta}}(t_j(\boldsymbol{x_i})), j \in \{1, 2\}$. Then, classical contrastive learning can be formulated as

$$\ell_{\text{CL}}(\boldsymbol{u}, \boldsymbol{u}^+) = -\log \frac{\exp(cos(\boldsymbol{u}, \boldsymbol{u}^+)/\tau)}{\exp(cos(\boldsymbol{u}, \boldsymbol{u}^+)/\tau) + \sum_{\boldsymbol{v} \in \mathcal{N}(\boldsymbol{u})} \exp(cos(\boldsymbol{u}, \boldsymbol{v})/\tau)}, \tag{1}$$

where $\boldsymbol{u}$ denotes the projection vector of the anchor point, $\boldsymbol{u}^+$ represents the projection vector of $\boldsymbol{u}$'s positive sample, $\mathcal{N}(\boldsymbol{u})$ stands for the vector set of $\boldsymbol{u}$'s negative samples, and τ is a temperature parameter. When adapted to substitute model training, $\boldsymbol{u}$ and $\boldsymbol{u}^+$ are instantiated with $\boldsymbol{z_i^1}$ and $\boldsymbol{z_i^2}$, then the learning objective can be briefly written as

$$\mathcal{L}_{\text{CL}}(\boldsymbol{\phi} \circ \boldsymbol{\theta}; \boldsymbol{X}) := \frac{1}{2N} \sum_{i=1}^{N} (\ell_{\text{CL}}(\boldsymbol{z_i^1}, \boldsymbol{z_i^2}) + \ell_{\text{CL}}(\boldsymbol{z_i^2}, \boldsymbol{z_i^1})). \tag{2}$$

Overall Framework

In this paper, we propose a new substitute training framework, named AGS, as illustrated in Fig. 2. Each instance in input batch for AGS is augmented to obtain three views, i.e., two ordinary views obtained by normal augmentations w.r.t. the benign example and one adversarial view obtained by weak augmentation w.r.t the adversarial example. Then, they are sequentially processed by the substitute model and the projection head to extract the projection vectors. Then, two proposed optimization schemes are carried out. **Firstly**, the substitute model is trained by adversary-centric contrastive learning (ACCL), where three types of views from the same instance are pulled together and that from different instances are pushed away. Particularly, the adversarial examples for training are generated via maximizing ACCL in the inner loop. **Meanwhile**, the substitute model is optimized by the elaborated adversarial invariant learning (AIL), where the representations of the adversarial views under augmentations with various strengths are desired to be invariant. **Finally**, adversarial examples are crafted by perturbing the intermediate features of the trained substitute model in the cross-paradigm manner.

Adversary-Centric Contrastive Learning

Although classical contrastive learning has exhibited promising performance in the realm of self-supervised learning, its optimality for substitute training remains questionable. Given that substitute training aims to enhance adversarial transferability, classical contrastive learning lacks a

dedicated design to fulfill this objective. Our basic idea is to incorporate some adversarial components into contrastive learning process to explicitly establish a connection between contrastive learning and adversarial transferability. Inspired by recent studies (Springer, Mitchell, and Kenyon 2021; Malik et al. 2022), we incorporate samples with slight adversarial perturbations, which are crafted with a small adversarial budget, into classical contrastive learning, to enhance its alignment with adversarial transferability.

Actually, slight adversarial examples could naturally serve as an additional type of positive view of input images. In formal, with respect to $\boldsymbol{x_i}$, the adversarial example $\boldsymbol{x_i^*}$ is crafted as $\boldsymbol{x_i} + \boldsymbol{\delta_i}$. Its projection vector $\boldsymbol{z_i^a}$ can be obtained by: $\boldsymbol{z_i^a} = g_\phi \circ f_{\boldsymbol{\theta}}(\boldsymbol{x_i} + \boldsymbol{\delta_i})$. Subsequently, we simultaneously pull $\boldsymbol{z_i^a}$, $\boldsymbol{z_i^1}$ and $\boldsymbol{z_i^2}$ together, while push $\boldsymbol{z_i^a}$ far away from its negative view set $\mathcal{N}(\boldsymbol{z_i^a})$, as

$$\ell_{\mathrm{CL}}(\boldsymbol{z_i^a}, \boldsymbol{z_i^+}) = -\log \frac{\exp(\frac{cos(\boldsymbol{z_i^a}, \boldsymbol{z_i^+})}{\tau})}{\exp(\frac{cos(\boldsymbol{z_i^a}, \boldsymbol{z_i^+})}{\tau}) + \sum_{\boldsymbol{z_i^-} \in \mathcal{N}(\boldsymbol{z_i^a})} \exp(\frac{cos(\boldsymbol{z_i^a}, \boldsymbol{z_i^-})}{\tau})}. \quad (3)$$

By considering all the positive pairs, we can obtain

$$\ell_{\mathrm{ACCL}}(\boldsymbol{z_i^1}, \boldsymbol{z_i^a}, \boldsymbol{z_i^2}) = \frac{1}{|\mathcal{P}(\boldsymbol{z_i^a})|} \sum_{\boldsymbol{z_i^+} \in \mathcal{P}(\boldsymbol{z_i^a})} \ell_{\mathrm{CL}}(\boldsymbol{z_i^a}, \boldsymbol{z_i^+}), \quad (4)$$

where the positive view set $\mathcal{P}(\boldsymbol{z_i^a}) = \{\boldsymbol{z_i^a}, \boldsymbol{z_i^1}, \boldsymbol{z_i^2}\} - \{\boldsymbol{z_i^a}\}$ with cardinality of 2, the negative view set $\mathcal{N}(\boldsymbol{z_i^a}) = \{..., \boldsymbol{z_k^a}, \boldsymbol{z_k^1}, \boldsymbol{z_k^2}, ...\}(k \neq i)$ with cardinality of $(3N-3)$. Note that both $\mathcal{P}(\cdot)$ and $\mathcal{N}(\cdot)$ are enriched compared to their counterparts in classical contrastive learning, i.e., $|\mathcal{P}(\cdot)| : 1 \rightarrow 2; |\mathcal{N}(\cdot)| : (2N-2) \rightarrow (3N-3)$. This enrichment has two benefits. Firstly, the larger negative set will lead to increasing instance discrimination (He et al. 2020; Khosla et al. 2020); Secondly, Eq. (4) encourages substitute model to pull the projection vector of adversarial example, $\boldsymbol{z_i^a}$, and two clean samples, $\boldsymbol{z_i^1}$ and $\boldsymbol{z_i^2}$, together, which enhances the adversarial robustness. Furthermore, to foster comprehensive interactions between the adversarial and clean samples, we consider the all cases where each positive view in $\mathcal{P}(\boldsymbol{z_i^a})$ has the opportunity to be the central term and optimize the substitute model by minimizing the following objective

$$\mathcal{L}_{\mathrm{ACCL}}(\boldsymbol{\phi} \circ \boldsymbol{\theta}; \boldsymbol{X}) := \frac{1}{3N} \sum_{i=1}^{N} \{\ell_{\mathrm{ACCL}}(\boldsymbol{z_i^1}, \boldsymbol{z_i^a}, \boldsymbol{z_i^2}) + \ell_{\mathrm{ACCL}}(\boldsymbol{z_i^a}, \boldsymbol{z_i^1}, \boldsymbol{z_i^2}) + \ell_{\mathrm{ACCL}}(\boldsymbol{z_i^1}, \boldsymbol{z_i^2}, \boldsymbol{z_i^a})\}. \quad (5)$$

In Eq. (5), when it is compared to Eq. (2), the inclusion of the samples with slight adversarial perturbations contributes to the enhancement of both the instance discrimination and adversarial robustness in the substitute model.

Concurrently, we generate the slight adversarial perturbation $\boldsymbol{\delta_i}$ w.r.t. $\boldsymbol{x_i}$ for substitute training via

$$\boldsymbol{\delta_i} = \underset{\|\delta\|_2 \leq \epsilon_{train}}{\operatorname{argmax}} \mathcal{L}_{\mathrm{ACCL}}(\boldsymbol{\phi} \circ \boldsymbol{\theta}; \boldsymbol{x_i}), \quad (6)$$

where $\|\cdot\|_2$ denotes ℓ_2-norm of a vector, ϵ_{train} represents the adversarial budget used in the substitute training stage.

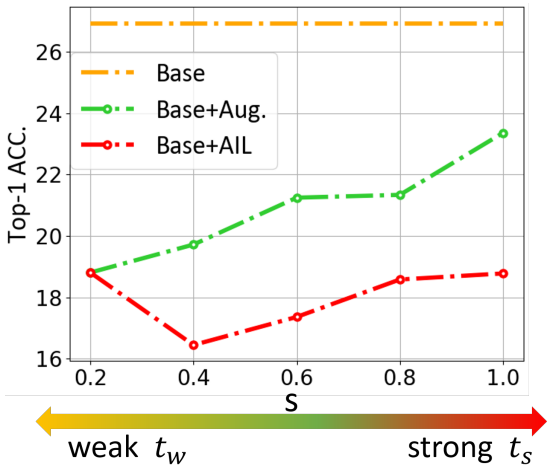


Figure 3: Effectiveness of the proposed AIL. The training data of the substitute model consists of 4k samples randomly drawn from the COCO (40k) dataset.

Adversarial Invariant Learning

In the absence of annotated data, invariant learning (Hadsell, Chopra, and LeCun 2006), which aims to maintain the invariance of the representations against various input transformations, has been confirmed to be effective in enhancing the model generalization in many downstream tasks (Jaiswal et al. 2020; Foster, Pukdee, and Rainforth 2021). This inspires us to explore whether maintaining the transformation invariance of adversarial examples can promote substitute training. We start with exploring the effects of performing input transformations to the samples with slight adversarial perturbations in substitute training.

Firstly, we apply input transformations with various strengths to the samples with slight adversarial perturbations to train the substitute model by minimizing Eq. (5) (the strategies for controlling the quantization strengths of the input transformations are consistent with that in (Jang et al. 2022)). The impacts of different transformation strengths on substitute training are presented in Fig. 3. According to the green curve in Fig. 3, the transformed adversarial samples can indeed improve substitute training compared to the baseline, i.e., no transformations are applied to the samples with slight adversarial perturbations. However, as the strength of transformations increases, the gain to the baseline becomes less significant. We attribute this to the extreme transformations, i.e., they may cause difficulties in convergence. To alleviate this problem, for substitute training, we apply weak transformations to the samples with slight adversarial perturbations, and additionally regulate the representations with the strong transformations to be invariant to the representations with the weak transformations. As shown in Fig. 3, the red curve reveals that the attack performance of this invariant regularization against the transformations with different strengths is obviously better than directly applying strong transformations. By assigning different weight to all the transformations with different strengths, the unified in-

variant regularization can be formulated in a cost-sensitive manner, as

$$\mathcal{L}_{\mathrm{AIL}}(\boldsymbol{\phi} \circ \boldsymbol{\theta}; \boldsymbol{X}) := \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{s \sim U(0,1)} \{ \mathcal{C}(s) \cdot (1 - cos(\boldsymbol{z_i^{a,s}}, \boldsymbol{z_i^{a,s_0}})) \}, \tag{7}$$

where the hyper-parameter s denotes the random variable controlling the strength of input transformation, and s_0 stands for the strength of weak transformation. It holds $\boldsymbol{z_i^{a,s}} = g_{\boldsymbol{\phi}} \circ f_{\boldsymbol{\theta}}(t_s(\boldsymbol{x_i} + \boldsymbol{\delta_i}))$, $\boldsymbol{z_i^{a,s_0}} = g_{\boldsymbol{\phi}} \circ f_{\boldsymbol{\theta}}(t_w(\boldsymbol{x_i} + \boldsymbol{\delta_i}))$, $t_w \leftrightarrow t_{s_0}, t_s \sim \mathcal{T}^*$, where $\mathcal{T}^*$ denotes the set of transformations for adversarial examples. For t_s, when $s = 0$, no transformation is applied, i.e., the original input is utilized. When $s = 1$, the strongest transformations are applied. When $s \in (0, 1)$, we multiply s by the probability that the augmentation happens, to manipulate t_s. $\mathcal{C}(s)$ is the cost-sensitive function of s and it is defined as

$$\mathcal{C}(s) = \exp\left(-\frac{(s - s_0)^2}{2\sigma^2} \right), \tag{8}$$

where we set s_0=0.2 and σ=1 in this paper.

In general, the overall objective of our substitute training is formulated as

$$\min_{\boldsymbol{\theta}, \boldsymbol{\phi}} \mathcal{L}_{\mathrm{ACCL}}(\boldsymbol{\phi} \circ \boldsymbol{\theta}; \boldsymbol{X}) + \lambda \cdot \mathcal{L}_{\mathrm{AIL}}(\boldsymbol{\phi} \circ \boldsymbol{\theta}; \boldsymbol{X}), \tag{9}$$

where $\mathcal{L}_{\mathrm{ACCL}}$ is computed via Eq. (5), and λ is a hyper-parameter balancing the roles of $\mathcal{L}_{\mathrm{ACCL}}$ and $\mathcal{L}_{\mathrm{AIL}}$.

Cross-Paradigm Transfer-based Attack

Once the substitute model is trained, transfer-based attack can be launched. However, in the absence of label information, the common schemes, such as maximizing cross-entropy loss to generate adversarial examples, are not applicable. The most feasible approach is perturbing the intermediate features of the trained substitute model. In particular, for an input image $\boldsymbol{I}$, its adversarial example can be generated via

$$\underset{\boldsymbol{\delta}}{\operatorname{argmin}}\, cos\left(f_{\boldsymbol{\theta}}^{l}(\boldsymbol{I} + \boldsymbol{\delta}), f_{\boldsymbol{\theta}}^{l}(\boldsymbol{I}) \right), s.t, \|\boldsymbol{\delta}\|_\infty \leq \epsilon_{test}, \tag{10}$$

where $f_{\boldsymbol{\theta}}^{l}(\cdot)$ denotes the intermediate feature of the l-th layer in the substitute model $f_{\boldsymbol{\theta}}$, $\boldsymbol{\delta}$ stands for the adversarial perturbation w.r.t. the image $\boldsymbol{I}$, $\|\cdot\|_\infty$ represents the ℓ_∞-norm of a vector and ϵ_{test} is the adversarial budget used in the attack stage. In this paper, MI-FGSM (Dong et al. 2018) is employed to solve Eq. (10) to obtain the final adversarial example $\boldsymbol{I} + \boldsymbol{\delta}$.

Experiments

Experimental Settings

Dataset. We train our substitute models on three unlabeled datasets, i.e., COCO (40k samples)(Lin et al. 2014), Comics (50k samples)(Cenk Bircanoglu 2017) and Paintings (79k samples)(Painter by Number 2017), respectively. For evaluation, we draw 5k images from the validation set of ImageNet (Russakovsky et al. 2015). The training and evaluation protocols are identical to the compared methods, No-box (Li, Guo, and Chen 2020) and APR (Malik et al. 2022).

Models Architectures. Our substitute model adopts the architecture of ResNet-50 (He et al. 2016a) without pre-training. Two types of architectures, i.e., Convolution Neural Networks (CNNs) and Vision Transformers (ViTs), are employed as the target models. Specifically, VGG-19 (Simonyan and Zisserman 2015), Inception-v3 (Szegedy et al. 2016), ResNet-152 (He et al. 2016a), Dense-121 (Huang et al. 2017), SeNet (Hu, Shen, and Sun 2018), Wide-ResNet-50 (Zagoruyko and Komodakis 2016) and MobileNet-V2 (Sandler et al. 2018) are selected as the CNN-type of target models. ViT-T, ViT-S (Dosovitskiy et al. 2021), DeiT-T, DeiT-S (Touvron et al. 2021) are selected as the ViT-type of target models. All the target models are pretrained on the ImageNet training set.

Substitute Training & Attack Settings. Our substitute model is randomly initialized. We train it via classical SGD algorithm with a fixed learning rate of 0.1. The batch size is set to 64. The weight decay is set to 1e-4. We set ϵ_{train} in Eq. (6) as 1.0 and λ in Eq. (9) as 0.1. The total training epoch is set to 100. The number of iterations for Eq. (6) is set to 1. For the training of one AGS model, about 17 hours are demanded on a single RTX 3090 GPU, which is an affordable cost compared to training a pretrained ResNet50 model. After the substitute model is trained, we conduct transfer-based attack, where ϵ_{test} is set to 0.1, the step size is set to 1/255 and the number of iterations is set to 300. The intermediate layer l selected in Eq. (10) is '$layer2$'. Top-1 (%) accuracy is selected as the evaluation metric.

Main Results

In this section, we conduct experiments to attack various target models, including normally trained CNNs and ViTs, robustly trained CNNs, as well as object detection and segmentation models, based on our substitute model. We compare our AGS with the closely related state-of-the-art methods, i.e., No-box (Li, Guo, and Chen 2020) and APR (Malik et al. 2022). It is noteworthy that we consider two variants of their methods, 'Rotation' mode and 'Jigsaw' mode, abbreviated as 'R' and 'J' in the rest of this paper. Additionally, we also compare AGS with two unsupervised baselines, i.e., training the substitute model via classical contrastive learning (denoted as "CL") and its adversarial version (denoted as "CL+adv", where $\boldsymbol{z_i^1}, \boldsymbol{z_i^2}$ in Eq. (2) are replaced with the projection vectors of the adversarial examples). Furthermore, we evaluate the performance of AGS when the substitute model is a ResNet50 (He et al. 2016a) pretrained on the entire ImageNet training set (Since the ImageNet training set lies in the target domain, intuitively, a pretrained ResNet50 is more likely to achieve a better performance than our method). For all the compared schemes, the strategy for generating adversarial examples remains consistent, i.e., Eq. (10). The sole distinctions lie in the substitute models employed.

Results on CNNs and ViTs. The attack performance on CNNs and ViTs across three unlabeled datasets is presented in Tab. 1, Tab. 2 and Tab. 3, respectively. Four key observations can be drawn from these results. Firstly, our baseline schemes, i.e., "CL" and "CL+adv", outperform existing closely related methods (No-box and APR) in most cases.

Models	CNNs								ViTs				
	VGG-19	Inc-v3	Res-152	Dense-121	SeNet	WRN	MNet-V2	Avg.	Deit-T	Deit-S	ViT-T	ViT-S	Avg.
ResNet50	0.72	4.62	0.84	0.82	2.36	0.84	0.60	1.54	24.52	36.38	12.32	37.78	27.75
No-box (R)	23.20	31.86	37.32	23.90	33.94	34.34	15.44	28.57	29.46	48.78	27.28	49.16	38.67
No-box (J)	30.18	49.30	53.98	44.66	59.48	53.58	16.80	43.40	49.68	65.22	44.74	65.20	56.21
APR (R)	17.02	27.70	26.92	22.80	34.30	21.72	11.82	23.18	34.72	49.32	14.78	41.22	35.01
APR (J)	19.74	37.82	39.10	29.92	42.94	36.42	13.16	31.30	41.72	58.42	25.18	57.28	45.65
CL	13.30	13.52	6.30	6.72	20.42	8.88	2.50	10.23	30.26	46.16	17.70	48.80	35.73
CL+adv	6.12	5.30	7.36	5.90	9.88	5.02	3.32	6.13	19.88	34.56	5.98	26.12	21.64
OCCL (ours)	6.14	5.54	5.16	3.72	11.60	3.58	1.64	5.34	26.04	42.30	10.80	45.44	31.15
ACCL (ours)	1.90	2.88	3.68	2.52	**5.12**	2.26	**0.90**	2.75	18.10	30.52	3.20	27.28	19.78
AGS (ACCL+AIL)	**1.88**	**2.08**	**2.02**	**1.72**	5.24	**1.50**	1.14	**2.23**	**14.70**	**26.50**	**2.96**	**22.66**	**16.71**

Table 1: Top-1 (%) accuracy of CNNs and ViTs on 5k adversarial examples with $\epsilon_{test} \leq 0.1$, from ImageNet validation set. The substitute models except ResNet50 are trained on the unlabeled COCO (40k) dataset (the lower, the better).

Models	CNNs								ViTs				
	VGG-19	Inc-v3	Res-152	Dense-121	SeNet	WRN	MNet-V2	Avg.	Deit-T	Deit-S	ViT-T	ViT-S	Avg.
ResNet50	0.72	4.62	0.84	0.82	2.36	0.84	0.60	1.54	24.52	36.38	12.32	37.78	27.75
No-box (R)	49.64	59.12	66.10	58.92	70.88	66.46	36.50	58.23	61.32	74.86	60.06	73.44	67.42
No-box (J)	61.20	67.28	72.96	69.16	78.90	73.90	50.66	67.72	68.64	79.54	70.80	78.50	74.37
APR (R)	15.30	28.00	30.38	25.04	32.56	27.24	10.26	24.11	39.88	55.00	20.78	45.92	40.39
APR (J)	30.78	47.40	48.58	41.68	53.48	49.02	18.90	41.41	50.34	65.82	39.62	63.52	54.83
CL	20.44	24.82	20.48	13.08	30.92	17.56	5.56	18.98	37.70	51.76	25.10	52.54	41.78
CL+adv	5.94	**9.80**	9.82	6.60	15.14	6.24	2.54	8.01	28.66	43.98	**6.90**	**33.20**	28.19
OCCL (ours)	19.60	25.80	18.84	11.60	34.36	15.60	4.36	18.59	39.12	49.42	25.32	53.02	41.72
ACCL (ours)	6.56	19.40	15.32	6.98	18.20	13.76	2.06	11.75	30.04	43.72	16.32	40.48	32.64
AGS (ACCL+AIL)	**3.54**	11.08	**7.30**	**3.94**	**12.70**	**5.06**	**1.40**	**6.43**	**23.76**	**34.90**	7.64	34.60	**25.23**

Table 2: Top-1 (%) accuracy of CNNs and ViTs on 5k adversarial examples, under $\epsilon_{test} \leq 0.1$, from ImageNet validation set. The substitute models except ResNet50 are trained on unlabeled Comics (50k) dataset (the lower, the better).

Models	CNNs								ViTs				
	VGG-19	Inc-v3	Res-152	Dense-121	SeNet	WRN	MNet-V2	Avg.	Deit-T	Deit-S	ViT-T	ViT-S	Avg.
ResNet50	0.72	4.62	0.84	0.82	2.36	0.84	0.60	1.54	24.52	36.38	12.32	37.78	27.75
No-box (R)	20.66	27.12	37.30	24.62	39.44	33.94	11.60	27.81	29.62	44.64	26.88	46.20	36.84
No-box (J)	27.62	48.86	51.98	45.60	59.00	52.48	21.52	43.87	48.76	65.12	43.32	62.54	54.94
APR (R)	13.04	18.58	21.26	15.66	30.40	17.80	7.36	17.73	26.08	42.70	11.14	37.02	29.24
APR (J)	22.86	36.30	42.58	29.98	46.42	39.66	16.36	33.45	43.44	61.32	28.88	58.00	47.91
CL	18.70	20.38	13.36	12.50	28.84	18.30	7.86	17.13	38.32	54.34	32.46	57.46	45.65
CL+adv	8.00	7.50	9.44	7.84	15.22	6.30	4.30	8.37	22.08	36.74	8.74	29.54	24.28
OCCL (ours)	12.58	16.28	8.16	8.72	24.76	9.56	4.88	12.13	35.66	52.14	24.82	53.04	41.42
ACCL (ours)	4.12	4.26	4.14	2.92	7.06	3.50	1.82	3.97	17.52	30.58	5.90	28.80	20.70
AGS (ACCL+AIL)	**3.10**	**3.06**	**3.10**	**2.54**	**5.34**	**1.86**	**1.62**	**2.95**	**15.60**	**27.44**	**4.76**	**24.50**	**18.08**

Table 3: Top-1 (%) accuracy of CNNs and ViTs on 5k adversarial examples, under $\epsilon_{test} \leq 0.1$, from ImageNet validation set. The substitute models except ResNet50 are trained on unlabeled Paintings (79k) dataset (the lower, the better).

This could be attributed to the more advanced substitute training mechanisms, where contrastive learning is superior to the auto-encoder approach adopted by No-box and APR in this task. Secondly, "CL+adv" demonstrates better performance than "CL", especially on ViTs, indicating the adversarial transferability of substitute model benefits from slight adversarial training. Thirdly, taking COCO dataset as an example, our method "ACCL+AIL" outperforms the strongest baseline "CL+adv" with a margin of 3.9% and 4.9% on CNNs and ViTs, respectively. Our superiority also exists and it is even more significant on the other two datasets, showcasing the effectiveness of our proposed "adversary-centric" training strategy. Fourthly, we can observe that the attack performance of our substitute model on CNNs approximates the supervised baseline, i.e., the pretrained ResNet50 (2.23 vs. 1.54). Meanwhile, our model surpasses ResNet50 on ViTs with a large margin, i.e., 16.72 vs. 27.75. These phenomenons preliminarily indicates that the models pretrained on large-scale annotated datasets are not necessarily optimal for transfer-based adversarial attacks.

Results on Robustly Trained CNNs. Here, we evaluate our method against the robustly trained CNNs, which are

Models	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens}$
Original ACC.	78.12	78.18	81.22
ResNet50	55.36	57.48	63.74
No-box(J)	60.36	62.56	66.74
No-box(R)	44.36	50.32	54.64
APR(J)	45.78	47.08	52.62
APR(R)	34.10	32.70	40.02
AGS (Paintings)	23.96	29.80	36.14
AGS (Comics)	47.00	47.96	54.98
AGS (COCO)	**12.74**	**17.64**	**25.44**

Table 4: Top-1 (%) accuracy of robustly trained CNNs on 5k adversarial examples with $\epsilon_{test} \leq 0.1$, which are generated from the ImageNet validation set.

more challenging to be attacked than the normally trained CNNs. Three typical robust models, i.e., Inc-v3$_{ens3}$, Inc-v3$_{ens4}$ and IncRes-v2$_{ens}$ (Tramr et al. 2018) are selected as the target models. The results are shown in Tab. 4. As observed, the adversarial examples generated by the pretrained ResNet50 only induce a marginal drop in original accuracy, which implies the challenge of this scenario. On the contrary, adversarial examples generated by AGS can lead to significantly lower accuracies. For instance, AGS achieves an improvement of over 40% from the pretrained ResNet50 on the COCO (40k) dataset. The primary reason is that our substitute training method facilitates comprehensive interaction between adversarial and benign examples, allowing the capture of robust features, which is beneficial for effectively attacking the robustly trained models.

Results on Object Detection and Segmentation Models. To validate the generalizability of our method, we further conduct adversarial attack on object detection and video segmentation models. Following (Malik et al. 2022), we select two representative models, DETR (Carion et al. 2020) for object detection, and DINO (Caron et al. 2021) for video segmentation. The datasets for evaluation are the MS COCO validation set (Lin et al. 2014) and the DAVIS validation set (Pont-Tuset et al. 2017), respectively. The evaluation metrics include mean average precision (mAP), mean average recall (mAR) and the Jacard index metrics, $(J\&F)_m$ and $(J\&F)_r$. The results are shown in Tabs. 5 and 6. Two conclusions can be obtained: 1) AGS outperforms the best compared method by an average margin of 9.3% and 6.5%, indicating its superior adversarial transferability across various vision tasks; 2) AGS also performs comparably with pretrained ResNet50 in most of the cases. Notably, the pretrained ResNet50 performs slightly better than ours in terms of mAP on the detection task (2.6 vs. 2.8). This is due to the backbone of the DETR model is exactly the pretrained ResNet50 model. In contrast, AGS, requiring no annotated data in the training process, demonstrates its suitability for transfer-based adversarial attacks compared to pretrained ResNet50.

Ablation Study

In this section, we assess the effectiveness of each proposed scheme, i.e., ACCL and AIL, in our substitute training framework. Firstly, to observe the influence of "adversary-centric" in ACCL, we replace the adversary $\boldsymbol{z_i^a}$ in Eq. (5) with its original example $\boldsymbol{z_i}$ to train the substitute model, which is denoted as "OCCL". Then, to validate the effectiveness of AIL, we train the substitute model only with Eq. (5), which is denoted as "ACCL". At last, we denote the proposed substitute training framework AGS as "ACCL+AIL".

Models	COCO		Paintings		Comics	
	mAP	mAR	mAP	mAR	mAP	mAR
No-box (R)	19.3	-	17.2	-	34.3	-
No-box (J)	24.7	-	24.1	-	38.0	-
APR (R)	14.6	-	11.9	-	13.3	-
APR (J)	14.5	-	14.0	-	20.8	-
ResNet50	**2.6**	7.4	**2.6**	**7.4**	**2.6**	**7.4**
AGS (ours)	2.8	**6.9**	3.8	9.5	5.2	12.1

Table 5: Evaluations on object detection model. The target detector is DETR with the backbone of Resnet50. The mean average precision (mAP) and recall (mAR) at [0.5:0.95] on MS COCO validation set are reported.

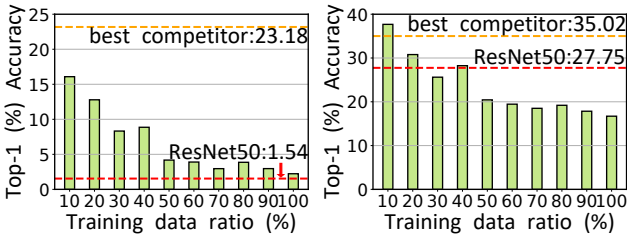
Models	COCO		Paintings		Comics	
	$(J\&F)_m$	$(J\&F)_r$	$(J\&F)_m$	$(J\&F)_r$	$(J\&F)_m$	$(J\&F)_r$
No-box (R)	53.2	-	52.6	-	57.8	-
No-box (J)	53.9	-	53.2	-	58.3	-
APR (R)	48.9	-	46.9	-	47.8	-
APR (J)	46.6	-	48.5	-	51.7	-
ResNet50	45.3	47.9	45.3	47.9	45.3	47.9
AGS (ours)	**41.8**	**42.6**	**43.9**	**45.8**	**36.1**	**31.5**

Table 6: Evaluations on video segmentation model. The target model is DINO with the backbone of ViT-S. Two Jacard index metrics on the DAVIS validation set are reported.

As shown in the last three lines of Tab. 1, Tab. 2 and Tab. 3, "ACCL+AIL" performs better than "OCCL" and "ACCL". Taking COCO (40k) as an example, "ACCL" improves "OCCL" by a margin of 2.59% and 11.37% on CNNs and ViTs, respectively. It indicates that our "adversary-centric" mechanism is more effective than the "original example-centric" mechanism, especially on ViTs, which have larger architecture gaps to the source model. Besides, with the help of AIL, "ACCL+AIL" (AGS) further improves "ACCL" by a margin of 0.52% and 3.07% on CNNs and ViTs, respectively. As can be observed from Fig. 5(b) and Fig. 5(c), AIL makes the loss landscape of substitute model flatter than the baseline, which is probably the reason of AGS's superiority.

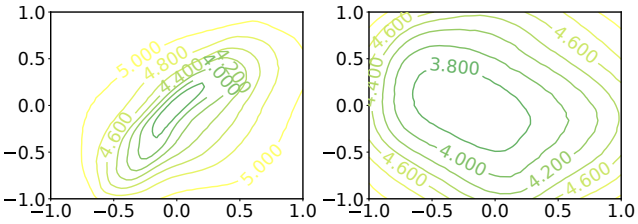
More Analysis

The Scale of Training Data. One of the aims of this paper is to train the substitute model only with the unlabeled training data as little as possible, i.e., striving for affordability. To this end, we explore the performance change trend of our AGS when reducing the training data. Taking COCO (40K) as an example, we firstly utilize it entirely as the training set. Then, we divide it into different parts to train our substitute model. The averaged Top-1 (%) accuracies on CNNs



(a) CNNs (b) ViTs

Figure 4: The attack performances with different scales o training data for the substitute model, where the training dat is from the COCO (40k) dataset. Averaged Top-1 (%) accu racy on (a) CNNs and (b) ViTs are presented.



(a) OCCL (b) ACCL+AIL

Figure 5: Visualization of loss landscape geometry.

and ViTs are shown in Figs. 4(a) and 4(b), respectively. For CNNs, our method outperforms previous best competitor, APR (R), by a margin of more than 5%, only with 10% of COCO (40K). Besides, as shown in Fig. 4(b), our model achieves better performance than the pretrained ResNet50 model against ViTs only with 50% of COCO (40K). These results validate the affordability of our AGS framework.

Loss Landscape Geometry. Here, we try to explain the superior adversarial transferability of our substitute model from the perspective of loss landscape geometry, which highly affects the generalizability of the model (Li et al. 2018). Specifically, the loss surface of Eq. (5) is visualized via the technique in (Li et al. 2018). As can be observed from Fig. 5(a) to Fig. 5(b), the distance between the contours gradually increases and our proposed AGS further widens these spaces, which indicates that a flatter minimum is found and the generalizability of the model is improved.

Visualizations. In this section, some qualitative explanations of our AGS is given via GradCAM (Selvaraju et al. 2017). As shown in Fig. 6, the adversarial examples generated by our method can confuse the attentions of the target model, which leads to wrong predictions.

Conclusion

In this paper, we bring a new insight that the models pretrained on large-scale annotated datasets, e.g., ResNet50 pretrained on ImageNet, are not always appropriate for transfer-based attacks. Without any annotated data, we develop a substitute training framework for transfer-based attacks, named AGS, with which the affordability and generalizability of the substitute model are simultaneously

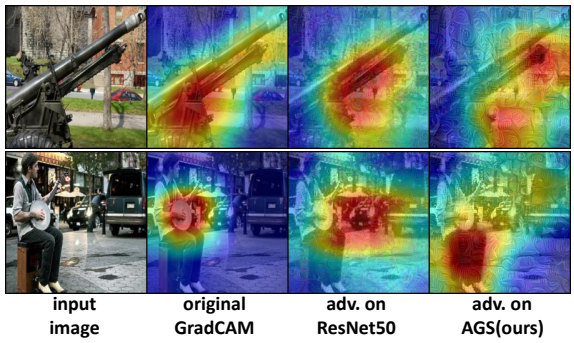


Figure 6: The illustration of attention shift. Adversarial examples are generated on two types of substitute models, i.e., pretrained ResNet50 and our proposed AGS, respectively, with $\epsilon_{test} \leq 0.1$. All the activation maps are generated on ImageNet pretrained ResNet152 model.

achieved. Comprehensive experiments demonstrate the superiority of our proposed work. More importantly, we hope our explorations can inspire the future work on rethinking and developing more suitable substitute training methods for transfer-based black-box attacks.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62272020 and U20B2069, in part by the National Social Science Fund of China under Grant 22VMG037, in part by the State Key Laboratory of Software Development Environment under Grant SKLSDE2023ZX-16, and in part by the Fundamental Research Funds for Central Universities.

References

Aich, A.; Khang-Ta, C.; Gupta, A.; Song, C.; Krishnamurthy, S. V.; Asif, M. S.; and Roy-Chowdhury, A. K. 2022. GAMA: Generative Adversarial Multi-Object Scene Attacks. In *Advances in Neural Information Processing Systems*, 36914–36930.

Ban, Y.; and Dong, Y. 2022. Pre-trained Adversarial Perturbations. In *Advances in Neural Information Processing Systems*, 1196–1209.

Byun, J.; Cho, S.; Kwon, M.-J.; Kim, H.-S.; and Kim, C. 2022. Improving the Transferability of Targeted Adversarial Examples through Object-based Diverse Input. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 15244–15253.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-End Object Detection with Transformers. In *Proceedings of the European Conference on Computer Vision*, 213–229.

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging Properties in Self-Supervised Vision Transformers. In *Proceedings of the IEEE International Conference on Computer Vision*, 9650–9660.

Cenk Bircanoglu. 2017. Comic Books Classification Dataset. https://www.kaggle.com/datasets/cenkbircanoglu/comic-books-classification.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A Simple Framework for Contrastive Learning of Visual Representations. In *International Conference on Machine Learning*, 1597–1607.

Chen, X.; and He, K. 2021. Exploring Simple Siamese Representation Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 15750–15758.

Dong, Y.; Liao, F.; Pang, T.; Su, H.; Zhu, J.; Hu, X.; and Li, J. 2018. Boosting Adversarial Attacks with Momentum. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 9185–9193.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

Foster, A.; Pukdee, R.; and Rainforth, T. 2021. Improving Transformation Invariance in Contrastive Representation Learning. In *International Conference on Learning Representations*.

Gubri, M.; Cordy, M.; Papadakis, M.; Traon, Y. L.; and Sen, K. 2022. Lgv: Boosting adversarial example transferability from large geometric vicinity. In *Proceedings of the European Conference on Computer Vision*, 603–618.

Hadsell, R.; Chopra, S.; and LeCun, Y. 2006. Dimensionality Reduction by Learning an Invariant Mapping. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1735–1742.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum Contrast for Unsupervised Visual Representation Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 9729–9738.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016a. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016b. Identity Mappings in Deep Residual Networks. In *Proceedings of the European Conference on Computer Vision*, 630–645.

Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-Excitation Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7132–7141.

Huang, G.; Liu, Z.; Van Der Maaten, L.; and Weinberger, K. Q. 2017. Densely Connected Convolutional Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4700–4708.

Huang, Q.; Katsman, I.; He, H.; Gu, Z.; Belongie, S.; and Lim, S.-N. 2019. Enhancing adversarial Example Transferability with an Intermediate Level Attack. In *Proceedings of the IEEE International Conference on Computer Vision*, 4733–4742.

Huang, Y.; and Kong, A. W.-K. 2022. Transferable Adversarial Attack Based on Integrated Gradients. In *International Conference on Learning Representations*.

Inkawhich, N.; Liang, K. J.; Wang, B.; Inkawhich, M.; Carin, L.; and Chen, Y. 2020. Perturbing Across the Feature Hierarchy to Improve Standard and Strict Black-box Attack Transferability. In *Advances in Neural Information Processing Systems*, 20791–20801.

Jaiswal, A.; Moyer, D.; Ver Steeg, G.; AbdAlmageed, W.; and Natarajan, P. 2020. Invariant Representations through Adversarial Forgetting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 4272–4279.

Jang, Y. K.; Gu, G.; Ko, B.; Kang, I.; and Cho, N. I. 2022. Deep Hash Distillation for Image Retrieval. In *Proceedings of the European Conference on Computer Vision*, 354–371.

Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; and Krishnan, D. 2020. Supervised Contrastive Learning. In *Advances in Neural Information Processing Systems*, 18661–18673.

Kumar, C.; Ramesh, J.; Chakraborty, B.; Raman, R.; Weinrich, C.; Mundhada, A.; Jain, A.; and Flohr, F. B. 2021. Vru Pose-SSD: Multi-Person Pose Estimation for Automated Driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 15331–15338.

Li, H.; Xu, Z.; Taylor, G.; Studer, C.; and Goldstein, T. 2018. Visualizing the loss landscape of neural nets. In *Advances in Neural Information Processing Systems*, 6391–6401.

Li, Q.; Guo, Y.; and Chen, H. 2020. Practical No-box Adversarial Attacks against DNNs. In *Advances in Neural Information Processing Systems*, 12849–12860.

Li, Z.; Wu, W.; Su, Y.; Zheng, Z.; and Lyu, M. R. 2023. CDTA: A Cross-Domain Transfer-Based Attack with Contrastive Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 1530–1538.

Lin, J.; Song, C.; He, K.; Wang, L.; and Hopcroft, J. E. 2020. Nesterov Accelerated Gradient and Scale Invariance for Adversarial Attacks. In *International Conference on Learning Representations*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft Coco: Common Objects in Context. In *Proceedings of the European Conference on Computer Vision*, 740–755.

Malik, H. S.; Kunhimon, S. K.; Naseer, M.; Khan, S.; and Khan, F. S. 2022. Adversarial Pixel Restoration as a Pretext Task for Transferable Perturbations. In *British Machine Vision Conference*, 546–561.

Naseer, M.; Khan, S.; Hayat, M.; Khan, F. S.; and Porikli, F. 2021. On Generating Transferable Targeted Perturbations. In *Proceedings of the IEEE International Conference on Computer Vision*, 7708–7717.

Naseer, M. M.; Khan, S. H.; Khan, M. H.; Shahbaz Khan, F.; and Porikli, F. 2019. Cross-domain Transferability of Adversarial Perturbations. In *Advances in Neural Information Processing Systems*, 12885–12895.

Painter by Number. 2017. https://www.kaggle.com/c/painter-by-numbers/data.

Pont-Tuset, J.; Perazzi, F.; Caelles, S.; Arbeláez, P.; Sorkine-Hornung, A.; and Van Gool, L. 2017. The 2017 Davis Challenge on Video Object Segmentation. In *arXiv preprint arXiv:1704.00675*.

Qin, Z.; Fan, Y.; Liu, Y.; Shen, L.; Zhang, Y.; Wang, J.; and Wu, B. 2022. Boosting the transferability of adversarial attacks with reverse adversarial perturbation. In *Advances in Neural Information Processing Systems*, 29845–29858.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster R-cnn: Towards Real-time Object detection with Region Proposal Networks. In *Advances in Neural Information Processing Systems*, 91–99.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; and Bernstein, M. 2015. ImageNet Large Scale Visual Recognition Challenge. In *International Journal of Computer Vision*, 211–252.

Salzmann, M.; et al. 2021. Learning Transferable Adversarial Perturbations. In *Advances in Neural Information Processing Systems*, 13950–13962.

Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; and Chen, L.-C. 2018. Mobilenetv2: Inverted Residuals and Linear Bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4510–4520.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual Explanations from Deep Networks via Gradient-based Localization. In *Proceedings of the IEEE International Conference on Computer Vision*, 618–626.

Simonyan, K.; and Zisserman, A. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *International Conference on Learning Representations*.

Springer, J.; Mitchell, M.; and Kenyon, G. 2021. A Little Robustness Goes a Long Way: Leveraging Robust Features for Targeted Transfer Attacks. In *Advances in Neural Information Processing Systems*, 9759–9773.

Sun, C.; Zhang, Y.; Chaoqun, W.; Wang, Q.; Li, Y.; Liu, T.; Han, B.; and Tian, X. 2022. Towards Lightweight Black-Box Attacks against Deep Neural Networks. In *Advances in Neural Information Processing Systems*, 19319–19331.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the Inception Architecture for Computer Vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2818–2826.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; and Fergus, R. 2014. Intriguing Properties of Neural Networks. In *International Conference on Learning Representations*.

Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; and Jégou, H. 2021. Training Data-Efficient Image Transformers & Distillation through Attention. In *International Conference on Machine Learning*, 10347–10357.

Tramr, F.; Kurakin, A.; Papernot, N.; Goodfellow, I.; Boneh, D.; and McDaniel, P. 2018. Ensemble Adversarial Training: Attacks and Defenses. In *International Conference on Learning Representations*.

Vakhshiteh, F.; Ramachandra, R.; and Nickabadi, A. 2020. Threat of Adversarial Attacks on Face Recognition: A Comprehensive Survey. In *arXiv preprint arXiv:2007.11709*.

Wang, X.; and He, K. 2021. Enhancing the Transferability of Adversarial Attacks through Variance Tuning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1924–1933.

Wang, X.; He, X.; Wang, J.; and He, K. 2021a. Admix: Enhancing the Transferability of Adversarial Attacks. In *Proceedings of the IEEE International Conference on Computer Vision*, 16158–16167.

Wang, Y.; Wang, J.; Yin, Z.; Gong, R.; Wang, J.; Liu, A.; and Liu, X. 2022. Generating Transferable Adversarial Examples against Vision Transformers. In *Proceedings of the 30th ACM International Conference on Multimedia*, 5181–5190.

Wang, Z.; Guo, H.; Zhang, Z.; Liu, W.; Qin, Z.; and Ren, K. 2021b. Feature Importance-Aware Transferable Adversarial Attacks. In *Proceedings of the IEEE International Conference on Computer Vision*, 7639–7648.

Wu, D.; Wang, Y.; Xia, S.-T.; Bailey, J.; and Ma, X. 2019. Skip Connections Matter: On the Transferability of Adversarial Examples Generated with ResNets. In *International Conference on Learning Representations*.

Wu, L.; Zhu, Z.; Tai, C.; et al. 2018. Understanding and enhancing the transferability of adversarial examples. In *arXiv preprint arXiv:1802.09707*.

Wu, W.; Su, Y.; Chen, X.; Zhao, S.; King, I.; Lyu, M. R.; and Tai, Y.-W. 2020. Boosting the Transferability of Adversarial Samples via Attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1161–1170.

Xie, C.; Zhang, Z.; Zhou, Y.; Bai, S.; Wang, J.; Ren, Z.; and Yuille, A. L. 2019. Improving Transferability of Adversarial Examples with Input Diversity. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2730–2739.

Yang, R.; Guo, Y.; Wang, R.; Zhao, X.; and Wang, Y. 2022. Exploring the Impact of Adding Adversarial Perturbation onto Different Image Regions. In *IEEE International Symposium on Circuits and Systems*, 2363–2367.

Zagoruyko, S.; and Komodakis, N. 2016. Wide Residual Networks. In *British Machine Vision Conference*.

Zhang, J.; Wu, W.; Huang, J.-t.; Huang, Y.; Wang, W.; Su, Y.; and Lyu, M. R. 2022a. Improving Adversarial Transferability via Neuron Attribution-based Attacks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 14993–15002.

Zhang, Q.; Li, X.; Chen, Y.; Song, J.; Gao, L.; He, Y.; and Xue, H. 2022b. Beyond ImageNet Attack: Towards Crafting Adversarial Examples for Black-box Domains. In *International Conference on Learning Representations*.