Boosting Adversarial Transferability across Model Genus by Deformation-Constrained Warping

Qinliang Lin[1*], Cheng Luo[1*], Zenghao Niu[1], Xilin He[1], Weicheng Xie[1, 2, 3†]
Yuanbo Hou[4], Linlin Shen[1, 2, 3], Siyang Song[5]

[1]Computer Vision Institute, School of Computer Science & Software Engineering, Shenzhen University
[2]Shenzhen Institute of Artificial Intelligence and Robotics for Society
[3]Guangdong Key Laboratory of Intelligent Information Processing
[4]WAVES Research Group, Ghent University, Belgium
[5]University of Leicester, UK.
2017192020@email.szu.edu.cn, wcxie@szu.edu.cn

Abstract

Adversarial examples generated by a surrogate model typically exhibit limited transferability to unknown target systems. To address this problem, many transferability enhancement approaches (*e.g.,* input transformation and model augmentation) have been proposed. However, they show poor performances in attacking systems having different model genera from the surrogate model. In this paper, we propose a novel and generic attacking strategy, called Deformation-Constrained Warping Attack (DeCoWA), that can be effectively applied to cross model genus attack. Specifically, DeCoWA firstly augments input examples via an elastic deformation, namely Deformation-Constrained Warping (DeCoW), to obtain rich local details of the augmented input. To avoid severe distortion of global semantics led by random deformation, DeCoW further constrains the strength and direction of the warping transformation by a novel adaptive control strategy. Extensive experiments demonstrate that the transferable examples crafted by our DeCoWA on CNN surrogates can significantly hinder the performance of Transformers (and vice versa) on various tasks, including image classification, video action recognition, and audio recognition. Code is made available at https://github.com/LinQinLiang/DeCoWA.

Introduction

In the past decade, various deep network architectures, including Convolutional Neural Networks (CNNs) (He et al. 2016), LSTMs (Hochreiter and Schmidhuber 1997), Transformers (Dosovitskiy et al. 2021), etc., have demonstrated revolutionary performances in recognition or classification of real-world signals, including images (He et al. 2016), videos (Bertasius, Wang, and Torresani 2021), audio (Hwang et al. 2022), etc. Nevertheless, almost all these networks are vulnerable to adversarial examples crafted by adversaries, potentially leading to severe security threats. To this end, investigating generic features/vulnerabilities among different network architectures is crucial and urgent.

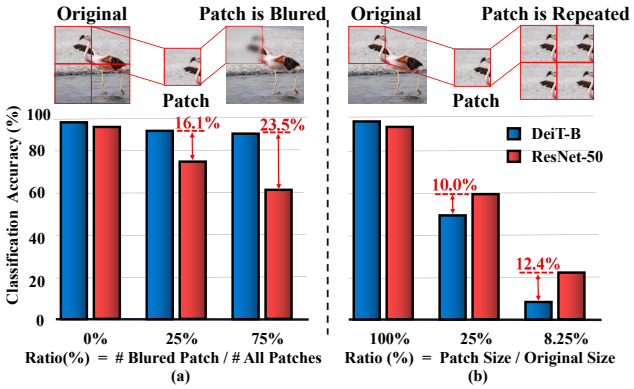

Figure 1: Illustration of distinctions between CNNs and ViTs. We test recognition accuracy by two different model genera using (a) local detail blurred images and (b) global structure damaged images, respectively. (a) As more local details are blurred, the performance of ResNet-50 significantly drops while DeiT-B is still robust. (b) When an image patch of a smaller size remains, ResNet-50 achieves higher classification accuracy than DeiT-B. # denotes counting the number.

Some approaches aim to generate adversarial examples with transferability to fool target systems in black-box scenarios, where network details of the target model are unknown. Most of the relevant research (Dong et al. 2018; Lin et al. 2020; Dong et al. 2019; Wang and He 2021) lies on adversarial transferability across CNN architectures. Meanwhile, a small part of recent studies (Wei et al. 2022a; Naseer et al. 2022; Wang et al. 2022; Han et al. 2022) has sparked a discussion about boosting the transferability across different Vision Transformers (ViTs). Despite the above advances, there is still a lack of effective attacking methods that can achieve strong transferability in a wider and more practical setting (*i.e.,* transferability across **model genus**[1]).

*These authors contributed equally.
†Corresponding author.

[1]One Model Genus is defined as a set of deep neural networks that have similar architectures. This concept was introduced in

To the best of our knowledge, this is one of the pioneering works aiming at improving **cross-model genera attacking transferability** (*e.g.,* the surrogate model is CNN, whereas the target system is based on a Transformer model). This task is more challenging due to the huge architecture gap between different model genera and the consequent distinctions of their extracted features (Naseer et al. 2021; Raghu et al. 2021a; Shao et al. 2021). Hereby, we present a few toy experiments to reveal the key distinction between discriminative features learned by CNN and ViT, respectively.

(i) As shown in Figure 1(a), we divide each image into several equal-sized patches and randomly blur some of them. It can be seen that with more patches blurred, pretrained ResNet-50 (He et al. 2016) achieves lower classification accuracy, whereas **DeiT-B** (Touvron et al. 2021) is **more robust to local detail loss**. (ii) Figure 1(b) shows that when we pick one of the partitioned patches to fill the whole image so as to retain only local image features as input to classification models, the results demonstrate that **ResNet-50** can still **utilize local details as classification cues** and achieve relatively higher accuracy. These toy experiments suggest that both Transformer and CNN models make predictions mainly based on global features, while CNN models frequently overfit to local patterns. This also inspired us to pay attention to algorithms that are perceptive to local characteristics in images.

During transferable adversarial attacks, to the best of our knowledge, cross model genus attacks have not been well explored in depth. Although model augmentation methods (Lin et al. 2020) is a kind of effective solution that can simulate the decision space of similar architectures by input transformation. However, existing model augmentation methods fail to enhance transferability across model genera. The reason is that they are specifically designed for the unique properties of a model genus and neglect the general and invariant features of different model genera (**Problem 1**). For example, scale (Lin et al. 2020) and translation invariance (Dong et al. 2019) of convolutional operation is utilized to produce augmented samples, which can generalize the feature space of a CNN surrogate to another of the same model genus. However, these augmentation methods bring less benefit to the generalization to a different model genus (*e.g.,* ViTs). Furthermore, previous works such as affine/linear (*e.g.,*, Admix (Wang et al. 2021a)), intensity (*e.g.,* SI (Lin et al. 2020)) and spectrum (*e.g.,* S^2I (Long et al. 2022)) transformation techniques, have primarily focused on global contents, neglecting the importance of diverse local regions (**Problem 2**). Hence, based on the aforementioned analysis, from the perspective of input transformation, we found that elastic deformation can adjust the local shape and contents while obtaining more augmented local patterns, thus extracting more local the general features.

Consequently, we propose a novel and generic input transformation approach called **Deformation-Constrained Warping (DeCoW)**, which applies a deformation to the local details of the target data (*e.g.,* an image and log mel spectrogram of an audio segment). In this way, the surrogate model has a tendency to rely on invariant features (*i.e.,* global features) from augmented samples. However, unconstrained elastic deformation may lead to excessive or unreasonable changes in image semantics. To tackle this issue, we further contribute an adaptive control strategy in DeCoW which can optimize the random deformation variable to a constrained point, ensuring the consistency of the global semantics of augmented samples and inputs. The main contributions and novelties of this paper can be summarised as

- We systematically investigated the task of cross model genus attack and revealed that the low transferability of adversarial samples generated by previous attack methods is due to insufficient manipulation on the local structure of the signal.
- We propose a generic Deformation-Constrained Warping (DeCoW) to boost the adversarial transferability across model genera. DeCoW can increase the diversity of local details such as local shape and contours through elastic deformation and adaptively constrains the magnitude and direction of warping transformation.
- Deformation-Constrained Warping Attack (DeCoWA) is proposed by integrating DeCoW into a gradient-based attack method. It is an approach that can be applied to various modalities of data such as image, video, and audio, and achieves superior transferability over the state-of-the-art attack methods by a significant margin.

Related Work

Intriguing Properties of CNN and Transformer

It is noted that CNNs excels at capturing high-frequency components of an image (Wang et al. 2020) since their primary parts, convolutional layers, function as individual high-pass filters (Park and Kim 2022). By contrast, Multi-Head Self-Attentions (MSAs), which bring key benefits to a transformer, serve as low-pass filters for spatial smoothing (Park and Kim 2022). This difference between the main components of these two model genera unveils why CNNs are more sensitive to trivial details and noise as compared to ViTs. Furthermore, ViTs are confirmed to have a stronger bias to object shape than CNNs, whereas CNNs show more bias to local textures (Naseer et al. 2021; Geirhos et al. 2019). From a different perspective, Raghu *et al.* (Raghu et al. 2021b) leveraged Centered Kernel Alignment (CKA) to measure the layer representation similarity between network blocks. They found that ViTs have more consistent representations across all layers than CNNs. These works expose different properties of the two model genera (*i.e.,* CNN and ViT), which are beneficial to get out of the predicament of low adversarial transferability across different genera.

Deformation-based Data Augmentation

Data augmentation is one of the most valid approaches to boost model generalization. For instance, some model-free transformations like affine transformation (*e.g.* scale, translation) and intensity transformation (*e.g.* blurring and adding

(Mahmood, Mahmood, and van Dijk 2021), existing models were categorized into several model genera, such as ViT model genus, and CNN model genus.

noise) are commonly used for training various models (He et al. 2016, 2020; Sohn et al. 2020). In addition to these methods, blending multiple images such as mixup (Zhang et al. 2018) have also been proposed to improve model generalization. Elastic deformation is another approach that is gradually being applied to model training because it can alter the shape or posture of objects (Xu et al. 2023).

Transferable Adversarial Attack

Transferable Attack on CNNs. For transferable attacks across different CNN models, there is a large body of works spanning from gradient-based enhanced methods (Dong et al. 2018; Lin et al. 2020; Wang et al. 2021b; Wei et al. 2022b, 2020), variance-tuning methods (Wang and He 2021; Xiong et al. 2022), knowledge-based methods (Gao et al. 2021; Wu et al. 2020; Ganeshan, S., and Radhakrishnan 2019; Huang et al. 2019; Inkawhich et al. 2019; Wang et al. 2021c; Zhou et al. 2018; Long et al. 2022; Luo et al. 2022), methods built on generative models (Nakka and Salzmann 2021; Yang et al. 2022), to input augmentation methods (Lin et al. 2020; Xie et al. 2019; Dong et al. 2019; Wang et al. 2021a). Among these, MIM (Dong et al. 2018), a gradient-based enhanced method, uses a momentum term to keep the gradient directions, and SI-NI-FGSM (Lin et al. 2020) further modifies it with a Nesterov momentum and variance tuning. Recent research points to the great potential of augmentation techniques to boost adversarial transferability. For example, differentiable stochastic transformations, DI-FGSM (Xie et al. 2019), is applied to aggregate diverse directions of gradients, Admix (Wang et al. 2021a) incorporates a small portion of images of other classes into the input example to increase gradient diversity, and S^2I-FGSM (Long et al. 2022) augments input from a perspective of the frequency domain. (Zhang et al. 2023a) constructed a candidate augmentation path pool to augment images from multiple image augmentation paths. And (Liang and Xiao 2023) advocated using stylized networks to prevent adversarial examples from using non-robust style features.

Transferable Attack on Transformers. The popularity of ViT incurs a rapidly expanding body of studies on developing effective algorithms to improve the transferability of adversarial examples across various ViT models. Naseer *et al.* (Naseer et al. 2022) propose self-ensemble (SE) and token refinement (TR) strategies to alleviate the sub-optimal results. However, Wei *et al.* (Wei et al. 2022a) point out that there are not enough class tokens to construct an SE in the real world. To avoid it, the skills of Pay No Attention (PNA) and PatchOut are designed to manipulate the attention mechanism and image features in parallel. Besides, Han *et al.* (Han et al. 2022) also utilize partial encoder blocks to replace all encoder blocks, effectively reducing overfitting to a specific surrogate model. Wang *et al.* (Wang et al. 2022) proposed an Architecture-Oriented Transferable Attacking (ATA) framework to take the architectural features of vision transformers into account. Zhang *et al.* (Zhang et al. 2023b) proposed the Token Gradient Regularization (TGR) to reduce the variance of the back-propagated gradient and utilizes the regularized gradient to generate adversarial samples.

Methodology

Preliminary and A Unified Paradigm

Formally, take the image classification model as an example, let $\mathcal{M}_\phi : x \to y$ represents a classifier with learned parameters ϕ, $x \in \mathbb{R}^{H\times W\times C}$ and $y \in \mathcal{Y} = \{1, 2, .., \#class\}$ denotes clean input and ground truth, respectively, where $\#class$ represents the number of classes. Unlike the unconstrained spatial transformation perturbation (Xiao et al. 2018), our goal is to craft an adversarial example $x^{adv} = x + \delta$ with perturbations δ, which can mislead the classifier to make a wrong decision, *i.e.*, $\mathcal{M}_\phi(x^{adv}) \neq y$ (untargeted attack). To limit attack strength, x^{adv} is supposed to be constrained in a ℓ_p-norm ball centered at x with a radius ϵ. Following the widely-used setting in prior research (Dong et al. 2018; Lin et al. 2020; Wang et al. 2021a; Dong et al. 2019), we restrict perturbations in a ℓ_∞-norm ball in this paper. Therefore, the generation of adversarial examples can be formulated as an optimization problem:

$$\arg\max_{\boldsymbol{x}^{adv}} \mathcal{L}(\mathcal{M}_\phi(x^{adv}), y), \quad \text{s.t. } \|\delta\|_\infty \leq \epsilon, \tag{1}$$

where $\mathcal{L}$ is the cross-entropy loss, which is commonly applied to the classification model. Nonetheless, in the black-box scenario, it is impossible to directly optimize Eq. (1) because parameters of $\mathcal{M}_\phi$ are unknown. To address this issue, a common practice is to generate adversarial examples via a surrogate model $\mathcal{S}_\theta$ and mislead the target model by the transferability of adversarial examples. According to the I-FGSM (Kurakin, Goodfellow, and Bengio 2017), adversarial example at $(t+1)$-th optimization iteration can be formulated as:

$$\boldsymbol{x}_{t+1}^{adv} = \mathrm{Clip}_x^\epsilon\{\boldsymbol{x}_t^{adv} + \alpha \cdot \mathrm{sign}(\nabla_{x_t^{adv}} \mathcal{L}(\mathcal{S}_\theta(x_t^{adv}, y)))\}, \tag{2}$$

where $\mathrm{Clip}_x^\epsilon\{\cdot\}$ is an operation to constrain attack strength to be within a ϵ-ball, α is the step size, $\mathrm{sign}(\cdot)$ denotes the sign function and $\nabla_{x_t^{adv}} \mathcal{L}(\mathcal{S}_\theta(x_t^{adv}, y))$ represents the gradient of the loss. Moreover, if data transformation $\mathcal{T}(\cdot)$ is leveraged to boost the transferability of the adversarial examples, we can modify Eq. (2) as a more unified paradigm:

$$\boldsymbol{x}_{t+1}^{adv} = \mathrm{Clip}_x^\epsilon\{\boldsymbol{x}_t^{adv} + \alpha \cdot \mathrm{sign}(\nabla_{x_t^{adv}} \mathcal{L}(\mathcal{S}_\theta(\mathcal{T}(x_t^{adv}), y)))\}. \tag{3}$$

Vanilla Warping Transformation (VWT)

We have revealed in the Introduction that the warping transformation can achieve a greater diversity of local details and contents of data. Hence, we propose Vanilla Warping Transformation (VWT) ($\mathcal{T}_v(x;\xi)$) controlled by random noise map ξ to replace the $\mathcal{T}(\cdot)$ in Eq. (3). To perform VWT, according to the core of TPS algorithm (Bookstein 1989; Donato and Belongie 2002), we set two interpolation functions Φ_x and Φ_y for the coordinate offsets in the x-direction and y-direction, respectively. Then, two sets of control points are manually set to acquire the unknown parameter coefficients for the Φ_x and Φ_y functions. Specifically, the original control points are defined as $O \in \mathbb{R}^{M\times 2}$, where M

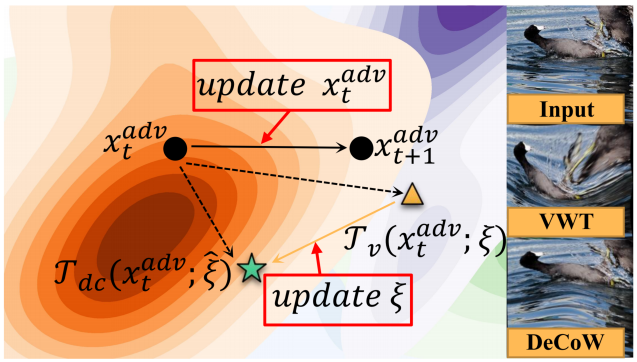

Figure 2: The process of updating ξ and x_t^{adv}. The left part shows a diagram of the update process. The right column enumerates the input sample and its result after VWT and DeCoW, respectively.

is the number of control points. The target control points $P \in \mathbb{R}^{M\times 2}$ are generated by slightly perturbing the original control points:

$$P = O + \xi, \tag{4}$$

where $\xi \in \mathbb{R}^{M\times 2}$ are randomly sampled from a uniform distribution. Then, through these two group control points, the TPS coefficients of Φ_x and Φ_y are obtained. For brevity, we put more details of the TPS algorithm in the supplementary material. Finally, we are able to achieve the transformation $\mathcal{T}_v(x;\xi)$ by interpolation as:

$$\mathcal{T}_v(x_t^{adv}[m,n];\xi) = x_t^{adv}[m+\Phi_x(m), n+\Phi_y(n)], \tag{5}$$

where (m,n) is the coordinate of any pixel on x_t^{adv}, and $x_t^{adv}[m,n]$ denotes the pixel value corresponding to coordinate (m,n). We incorporate VWT into Eq. (1) and get an optimization objective to craft adversarial examples:

$$\max_{\|\delta\|_\infty \le \epsilon} \mathcal{L}(\mathcal{S}_\theta(\mathcal{T}_v(x_t^{adv};\xi), y)). \tag{6}$$

Deformation-Constrained Warping (DeCoW)

And yet, the magnitude and direction of VWT are dependent on a random noise map ξ from a uniform distribution. Excessive and unreasonable deformation caused by ξ may be posed on some image regions or video/audio segments, invalidating their global content and semantics. To constrain the deformation variable ξ, we modify VWT with an adaptive control strategy. We denote the new transformation, Deformation-Constrained Warping (DeCoW), as $\mathcal{T}_{dc}(x;\hat{\xi})$. As illustrated in Figure 2, the content of the augmented sample with VWT is drastically changed while DeCoW can refine the deformation to an optimized point.

To achieve DeCoW, we first optimize initial point ξ by minimizing an objective, which is a reverse optimization to the Eq.(6):

$$\hat{\xi} = \arg\min_{\xi} \mathcal{L}(\mathcal{S}_\theta(\mathcal{T}_v(x_t^{adv};\xi)), y). \tag{7}$$

Through this process, we can update random deformation noise ξ to a secure point $\hat{\xi}$, reducing the variations of global semantics during elastic transformation. More specifically, we implement an iterative update with back-propagated gradients to achieve the optimization, which is formulated as follows:

$$\hat{\xi} = \xi - \beta \cdot \nabla_{\boldsymbol{\xi}}(\mathcal{L}(\mathcal{S}_\theta(\mathcal{T}_v(x_t^{adv};\xi), y)), \tag{8}$$

where β is the learning rate. Based on this, we can get a new set of control points:

$$P^{'} = O + \hat{\xi}. \tag{9}$$

Consequently, two new interpolation functions $\hat{\Phi}'_x$ and $\hat{\Phi}'_y$ can be obtained to augment samples as follow:

$$\mathcal{T}_{dc}(x^{adv_t}[m,n];\hat{\xi}) = x_t^{adv}[m+\hat{\Phi}'_x(m), n+\hat{\Phi}'_y(n)]. \tag{10}$$

Finally, we formulate this new attack loss as:

$$\max_{\|\delta\|_\infty \le \epsilon} \min_{\xi} \mathcal{L}(\mathcal{S}_\theta(\mathcal{T}_v(x_t^{adv};\xi), y)). \tag{11}$$

The defined max-min optimization can craft adversarial examples (maximization problem) with desire augmentation that has limited elastic deformation (minimization problem) to increase the diversity of local details and can keep image global semantics.

Warping for Video and Audio. As a general approach, DeCoW can also be applied to augment data from other modalities (*e.g.* video or audio). In processing these data with temporal information, we have made slight adjustments to allow it to be applied to the deformation of time series data. Specifically, suppose we want to apply DeCoW to a video $x \in \mathbb{R}^{K\times H\times W\times C}$ with K frames. Here we first sample continuous random noise within a periodic function *cosine function*, so that the noises have periodic prior information and relationships with each other. It can be formulated as follows:

$$\xi_\tau = \{\xi_\tau^{(1)}, \xi_\tau^{(2)}, ..., \xi_\tau^{(K)}\}. \tag{12}$$

Here $\xi_\tau \in \mathbb{R}^{K\times M\times 2}$, K is the number of video frames. And then, with this prior noise, we first perform an initial update on ξ_τ using Eq. (8) and get $\hat{\xi}_\tau = \{\hat{\xi}_\tau^{(1)}, \hat{\xi}_\tau^{(2)}, ..., \hat{\xi}_\tau^{(K)}\}$. Subsequently, in order to achieve a more smooth noise between frames, we adopt a momentum accumulation fashion and perform another update on the $\hat{\xi}$ to strengthen the temporal correlation as:

$$\hat{\xi}_\tau^{(i+1)} = d \cdot \hat{\xi}_\tau^{(i)} + (1-d) \cdot \hat{\xi}_\tau^{(i+1)}, \tag{13}$$

where d is a hyperparameter. Hence, through Eq. (13), we can perform continuous DeCoW for video clips and exploit the temporal information between consecutive frames.

Attack Algorithm

In this section, we proposed **Deformation-Constrained Warping Attack (DeCoWA)** algorithm by integrating DeCoW into MI-FGSM (Dong et al. 2018) method. Firstly, every time we obtain the adversarial gradient $g^{'}$, a maximization-minimization operation is required, updating the random noise ξ to obtain $\hat{\xi}$ via the minimization, and deriving the adversarial gradient $g^{'}$ based on $\hat{\xi}$ via the maximization. In addition, during the attack process, we apply

Surrogate	Method	ViT-B/16	DeiT-B	LeViT-256	PiT-B	CaiT-S-24	ConViT-B	TNT-S	Visformer-S	Avg
	Clean	88.70	96.00	94.60	93.50	97.20	94.40	89.90	95.20	93.69
Inc-v3	DIM	64.80	69.30	53.90	66.90	66.70	70.80	47.40	55.10	61.86
	TIM	66.20	77.30	64.60	78.00	76.50	78.50	54.00	70.00	70.64
	SIM	64.20	69.70	55.30	69.00	67.10	70.40	47.80	59.60	62.89
	Admix	57.00	59.30	43.60	60.80	57.60	63.10	37.60	46.40	53.34
	S^2IM	53.70	55.50	38.90	54.70	51.90	59.80	33.40	41.10	<u>48.62</u>
	DeCoWA	**44.80**	**36.40**	**22.80**	**40.40**	**35.90**	**44.20**	**21.60**	**25.70**	**33.97**
R50	DIM	60.30	53.00	39.80	47.50	51.40	58.80	41.20	33.80	48.23
	TIM	62.90	65.20	59.30	65.80	66.00	68.00	50.40	55.20	61.60
	SIM	58.30	50.40	42.70	48.20	51.10	55.00	42.40	34.50	47.82
	Admix	52.10	38.60	27.00	36.60	38.00	45.70	30.80	20.90	36.21
	S^2IM	40.60	31.80	20.50	29.40	30.70	37.80	20.90	19.00	<u>28.84</u>
	DeCoWA	**37.90**	**23.50**	**13.90**	**20.60**	**23.50**	**30.30**	**15.20**	**11.00**	**21.99**

Table 1: Classification accuracy (%) against eight trained ViT models under the transferable adversarial attack with single input transformation, where all methods integrate MI-FGSM. 'Clean' indicates the accuracy before the attack. The best and the second-best performances are labeled in bold and underline, respectively. We abbreviate Inception-v3 to inc-v3 and abbreviate ResNet-50 to R50. 'Avg' is the average classification accuracy.

multiple warping transformations to the adversarial example x_t^{adv} as:

$$\bar{g}_{t+1} = \frac{1}{N}\sum_{j=0}^{N} g_j^{'} = \frac{1}{N}\sum_{j=0}^{N} \nabla_{x_t^{adv}} \mathcal{L}(\mathcal{S}_\theta(\mathcal{T}_{dc}(x_t^{adv}; \hat{\xi}_j)), y), \tag{14}$$

where N is the number of adversarial warping transformations. Multiple transformations allow us to enhance the model features in different directions, enhancing the diversity of surrogate models. And then update the enhanced momentum:

$$g_{t+1} = \mu \cdot g_t + \frac{\bar{g}_{t+1}}{\|\bar{g}_{t+1}\|_1}, \tag{15}$$

where μ denotes the decay factor. Finally, based on the enhanced momentum g_{t+1} , the adversarial example is updated as:

$$x_{t+1}^{adv} = \text{Clip}_x^{\epsilon}\{x_t^{adv} + \alpha \cdot \text{sign}(g_{t+1})\}. \tag{16}$$

Experiments

Attack on Image Classification

In this section, we conduct extensive empirical evaluations on the trained image classifier and use the classification accuracy as the main evaluation metric, where the lower classification accuracy indicates better attack performance. Further details on the experimental setup and analysis are presented below.

Dataset. Following the previous works (Long et al. 2022), we evaluate the proposed method on images from the ImageNet-compatible dataset[2].

Baseline. We compare our approach with five state-of-the-art transfer-based attack methods, embracing Diverse Input Method (DIM) (Xie et al. 2019), Translation-Invariant Method (TIM) (Dong et al. 2019), Scale-Invariant Method (SIM) (Lin et al. 2020), Admix (Wang et al. 2021a) and Spectrum Simulation Attack Method (S^2IM) (Long et al. 2022) . They also boost the transferability of adversarial examples by employing data augmentation. Note that MI-FGSM (MIM) (Dong et al. 2018) is integrated into all the aforementioned methods.

Models. To realize the cross model genus attack, two CNN models are chosen as the surrogate models, including Inception-v3 (Szegedy et al. 2016), ResNet-50 (He et al. 2016). Then, the adversarial examples generated on CNN are tested on multiple ViT variants. More comparative experiments are provided in the supplementary materials, including using two ViT models ViT-B/16 (Dosovitskiy et al. 2021) and DeiT-B (Touvron et al. 2021) to attack multiple CNN variants.

Attack Setting. We follow the parameters setting (Dong et al. 2018). The perturbation budget is $\epsilon = 16.0$, the number of iterations is $T = 10$, and step size $\alpha = 1.6$. The decay factor for MIM is $\mu = 1.0$. The Gaussian kernel size for TIM is 7×7. The number of copies is 5 for SIM. The transformation probability for DIM is $p = 0.5$. The number of random samples for Admix is 3. In S^2IM, we set the number of spectrum transformations as 15. We set the number of DeCoW as $N = 15$, the number of control points is $M = 9$, learning rate $\beta = 0.02$.

Using CNNs to Attack ViTs. In Table 1, we used CNN as the surrogate model to attack various variants of the ViT model, which is a challenging task mentioned in (Mahmood, Mahmood, and van Dijk 2021; Shao et al. 2021; Bhojanapalli et al. 2021). Nevertheless, our proposed DeCoWA still achieves significant improvements. For instance, when applying Inception-v3 as the surrogate model, our attacking performance exceeds the existing best approaches S^2IM (48.62%, underline) by over **14.65%** on average.

Using CNNs to Attack CNNs. In Table 2, we realize a homologous model genera attack. It compares different adversarial examples generated from a CNN surrogate to other CNN targets (CNN $\rightarrow$ CNN). Here we have selected four classic CNN networks as the target models, i.e. ResNet-

[2]https://github.com/cleverhans-lab/cleverhans/tree/master/cleverhans_v3.1.0/examples/nips17_adversarial_competition/dataset

Surrogate	Mehtod	R101	VGG19	DN121	EfficientNet
	Clean	94.30	87.00	91.50	92.60
Inc-v3	DIM	49.00	26.90	31.40	31.80
	TIM	66.00	37.00	49.80	50.60
	SIM	53.80	30.90	33.30	36.10
	Admix	39.00	21.90	20.10	23.10
	S^2IM	36.20	18.80	19.30	21.40
	DeCoWA	**28.80**	**12.80**	**11.60**	**9.50**
R50	DIM	14.40	22.80	17.60	27.20
	TIM	32.30	32.30	34.70	41.40
	SIM	12.80	26.10	15.70	29.80
	Admix	5.80	16.40	8.40	17.10
	S^2IM	**5.40**	9.20	7.10	12.80
	DeCoWA	**5.40**	**6.80**	**3.90**	**7.80**

Table 2: Classification accuracy (%) against four trained CNN models under the transferable adversarial attack with single input transformation, where all methods integrate MI-FGSM. 'Clean' indicates the accuracy before the attack. The best performances are labeled in bold .

101 (R101), VGG19 (Simonyan and Zisserman 2014), DenseNet121(DN121 (Huang et al. 2017)) and EfficientNet (Tan and Le 2019). Under all settings, our DeCoWA achieved the best transferability as compared to existing state-of-the-art methods, which outperform the second-best method. Generally speaking, the methods that achieve good results in cross model genus attacks can also perform well in homologous model genera attacks.

Surrogate	Method	I3D	SlowFast	TimeS	Swin-S
I3D-50	BIM	0.64*	94.88	96.38	97.44
	DIM	0.85*	84.65	88.91	94.46
	SIM	0.00*	73.56	82.73	86.14
	DeCoWA	0.43*	**70.58**	**79.32**	**83.37**
SlowFast	BIM	89.98	1.71 *	94.88	98.72
	DIM	80.81	2.35*	85.29	93.60
	SIM	70.58	0.00*	75.69	91.90
	DeCoWA	**69.08**	2.13*	**73.35**	**88.49**
TimeS	BIM	88.49	91.26	0.00*	84.01
	DIM	69.51	73.99	0.21*	65.03
	SIM	55.22	53.94	0.00*	45.42
	DeCoWA	**44.56**	**49.04**	0.00*	**39.23**
Swin-S	BIM	92.54	94.03	95.10	0.00*
	DIM	78.89	79.96	85.93	0.21*
	SIM	63.11	70.36	78.25	0.00*
	DeCoWA	**44.99**	**50.75**	**66.74**	0.00*

Table 3: Classification accuracy (%) on four video recognition models. The adversaries are crafted on I3D-50, SlowFast, TimeSformer (TimeS), and Swin transformer. * indicates the attack performance under the white-box attack.

Attack on Video Recognition

In this section, we show that our DeCoWA also can be easily applied in attacking the video recognition models.

Attack Setting. We evaluate our approach using Kinetics-400 (Kay et al. 2017) (K400) datasets, which are widely used for action video recognition. 469 videos are chosen from the validation set to evaluate the effectiveness of our algorithm. Our proposed method is evaluated on four action video recognition models, i.e. I3D (Wang et al. 2018), SlowFast (Feichtenhofer et al. 2019), TimesFormer (Bertasius, Wang, and Torresani 2021) and Swin Transformer (Liu et al. 2021). All the models are trained on Kinetics-400. The spatial size of the input is 224×224. We make modifications based on the mmaction[3] to implement the attack for the frames. We skip every three frames to select 32 consecutive frames to construct an input clip, and we get 3 input clips for each video. We evaluate the performances against three popular transfer-based attacks, i.e. BIM (Kurakin, Goodfellow, and Bengio 2017), DIM (Xie et al. 2019) and SIM (Lin et al. 2020).

Experiment Analysis. Table 3 shows the comparison results, it shows that our DeCoW can process consecutive frames in a temporal manner, outperforming other input transformation methods in attacking video recognition models. This demonstrates the generality of our method. Meanwhile, we observed that our method has a more pronounced augmentation effect on ViT models and achieves stronger attack performance when used as a substitute model.

Surrogate	Method	Baseline	PANN	RGASC	ERGL
	Clean	68.30	73.40	77.40	75.50
Baseline	BIM	0.10*	42.60	39.50	52.10
	SI-FGSM	13.90*	46.80	43.80	54.80
	DeCoWA	0.20*	**40.80**	**37.60**	**50.80**
PANN	BIM	45.20	0.00*	8.90	22.70
	SI-FGSM	55.00	35.70*	39.90	43.80
	DeCoWA	**44.10**	0.00*	**7.90**	**21.40**
RGASC	BIM	33.30	57.50	17.90*	48.10
	SI-FGSM	37.40	57.90	24.30*	49.30
	DeCoWA	**28.70**	**55.50**	13.90*	**44.50**
ERGL	BIM	61.20	46.70	45.50	0.60*
	SI-FGSM	63.00	54.80	55.30	32.40*
	DeCoWA	**59.50**	**42.10**	**39.90**	0.40*

Table 4: Classification accuracy (%) on four audio recognition models. The adversaries are crafted on Baseline, PANN, ERGL, and RGASC, respectively. * indicates the performance under the white-box attack.

Attack on Audio Recognition

In this section, we show that our DeCoWA also can be easily applied in attacking the audio recognition models.

Attack Setting. Four acoustic scene classification models, i.e. Baseline[4], PANN (Kong et al. 2020), ERGL (Hou et al. 2022b) and RGASC (Hou et al. 2022a), and 2,518 audios selected from the validation set are used for the evaluation. We compared the proposed DeCoWA against two popular transfer-based attacks, i.e. BIM (Kurakin, Goodfellow, and Bengio 2017) and SI-FGSM (Lin et al. 2020). All the models are trained on TUT Urban Acoustic Scenes 2018. For this comparison, the image transformation method of DIM is not used, since it can not be applied to process speech signals directly, while our method can process speech signals easily.

Experiment Analysis. Table 4 shows that our DeCoWA consistently outperforms the state of the arts by crafting

[3]https://github.com/open-mmlab/mmaction2

[4]https://github.com/qiuqiangkong/dcase2018_task1

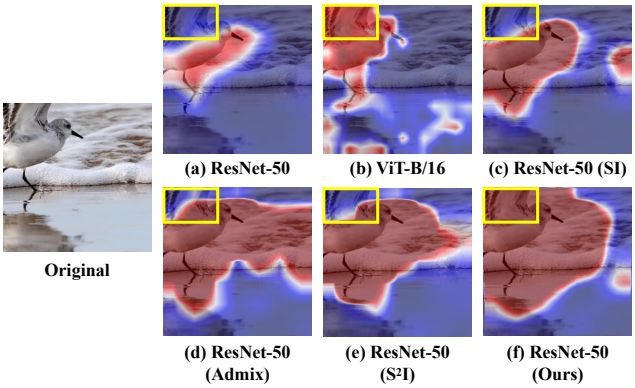

Figure 3: Visualization of Grad-CAM (Selvaraju et al. 2017) for two trained models ResNet-50 and ViT-B/16. (a)∼(b): the results for raw images on ResNet-50 and ViT-B/16. (c)∼(e): the results for SI (Lin et al. 2020), Admix (Wang et al. 2021a), S^2I (Long et al. 2022) images on ResNet-50. (f): the result for our DeCoW images on ResNet-50.

more generalized perturbations. Meanwhile, as far as we know, there are currently few methods aiming to specifically improve the transferability of audio adversarial samples, our algorithm provides a new approach to attack such systems. Note that here our DeCoW only combines with I-FGSM.

Visualization of Grad-CAM

To shed light on how our method works, we visualize the Grad-CAMs by ResNet-50, ViT-B in Figure 3. As shown in Figure 3 (a)∼(b) and (f), ResNet-50, which is prone to focus more on local and sparse regions of an object, can be transformed by our DeCoW to recognize an object in terms of its global appearance. For example, in the yellow box, DeCoW makes ResNet-50 pay attention to the bird's wing as well like the way of ViT-B/16, which enables the surrogate ResNet-50 to simulate the ViT-B/16 successfully. Still, other augmentation methods fail to achieve it. This indicates that the proposed DeCoW is a generalized transform and can explore more attention areas of the target system, and consequently better narrow the gap between CNNs and ViTs. Noticed that here we apply the CAM to the original sample instead of the enhanced sample to show how the augmented sample shift and extend the original CAM (Figure 3(a)). As a result, the backgrounds in Figure 3 are similar.

Visualization of Different Input Transformation

We present prior augmentation methods and their performance in Figure 4. Such methods augment input samples from various perspectives, including image size (Xie et al. 2019), translation (Dong et al. 2019), scale (Lin et al. 2020), linear multiple-image fusion (Wang et al. 2021a) and spectrum (Long et al. 2022). These transformations are prone to change global contents (*e.g.,* size, position, lighting, or color), while we seek a new transformation to preserve global semantics and increase the diversity of local details like local shape and contours, which are more general and invariant features (Mahmood, Mahmood, and van Dijk

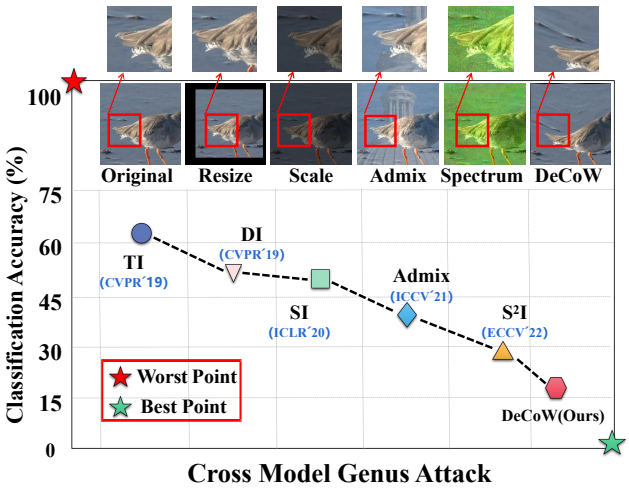

Figure 4: In comparison with other input transformation methods, our method makes profound changes with the local shape and contours (red box) thus accessing diverse localities, while others can only increase global diversity.

2021; Geirhos et al. 2019) to both CNN and Transformer-based models. Figure 4 depicts that warping transformation can cause deformation to the local region (red box), while the others are only able to make limited changes on the tail of the bird (lack of diversity).

Conclusion and Outlook

In this work, we argue that more attention should be paid to the task of cross model genus attacks. We proposed a novel technique, Deformation-Constrained Warping Attack (DeCoWA) to boost the adversarial transferability across model genera. It features applying constrained elastic deformation to input samples to simulate diverse models covering different model genera. Comprehensive experiments corroborate the superiority of DeCoWA for the task of cross model genus attack on data of various modalities. Therefore, our attack can serve as a strong baseline to compare future cross model genus attack. In the future, we will focus on the cross-data distribution attack in which the adversary can only access surrogate models trained on different data distributions and having distinct model genera from the target system.

Acknowledgments

The work was supported by the Natural Science Foundation of China under grants no. 62276170, 82261138629, the Science and Technology Project of Guangdong Province under grants no. 2023A1515011549, 2023A1515010688, the Science and Technology Innovation Commission of Shenzhen under grant no. JCYJ20220531101412030.

References

Bertasius, G.; Wang, H.; and Torresani, L. 2021. Is Space-Time Attention All You Need for Video Understanding? In

ICML 2021, volume 139 of *Proceedings of Machine Learning Research*, 813–824. PMLR.

Bhojanapalli, S.; Chakrabarti, A.; Glasner, D.; Li, D.; Unterthiner, T.; and Veit, A. 2021. Understanding Robustness of Transformers for Image Classification. In *ICCV 2021*, 10211–10221. IEEE.

Bookstein, F. L. 1989. Principal warps: Thin-plate splines and the decomposition of deformations. *IEEE TPAMI*, 11(6): 567–585.

Donato, G.; and Belongie, S. 2002. Approximate thin plate spline mappings. In *ECCV 2002*, 21–31. Springer.

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

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

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

Feichtenhofer, C.; Fan, H.; Malik, J.; and He, K. 2019. SlowFast Networks for Video Recognition. In *ICCV 2019*, 6201–6210. IEEE.

Ganeshan, A.; S., V. B.; and Radhakrishnan, V. B. 2019. FDA: Feature Disruptive Attack. In *ICCV 2019*, 8068–8078. IEEE.

Gao, L.; Huang, Z.; Song, J.; Yang, Y.; and Shen, H. T. 2021. Push & pull: Transferable adversarial examples with attentive attack. *IEEE Transactions on Multimedia*, 24: 2329–2338.

Geirhos, R.; Rubisch, P.; Michaelis, C.; Bethge, M.; Wichmann, F. A.; and Brendel, W. 2019. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *ICLR 2019*. OpenReview.net.

Han, Y.; Liu, J.; Liu, X.; Jiang, X.; Gu, L.; Gao, X.; and Chen, W. 2022. Enhancing adversarial transferability with partial blocks on vision transformer. *Neural Comput. Appl.*, 34(22): 20249–20262.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. B. 2020. Momentum Contrast for Unsupervised Visual Representation Learning. In *CVPR 2020*, 9726–9735. Computer Vision Foundation / IEEE.

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

Hochreiter, S.; and Schmidhuber, J. 1997. Long short-term memory. *Neural computation*, 9(8): 1735–1780.

Hou, Y.; Kang, B.; Hauwermeiren, W. V.; and Botteldooren, D. 2022a. Relation-guided acoustic scene classification aided with event embeddings. In *IJCNN 2022*, 1–8. IEEE.

Hou, Y.; Song, S.; Yu, C.; Song, Y.; Wang, W.; and Botteldooren, D. 2022b. Multi-dimensional Edge-based Audio Event Relational Graph Representation Learning for Acoustic Scene Classification. *CoRR*, abs/2210.15366.

Huang, G.; Liu, Z.; Van Der Maaten, L.; and Weinberger, K. Q. 2017. Densely connected convolutional networks. In *CVPR*, 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 *ICCV*, 4733–4742.

Hwang, J.; Han, D.; Heo, B.; Park, S.; Chun, S.; and Lee, J. 2022. Similarity of Neural Architectures Based on Input Gradient Transferability. *CoRR*, abs/2210.11407.

Inkawhich, N.; Wen, W.; Li, H. H.; and Chen, Y. 2019. Feature space perturbations yield more transferable adversarial examples. In *CVPR*, 7066–7074.

Kay, W.; Carreira, J.; Simonyan, K.; Zhang, B.; Hillier, C.; Vijayanarasimhan, S.; Viola, F.; Green, T.; Back, T.; Natsev, P.; Suleyman, M.; and Zisserman, A. 2017. The Kinetics Human Action Video Dataset. *CoRR*, abs/1705.06950.

Kong, Q.; Cao, Y.; Iqbal, T.; Wang, Y.; Wang, W.; and Plumbley, M. 2020. PANNs: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM TASLP*, 28: 2880–2894.

Kurakin, A.; Goodfellow, I. J.; and Bengio, S. 2017. Adversarial Machine Learning at Scale. In *ICLR 2017*. OpenReview.net.

Liang, K.; and Xiao, B. 2023. StyLess: Boosting the Transferability of Adversarial Examples. In *CVPR*, 8163–8172.

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

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In *ICCV 2021*, 9992–10002. IEEE.

Long, Y.; Zhang, Q.; Zeng, B.; Gao, L.; Liu, X.; Zhang, J.; and Song, J. 2022. Frequency Domain Model Augmentation for Adversarial Attack. In *ECCV 2022*, volume 13664 of *Lecture Notes in Computer Science*, 549–566. Springer.

Luo, C.; Lin, Q.; Xie, W.; Wu, B.; Xie, J.; and Shen, L. 2022. Frequency-driven Imperceptible Adversarial Attack on Semantic Similarity. In *CVPR 2022*, 15294–15303. IEEE.

Mahmood, K.; Mahmood, R.; and van Dijk, M. 2021. On the Robustness of Vision Transformers to Adversarial Examples. In *ICCV 2021*, 7818–7827. IEEE.

Nakka, K. K.; and Salzmann, M. 2021. Learning Transferable Adversarial Perturbations. In *NeurIPS 2021*, 13950–13962.

Naseer, M.; Ranasinghe, K.; Khan, S.; Hayat, M.; Khan, F. S.; and Yang, M. 2021. Intriguing Properties of Vision Transformers. In *NeurIPS 2021*, 23296–23308.

Naseer, M.; Ranasinghe, K.; Khan, S.; Khan, F. S.; and Porikli, F. 2022. On Improving Adversarial Transferability of Vision Transformers. In *ICLR 2022*. OpenReview.net.

Park, N.; and Kim, S. 2022. How Do Vision Transformers Work? In *ICLR 2022*. OpenReview.net.

Raghu, M.; Unterthiner, T.; Kornblith, S.; Zhang, C.; and Dosovitskiy, A. 2021a. Do Vision Transformers See Like Convolutional Neural Networks? In *NeurIPS 2021*, 12116–12128.

Raghu, M.; Unterthiner, T.; Kornblith, S.; Zhang, C.; and Dosovitskiy, A. 2021b. Do vision transformers see like convolutional neural networks? *Advances in Neural Information Processing Systems*, 34: 12116–12128.

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 *ICCV 2017*, 618–626. IEEE Computer Society.

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

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Sohn, K.; Berthelot, D.; Carlini, N.; Zhang, Z.; Zhang, H.; Raffel, C.; Cubuk, E. D.; Kurakin, A.; and Li, C. 2020. FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence. In *NeurIPS 2020*.

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

Tan, M.; and Le, Q. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *ICML*, 6105–6114. PMLR.

Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; and Jégou, H. 2021. Training data-efficient image transformers & distillation through attention. In *ICML*, 10347–10357. PMLR.

Wang, H.; Wu, X.; Huang, Z.; and Xing, E. P. 2020. High-Frequency Component Helps Explain the Generalization of Convolutional Neural Networks. In *CVPR 2020*, 8681–8691. Computer Vision Foundation / IEEE.

Wang, X.; Girshick, R. B.; Gupta, A.; and He, K. 2018. Non-Local Neural Networks. In *CVPR 2018*, 7794–7803. Computer Vision Foundation / IEEE Computer Society.

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

Wang, X.; He, X.; Wang, J.; and He, K. 2021a. Admix: Enhancing the Transferability of Adversarial Attacks. In *ICCV 2021*, 16138–16147. IEEE.

Wang, X.; Lin, J.; Hu, H.; Wang, J.; and He, K. 2021b. Boosting Adversarial Transferability through Enhanced Momentum. In *BMVC 2021*, 272. BMVA Press.

Wang, Y.; Wang, J.; Yin, Z.; Gong, R.; Wang, J.; Liu, A.; and Liu, X. 2022. Generating transferable adversarial examples against vision transformers. In *ACM MM*, 5181–5190.

Wang, Z.; Guo, H.; Zhang, Z.; Liu, W.; Qin, Z.; and Ren, K. 2021c. Feature importance-aware transferable adversarial attacks. In *ICCV*, 7639–7648.

Wei, Z.; Chen, J.; Goldblum, M.; Wu, Z.; Goldstein, T.; and Jiang, Y. 2022a. Towards Transferable Adversarial Attacks on Vision Transformers. In *AAAI 2022*, 2668–2676. AAAI Press.

Wei, Z.; Chen, J.; Wei, X.; Jiang, L.; Chua, T.; Zhou, F.; and Jiang, Y. 2020. Heuristic Black-Box Adversarial Attacks on Video Recognition Models. In *AAAI 2020*, 12338–12345. AAAI Press.

Wei, Z.; Chen, J.; Wu, Z.; and Jiang, Y. 2022b. Cross-Modal Transferable Adversarial Attacks from Images to Videos. In *CVPR 2022*, 15044–15053. IEEE.

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 *CVPR*, 1161–1170.

Xiao, C.; Zhu, J.-Y.; Li, B.; He, W.; Liu, M.; and Song, D. 2018. Spatially transformed adversarial examples. *arXiv preprint arXiv:1801.02612*.

Xie, C.; Zhang, Z.; Zhou, Y.; Bai, S.; Wang, J.; Ren, Z.; and Yuille, A. L. 2019. Improving Transferability of Adversarial Examples With Input Diversity. In *CVPR 2019*, 2730–2739. Computer Vision Foundation / IEEE.

Xiong, Y.; Lin, J.; Zhang, M.; Hopcroft, J. E.; and He, K. 2022. Stochastic Variance Reduced Ensemble Adversarial Attack for Boosting the Adversarial Transferability. In *CVPR 2022*, 14963–14972. IEEE.

Xu, M.; Yoon, S.; Fuentes, A.; and Park, D. S. 2023. A Comprehensive Survey of Image Augmentation Techniques for Deep Learning. *Pattern Recognit.*, 137: 109347.

Yang, X.; Dong, Y.; Pang, T.; Su, H.; and Zhu, J. 2022. Boosting Transferability of Targeted Adversarial Examples via Hierarchical Generative Networks. In *ECCV 2022*, volume 13664 of *Lecture Notes in Computer Science*, 725–742. Springer.

Zhang, H.; Cissé, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2018. mixup: Beyond Empirical Risk Minimization. In *ICLR 2018*. OpenReview.net.

Zhang, J.; Huang, J.-t.; Wang, W.; Li, Y.; Wu, W.; Wang, X.; Su, Y.; and Lyu, M. R. 2023a. Improving the Transferability of Adversarial Samples by Path-Augmented Method. In *CVPR*, 8173–8182.

Zhang, J.; Huang, Y.; Wu, W.; and Lyu, M. R. 2023b. Transferable Adversarial Attacks on Vision Transformers with Token Gradient Regularization. In *CVPR*, 16415–16424.

Zhou, W.; Hou, X.; Chen, Y.; Tang, M.; Huang, X.; Gan, X.; and Yang, Y. 2018. Transferable adversarial perturbations. In *(ECCV)*, 452–467.