Temporal Consistency Constrained Transferable Adversarial Attacks with Background Mixup for Action Recognition

Ping Li[1,2] , **Jianan Ni**[1] , **Bo Pang**[1]

[1]School of Computer Science and Technology, Hangzhou Dianzi University
[2]State Key Laboratory for Novel Software Technology, Nanjing University
{lpcs, njn, pbcs}@hdu.edu.cn

Abstract

Action recognition models using deep learning are vulnerable to adversarial examples, which are transferable across other models trained on the same data modality. Existing transferable attack methods face two major challenges: 1) they heavily rely on the assumption that the decision boundaries of the surrogate (*a.k.a.*, source) model and the target model are similar, which limits the adversarial transferability; and 2) their decision boundary difference makes the attack direction uncertain, which may result in the gradient oscillation, weakening the adversarial attack. This motivates us to propose a **B**ackground **M**ixup-induced **T**emporal **C**onsistency (**BMTC**) attack method for action recognition. From the input transformation perspective, we design a model-agnostic background adversarial mixup module to reduce the surrogate-target model dependency. In particular, we randomly sample one video from each category and make its background frame, while selecting the background frame with the top attack ability for mixup with the clean frame by reinforcement learning. Moreover, to ensure an explicit attack direction, we leverage the background category as guidance for updating the gradient of adversarial example, and design a temporal gradient consistency loss, which strengthens the stability of the attack direction on subsequent frames. Empirical studies on two video datasets, *i.e.*, *UCF101* and *Kinetics-400*, and one image dataset, *i.e.*, *ImageNet*, demonstrate that our method significantly boosts the transferability of adversarial examples across several action/image recognition models.

1 Introduction

Action recognition has established itself as a fundamental task in computer vision, and has widespread applications in many areas, such as video surveillance, robot, and virtual reality. In recent years, Deep Neural Networks (DNNs) have gained large popularity in developing action recognition models, such as SlowFast [Feichtenhofer *et al.*, 2019] and Video Vision Transformer (ViViT) [Arnab *et al.*, 2021], which focuses on capturing the spatiotemporal features of video. However, these model are vulnerable to adversarial examples [Goodfellow *et al.*, 2015], *i.e.*, adding human imperceptible perturbations to clean samples, which fool the classification model to make incorrect predictions. This raises the important security concerns from both academia and industry.

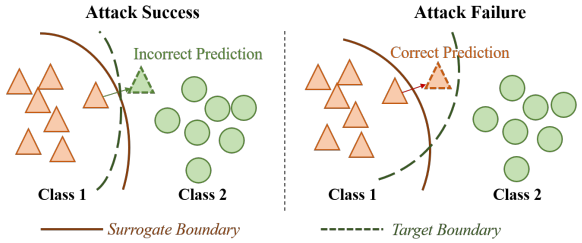

Figure 1: Illustration of the decision boundaries of the surrogate model and the target model.

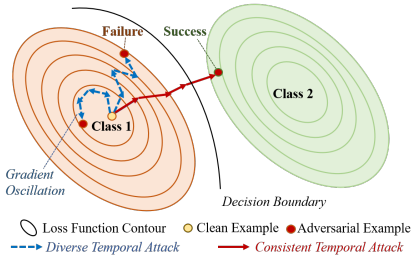

Figure 2: Illustration of diverse and consistent temporal attack.

In real-world scenarios, due to some privacy issue, the prior knowledge such as the parameters and the architecture of target model (*i.e.*, attacked model) are often unavailable to the attacker. Hence, we investigate the transferability of adversarial examples by adopting one white-box model as surrogate or source model to realize black-box attacks for action recognition models. To this end, there are only a few studies devoted to the transfer-based attack on video recognition models. For example, [Wei *et al.*, 2022] presents a Temporal Translation (TT) attack method that optimizes the adversarial perturbations on the temporal translated video clips to make the attack less sensitive to varying temporal patterns. From the cross-modal transfer perspective of image-to-video, [Wang *et al.*, 2023] improves the transferability of adversarial examples in black-box scenario by introducing global inter-frame inter-

action into attack process and disrupting the inherently local correlations of frames within a video; [Wei *et al.*, 2024] generates adversarial examples from white-box image models to attack video models, and optimize perturbations via reducing the similarity of intermedia features between clean frame and adversarial frame. However, the above methods have *two major drawbacks* including: 1) they update gradients to generate adversarial examples which makes them heavily rely on the assumption that the decision boundaries of the surrogate (*a.k.a.*, source) model and the target model are similar, *i.e.*, the transfer-based attack will fail (*e.g.*, adversarial example is correctly recognized by target model) when the two boundaries are isolated away (see Fig. 1), and 2) the existence of the boundary difference incurs the uncertainty problem of attack direction, and they make the attempt in many possible directions, resulting in possible large gradient oscillation as shown in Fig. 2 and thus weakening the adversarial attack.

For the *first* problem, from the input transformation perspective, we adopt the adversarial mixup strategy to generate adversarial examples, which makes the transferable attack less dependable on the assumption of the decision boundary similarity between the surrogate model and the target model. The so-called *mixup* acts as data augmentation that linearly interpolates two images and corresponding labels. Here we do not change the label as in [Wang *et al.*, 2021a]. By contrast, we consider the semantic perturbation on clean sample by making a set of background frames, each of which corresponds to one category. In particular, we add the background frame with the highest attack reward to the clean frame. All clean frames within a video may be mixed up with different background frames. To identify the background frame with the strongest attack ability, we leverage reinforcement learning by designing a reward function for the adversarial Mixer, which includes attack success award, transfer award, and temporal background consistency award. The background frame with the top score is selected for mixing up with the clean frame. Here the temporal background consistency award makes the selected background frames be similar or the same for nearby frames. These skills are wrapped in the model-agnostic *Background Adversarial Mixup* (BAM) module.

For the *second* problem, from the temporal attack perspective, some work [Wei *et al.*, 2022] presents a gradient-based temporal translation attack that optimizes the adversarial perturbations on temporally translated video clips, while some others [Wei *et al.*, 2024][Chen *et al.*, 2023] focus on feature-based attacks that minimize the cosine similarities between the intermediate features of the (warped) clean frame and the adversarial counterparts. However, they fail to consider the local relation of nearby frames. Meanwhile, [Wang *et al.*, 2023] disrupts the temporal local correlations by reducing the similarity of the adversarial examples of nearby frames, but it introduces the large diversity of adversarial frames by minimizing their feature cosine similarity. This makes the attack directions vary greatly, leading to possible inverse gradient directions as shown in Fig. 2. Thus, it adds the difficulty in pushing the adversarial example away from the current decision boundary. Hence, we design the *Background-induced Temporal Gradient enhancement* (BTG) module that leverages the background attack loss and the temporal gradient consistency loss to make nearby frames have similar gradients. This makes the attack directions be consistent across the frames along the temporal dimension, enhancing the attack ability of adversarial example gradually.

We briefly summarize the main contributions as follows:

- We study the transfer-based black-box attack on action recognition models, and propose a **B**ackground **M**ixup-induced **T**emporal **C**onsistency (**BMTC**) attack method to boost the transferability of adversarial examples.
- We perform a model-agnostic input transformation by considering background semantics, *i.e.*, adversarially mixup background frame from other categories with clean frame, while that background frame is determined by the reward function using reinforcement learning with good transferability and temporal consistency.
- We strengthen the attack direction across sequential frames in a progressive way, by imposing the temporal gradient consistency constraint on the loss and guiding the attack in the direction of the background category.

2 Related Works

2.1 Transfer-based Image Attacks

Previous methods often adopt white-box attacks, such as Fast Gradient Sign Method (FGSM) [Goodfellow *et al.*, 2015] that updates the gradient on the clean sample to maximize the loss function, Projected Gradient Descent (PGD) [Madry *et al.*, 2018] that is an iterative FGSM. To improve the transferability of adversarial examples on black-box models, there are three strategies: 1) *data augmentation* (*i.e.*,input transformation), *e.g.*, diversity input attack [Xie *et al.*, 2019], Scale-Invariant Method (SIM) [Lin *et al.*, 2020], Translation-Invariant (TI) attack [Dong *et al.*, 2019], Adversarial Mixup (AdMix) [Wang *et al.*, 2021a] that mixes two images using Mixup [Zhang *et al.*, 2018], and Path-Augmented Method (PAM) [Zhang *et al.*, 2023] that constructs a candidate augmentation path pool for path selection with greedy search; 2) *gradient modification*, *e.g.*, Momentum Iterative (IM) attack [Dong *et al.*, 2018], Skip Gradient Method (SGM) [Wu *et al.*, 2020], Variance Momentum Iterative tuning (VMI) [Wang and He, 2021], and Gradient-Related Adversarial Attack (GRA) [Zhu *et al.*, 2023]; 3) *feature disruption*, *e.g.*, Intermediate Level Attack (ILA) [Huang *et al.*, 2019], Feature Importance-Aware attack (FIA) [Wang *et al.*, 2021b], and ILA with Data Augmentation (ILA-DA) [Yan *et al.*, 2023].

2.2 Transfer-based Video Attacks

Transferable attacks on video models are less explored compared to that on image models. Existing methods attempt to break the temporal relations, *e.g.*, from the gradient update perspective, Temporal Translation (TT) [Wei *et al.*, 2022] does multiple translations on frames along the temporal dimension to generate the clips under different translation patterns and optimizes the perturbations on these clips; from the feature disruption perspective, some methods minimize the cosine similarities between the intermediate features of the clean frames [Wei *et al.*, 2024] or the warped clean frames

[Chen *et al.*, 2023] and the adversarial counterparts. Nevertheless, they treat video as a collection of unordered images, ignoring the inherent temporal continuity. In contrast, [Wang *et al.*, 2023] disrupts the temporal local correlations of nearby frames by reducing the similarity of the adversarial counterparts, and minimizes the cosine similarities among the intermediate features of the benign points lying on a convex hull, which leads to diverse adversarial frames. This may incur gradient oscillation, *e.g.*, some sample gradients update in reverse direction, because the uncertain attack directions vary greatly.

2.3 Action Recognition Models

Much progress has been made in deep neural network based action recognition models. Existing methods have two groups: 1) Convolutional Neural Networks (CNNs) methods, *e.g.*, Inflated 3D ConvNet (I3D) [Carreira and Zisserman, 2017], Non-Local neural networks (NL) [Wang *et al.*, 2018], SlowFast [Feichtenhofer *et al.*, 2019] that captures spatial semantics at low frame rate and the motion dynamics at fine temporal resolution, Temporal Pyramid Networks (TPN) [Yang *et al.*, 2020] that uses the source features and the fusion features to form a feature hierarchy to capture action instances at various tempos; 2) Vision Transformer (ViT) methods, *e.g.*, VideoTransformer Network (VTN) [Neimark *et al.*, 2021] that is built on top of any given 2D spatial network and attends to the entire video, Time-Space Transformer (TimeSformer) [Bertasius *et al.*, 2021] that leverages 3D self-attention over the space-time volume to capture long-range temporal dependency among frames, Motionformer [Patrick *et al.*, 2021] aggregates implicitly motion path information to model the temporal dynamic scenes, and VideoSwin [Liu *et al.*, 2022] that encourages an inductive bias of locality by adapting the Swin Transformer [Liu *et al.*, 2021] for a better speed-accuracy trade-off.

3 Methodology

This section introduces the proposed Background Mixup-induced Temporal Consistency (BMTC) attack method in the transfer-based black-box setting for action recognition models. The overall framework is illustrated in Fig. 3, which consists of two primary components, *i.e.*, Background Adversarial Mixup (BAM) module and Background-induced Temporal Gradient enhancement (BTG) module.

3.1 Problem Definition

Given a video sample x represented by a tensor $\mathbf{X} \in \mathbb{R}^{T\times H\times W\times C}$ with the ground-truth label $y \in \mathcal{Y} = \{1, 2, \cdots, K\}$ represented by a one-hot vector $\mathbf{y} \in \mathbb{R}^K$, where $\{N, H, W, C\}$ denote the number, height, width, and channel of the video, each frame is indexed by $t \in \{1, 2, \cdots, T\}$, and there are K action categories, transfer-based adversarial attack aims to generate an adversarial example $\mathbf{X}^{adv} = \mathbf{X} + \delta$ by the surrogate model $g(\cdot)$ to fool the target model $f(\cdot) : \mathbf{X} \mapsto \mathcal{Y}$ to make incorrect predictions, *i.e.*, $f(\mathbf{X}^{adv}) \neq y$ without knowing the gradients and architecture of the model, where the perturbation δ is restricted by the ℓ_p-norm $\|\delta\|_p \leq \epsilon$. Here, $\epsilon > 0$ is a constant that governs the perturbation magnitude, and we adopt $\ell_{\inf}$-norm and untargeted adversarial attacks as in [Wei *et al.*, 2024][Wei *et al.*, 2022]. The objective of untargeted adversarial attack is formulated as:

$$\arg\max_{\delta} J(f(x+\delta), y), s.\ t.\ \|\delta\|_{\inf} \leq \epsilon, \tag{1}$$

where the function $J(\cdot)$ often adopts cross-entropy loss. Since this work focuses on the black-box transfer-based attack, the attacker (*a.k.a.*, adversary) has no access to details of the target model $f(\cdot)$. We aim to improve the black-box transferability of video adversarial examples on other action recognition models.

3.2 Background Adversarial Mixup

To ensure the good transferability of attacks, the decision boundary of the target model is expected to approach that of the surrogate model. When their decision boundaries are far away, it results in poor transferability of adversarial examples, *e.g.*, they make different predictions on the same sample. Gradient modification methods heavily depends the above decision boundary assumption, since the gradients back-propagated during training have essential compacts on reshaping the decision boundary of the action recognition model. To circurvement this problem, we present a model-agnostic input transformation method, which adopts the adversarial mixup with background in video using a Mixer that consists of a feature extractor (*e.g.*, ResNet50) and a classification head (*e.g.*, fully-connected layer). The rationale behind this is that there exist strong correlations between action category and action background, *e.g.*, *surfing* in the blue sea and *riding* on a road, which makes it possible that mixing the background from other categories with a given video might mislead the model to yield wrong predictions. This raises a key problem that how to select the background frame. The working mechanism is shown below.

We construct a set of candidate background frames, which are obtained by randomly selecting one background frame of a randomly chosen video from each category. The selected background frame is made by applying one zero-shot video object segmentation method, *i.e.*, Isomerous transformer [1] [Yuan *et al.*, 2023], to abandon foreground part, such as person and action-related objects. Note that the label of each background frame is kept the same as that in the original video, and there is a one-to-one correspondence of label mapping between original video and background frame, *i.e.*, $y^{back} = y + K$, whose one-hot vector is $\mathbf{y}^{back} \in \mathbb{R}^K$ and its index is $k' = \{1, 2, \cdots, K\}$. The background frame is repeated to form the background video $\mathbf{X}^{back} \in \mathbb{R}^{T\times H\times W\times C}$ for a video $\mathbf{X}$. These background videos are added to the set of original videos and they are fed into the surrogate model for fine-tuning, whose loss $\mathcal{L}_{ft}$ is defined as

$$\mathcal{L}_{ft} = -\sum_{k=1}^{K} \mathbf{y}_k \log(\hat{\mathbf{y}}_k) - \lambda \sum_{k'=1}^{K} \mathbf{y}_{k'}^{back} \log(\hat{\mathbf{y}}_{k'}^{back}), \tag{2}$$

where the first term is action loss and the second term is background loss, $\{\hat{\mathbf{y}}_k, \hat{\mathbf{y}}_{k'}^{back}\}$ are the probability logits from

[1] https://github.com/DLUT-yyc/Isomer

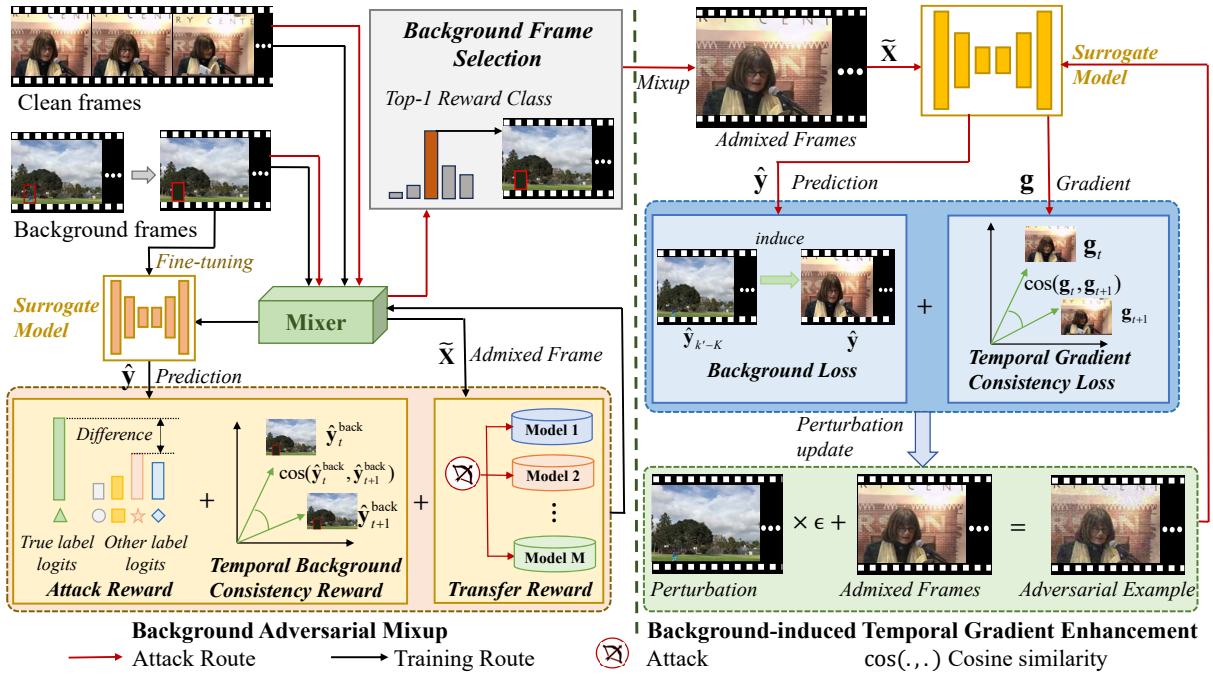

Figure 3: Overall framework of our Background Mixup-induced Temporal Consistency (BMTC) attack method for action recognition.

the model, the hyper-parameter λ (empirically set to 0.2) is used to balance the contribution of the two terms to the objective. Note that fine-tuning the surrogate model is to equip the surrogate model with some discriminative ability of different background categories, so the action loss should play a dominant role and the background loss serves as an auxiliary.

To select the background frame with the most powerful attack ability, we adopt the reinforcement learning to compute three reward functions, *i.e.*, attack reward R_{attack}, transfer reward $R_{transfer}$, and temporal background consistency reward R_{tbc}. These reward functions are used to encourage the Mixer to generate the admixed sample to successfully attack the surrogate model. Here the Mixer is used to mixup the selected background video with the original video, *i.e.*, $\tilde{\mathbf{X}} = (1-\gamma)\cdot\mathbf{X} + \gamma\cdot\mathbf{X}^{back}$, where the constant γ is empirically set to 0.2. For the frame-level mixup, the i-th admixed frame is $\tilde{\mathbf{X}}_i = (1-\gamma)\cdot\mathbf{X}_i + \gamma\cdot\mathbf{X}_i^{back}$. Actually, admixed sample serves as the vanilla adversarial example.

Attack Reward. For a video x with its label y and the predicted probability vector $\hat{\mathbf{y}} = [\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_K]$, this reward R_{attack} takes the difference of the label prediction probabilities before and after the adversarial mixup, *i.e.*,

$$R_{attack} = \max\{\hat{y}_{\setminus k}\} - \hat{y}_k, \ s.t.\ k = y, \tag{3}$$

where $\{\hat{y}_{\setminus k}\}$ is a set excluding the k-th entry, which indicates the true label k. When the probability of the true label $\hat{y}_k$ increases, the reward score decreases; when the maximum probability of other categories $\max\{\hat{y}_{\setminus k}\}$ increases, the reward score rises, *i.e.*, when the video is incorrectly predicted, we should give the model positive reward.

Transfer Reward. To improve the transferability of adversarial example across different models, we provide the transfer reward $R_{transfer}$ to attack M black-box target models, and averages their attack rewards, *i.e.*,

$$R_{transfer} = \frac{1}{K}\sum_{m=1}^{M} R_{attack}^{m}, \tag{4}$$

where R_{attack}^{m} is the attack award of the m-th target model.

Temporal Background Consistency Reward. Since similar samples have similar backgrounds, nearby frames are expected to have similar backgrounds. For a video x with T frames indexed by t, the neighbouring frames are expected to be mixed with similar or the same background frames, *i.e.*, temporal consistency among background frames, such that the attack strength will be boosted. We define the Temporal Background Consistency (TBC) reward R_{tbc} as

$$R_{tbc} = \frac{1}{T-1}\sum_{t=1}^{T-1}\frac{\hat{\mathbf{y}}_t^{back}\cdot\hat{\mathbf{y}}_{t+1}^{back}}{\|\hat{\mathbf{y}}_t^{back}\|_2\cdot\|\hat{\mathbf{y}}_{t+1}^{back}\|_2}, \tag{5}$$

where $\|\cdot\|_2$ denotes the ℓ_2-norm, and the vector $\hat{\mathbf{y}}_t^{back}\in\mathbb{R}^K$ is the predicted probability of the t-th background frame.

Therefore, the total reward of selecting the background frame is

$$R_{total} = R_{attack} + \alpha_1\times R_{transfer} + \alpha_2\times R_{tbc} \tag{6}$$

where the constants $\alpha_1 > 0$ (set to 0.3) and $\alpha_2 > 0$ (set to 0.1) are used to control the contributions of the transfer reward and the temporal background consistency reward, respectively. Finally, we select the background frame with the highest total reward score from K candidates to mixup with the original video frames sequentially.

3.3 Background-induced Temporal Gradient Enhancement

To further improve the transferability of vanilla adversarial example (*i.e.*, admixed sample), we introduces two adversar-

ial losses including the background attack loss and the temporal gradient consistency loss. The former is used to guide the surrogate model to generate the adversarial example, by learning the probability logits along the direction of the original action category, which is associated with the mixed background frame. In particular, we maximize the cross-entropy loss of other categories and minimize that of the background category (*i.e.*, fool the model to wrongly predict the sample as the selected background category), *i.e.*,

$$\mathcal{L}_{back} = -\sum_{j=1, j\neq k'\text{-}K}^{K} \mathbf{y}_j \log \hat{\mathbf{y}}_j + \mathbf{y}_{k'\text{-}K} \log \hat{\mathbf{y}}_{k'\text{-}K}, \quad (7)$$

where the background frame belongs to the $(k' - K)$-th category from the original K actions.

To makes the attack direction be consistent between nearby adversarial frames, we design the Temporal Gradient Consistency (TGC) loss to encourage the surrogate model to generate the adversarial counterparts with similar gradients for nearby frames. By this means, the attack strength is gradually enhanced across the frames along the temporal dimension. Mathematically, we adopt the cosine similarity to evaluate the temporal consistency between the gradients of two nearby adversarial frames, *i.e.*,

$$\mathcal{L}_{tgc} = \frac{1}{T-1} \sum_{t=1}^{T-1} \frac{\mathbf{g}^{t^\top} \cdot \mathbf{g}^{t+1}}{\|\mathbf{g}^t\|_2 \cdot \|\mathbf{g}^{t+1}\|_2}, \quad (8)$$

where $\mathbf{g}^t \in \mathbb{R}^{H\cdot W\cdot 3}$ denotes the gradient vector of the t-th adversarial frame $\mathbf{X}_t^{adv} \in \mathbb{R}^{H\times W\times C}$, which is calculated as

$$g^t = \nabla_{\mathbf{X}_t^{adv}} \mathcal{L}_{back}(\mathbf{X}_t^{adv}, \mathbf{y}; \theta), \mathbf{g}^t = \text{vec}(g^t), \quad (9)$$

where θ denotes the model parameter set, $\text{vec}(\cdot)$ vectorizes a tensor to the corresponding vector, and $\nabla_{\mathbf{X}_t^{adv}} L$ is the gradient of the loss $\mathcal{L}_{back}$ *w.r.t.*the t-th frame.

Therefore, the total temporal adversarial attack loss is:

$$\mathcal{L}_{total} = \mathcal{L}_{back} + \beta \mathcal{L}_{tgc}, \quad (10)$$

where the hyper-parameter $\beta \in [0, 1]$ (set to 0.1) controls the balance between the two terms during optimization.

3.4 Adversarial Example Generation

This work adopts the Projected Gradient Descent (PGD) algorithm [Madry *et al.*, 2018] to compute the sample gradients of the model, and the adversarial example at the i-th iteration is updated in the following way:

$$\mathbf{X}^{i+1} = \Pi_{\mathbf{X},\epsilon}\left[\mathbf{X}^i + \eta \cdot \text{sign}\left(\nabla_{\mathbf{X}^i} \mathcal{L}_{total}(\mathbf{X}^i, \mathbf{y})\right)\right], \quad (11)$$

where $\eta > 0$ is the step size, $\Pi_{\mathbf{X},\epsilon}[\cdot] = \min(\max(\mathbf{X}^i, \mathbf{X} - \epsilon), \mathbf{X} + \epsilon)$ is the projection function that controls the perturbation magnitude, $\mathbf{X}$ is the clean sample, and the initial $\mathbf{X}^0$ is the admixed sample. When the iteration achieves the maximum number N_{iter} (*e.g.*, set to 10 for a good tradeoff between performance and efficiency), we obtain the ultimate adversarial example to attack the target models.

4 Experiments

All experiments were performed on a server equipped with two 48G NVIDIA A6000 graphics cards. The codes are compiled with PyTorch 1.11, Python 3.8, and CUDA 11.4. Our code is available at Github.

4.1 Datasets and Evaluation Metric

We conduct experiments on two video benchmarks including UCF101 [Soomro *et al.*, 2012] and Kinetics-400 [Carreira and Zisserman, 2017], and one image benchmark ImageNet [Deng *et al.*, 2009]. Following [Wei *et al.*, 2022][Wei *et al.*, 2024], we adopt the Attack Success Rate (ASR) as the metric, which calculates the rate of adversarial examples misclassified by the black-box model. The higher the ASR, the better the transferability is. As a common practice, one video is randomly selected from each category and correctly classified by all target models.

4.2 Experimental Settings

For the adversarial Mixer, we train a classification model from scratch with ResNet50 as backbone on Kinetics-400 and ImageNet respectively, and the training epochs are 100. For the surrogate model fine-tuning, we adopt TPN and ResNet101 as the surrogate of video (20 epochs) and image (10 epochs) attack, respectively, and the hyper-parameter of the fine-tuning loss is $\lambda = 0.2$. For both of them, the initial learning rate is 0.1, momentum is 0.9, and the weight decay is 1e4. For the adversarial example generation, we set the maximum perturbation ϵ to 16 for video and 8 for image, the attack step size η to 1.6 for video and 0.8 for image, the maximum iteration number N_{iter} to 10.

For action recognition, we examine three CNN models with different architectures, *i.e.*, Non-local network (NL) [Wang *et al.*, 2018], SlowFast (SF) [Feichtenhofer *et al.*, 2019], and Temporal Pyramid Networks (TPN) [Yang *et al.*, 2020], which employ ResNet50 and ResNet101 as backbones. We train video models from scratch with Kinetics-400 and fine-tune them on UCF101. Following [Wei *et al.*, 2024], we skip every other frame from randomly selected clip with 64 consecutive frames to make input for Kinetics-400, and use 32 consecutive frames as input clips for UCF101. Moreover, we also examine four Transformer models including VideoTransformer Network (VTN) [Neimark *et al.*, 2021], Time-Space Transformer (TimeSformer) [Bertasius *et al.*, 2021], Motionformer [Patrick *et al.*, 2021], and VideoSwin [Liu *et al.*, 2022]. Following [Wei *et al.*, 2024], we sample one clip per video and each clip consists of 16 frames with the temporal stride 4. For image classification, we examine four models including ResNet18 [He *et al.*, 2016], ResNet101, ResNeXt50 (RNX50) [Xie *et al.*, 2017], and DenseNet121 (DN121) [Huang *et al.*, 2017]. Note that NL-101, SF-101 and TPN-101 using ResNet101 act as white-box models (surrogate) for generating adversarial examples to attack other black-box models (target). The input spatial size is 224×224. The model number M is 3 for video and 4 for image.

4.3 Quantitative Results

Action Recognition. The attack results of action recognition models are reported in Table 1 on UCF101, Tables 2 and 3 on Kinetics-400. The best records are highlighted in boldface, and the second best ones are underlined. Here, "RN" denotes ResNet, the records with the gray color are the results of white-box attack to be overlooked.

From these tables, we observe that our method achieves more satisfying performance across several action recogni-

Surrogate	Attack	Venue	NL		SlowFast		TPN	
			RN101	RN50	RN101	RN50	RN101	RN50
NL-101	PGD	ICCV'18	92.08	31.68	11.88	15.84	8.91	10.89
	AA	ICML'20	94.05	28.71	17.82	20.79	12.87	11.88
	TT	AAAI'22	81.19	62.38	48.52	58.42	37.62	39.60
	Ours		88.12	79.21	**71.28**	**69.31**	**53.47**	54.45
SF-101	PGD	ICCV'18	17.82	23.76	93.07	36.63	9.90	14.85
	AA	ICML'20	19.80	21.78	93.07	35.64	16.83	18.81
	TT	AAAI'22	54.46	62.38	80.10	52.48	36.63	38.61
	Ours		**73.26**	**78.22**	86.14	71.29	52.47	**66.34**
TPN-101	PGD	ICCV'18	11.88	10.89	9.90	14.85	84.16	33.67
	AA	ICML'20	13.86	17.82	10.89	18.81	89.11	24.75
	TT	AAAI'22	59.41	62.37	43.56	44.55	78.22	35.64
	Ours		69.31	76.24	64.35	67.33	82.17	73.27

Table 1: Performance comparison on UCF101.

Surrogate	Attack	Venue	NL		SlowFast		TPN	
			RN101	RN50	RN101	RN50	RN101	RN50
NL-101	PGD	ICCV'18	94.75	13.75	13.50	16.25	11.25	14.25
	AA	ICML'20	99.00	25.25	20.25	21.25	14.25	20.75
	TT	AAAI'22	97.25	77.25	78.25	75.75	57.25	62.00
	Ours		98.50	85.25	80.50	**83.00**	**75.25**	78.50
SF-101	PGD	ICCV'18	20.25	25.75	91.00	36.25	15.75	14.85
	AA	ICML'20	26.50	30.00	93.25	40.75	19.50	25.50
	TT	AAAI'22	59.75	61.25	94.00	75.25	53.50	62.75
	Ours		**67.00**	70.25	94.50	83.50	74.25	**83.75**
TPN-101	PGD	ICCV'18	14.00	10.25	15.50	14.85	94.00	33.50
	AA	ICML'20	19.75	21.50	24.75	31.50	99.00	24.75
	TT	AAAI'22	49.75	57.50	69.25	66.00	95.50	89.25
	Ours		63.25	**70.75**	**81.50**	80.25	98.50	84.25

Table 2: Performance comparison on Kinetics-400.

Attack	Surrogate	VTN	TimeSformer	Motionformer	VideoSwin
TT	NL	46.75	40.75	36.25	41.50
	SlowFast	49.50	45.50	41.50	32.25
	TPN	51.25	37.75	44.00	41.50
ENS-I2V-MF	Ensemble	53.50	42.00	36.75	56.25
AENS-I2V-MF	Ensemble	54.00	43.75	39.50	55.50
Ours	NL	67.25	61.25	59.25	**69.75**
	SlowFast	69.00	65.75	57.75	66.50
	TPN	**72.50**	**66.50**	**60.00**	68.25

Table 3: Performance comparison on Kinetics-400 (ViT models).

tion models including NL, SlowFast, and TPN, with different architectures. For example, when using the adversarial example from SF-101 to attack NL and TPN with RN50 as backbone, our method has a gain of 9.0% and 21.0% respectively, in comparison to TT on Kinetics-400; when using the adversarial example from TPN-101 to attack NL and SlowFast with RN50 as backbone, our method has an improvement of 13.87% and 22.78% respectively, compared to the most competitive alternative TT on UCF101. This demonstrates the superiority of the adversarial mixup strategy with reinforcement learning and the temporal consistency among nearby frames. Besides CNN models, the performance improvements are also found on ViT models in Table 3.

Meanwhile, the cross-modal attack results in terms of ASR are shown in Table 4 on UCF101 and Table 5 on Kinetics-400. Here, the surrogate models of the compared methods are AlexNet [Krizhevsky *et al.*, 2012], ResNet-101 [He *et al.*, 2016], SqueezeNet [Iandola *et al.*, 2016], and VGG-16 [Simonyan and Zisserman, 2015] in image domain, while we

Attack	Venue	SF/TPN→NL		NL/TPN→SF		NL/SF→TPN	
		RN101	RN50	RN101	RN50	RN101	RN50
I2V	CVPR'22	52.22	54.20	37.87	44.55	33.91	46.02
GCECA	AAAI'23	56.19	63.87	44.80	52.47	38.86	52.72
I2V-MF	TPAMI'24	56.93	57.67	37.87	48.26	37.87	48.26
Ours		**71.29**	**77.23**	**67.82**	**68.32**	**52.97**	**60.40**

Table 4: Average results on UCF101.

Attack	Venue	SF/TPN→NL		NL/TPN→SF		NL/SF→TPN	
		RN101	RN50	RN101	RN50	RN101	RN50
I2V	CVPR'22	44.25	54.13	64.13	63.94	65.38	72.19
GCECA	AAAI'23	54.56	63.88	71.88	70.63	72.19	77.31
I2V-MF	TPAMI'24	46.44	55.31	66.06	65.63	69.75	74.63
Ours		**65.13**	**70.50**	**81.00**	**81.63**	**74.75**	**81.13**

Table 5: Average results on Kinetics-400.

Surrogate	Attack	Venue	ResNet18			ResNet101			ResNeXt50			DenseNet121		
			Ori	+Ours	Δ	Ori	+Ours	Δ	Ori	+Ours	Δ	Ori	+Ours	Δ
RN18	MI-FGSM	CVPR'18	100	100	0.0	40.3	47.9	7.6	43.4	50.7	7.3	51.0	56.6	5.6
	AdMix	ICCV'21	100	100	0.0	62.1	64.1	2.0	65.1	65.2	0.1	73.5	73.9	0.4
	PAM	CVPR'23	100	100	0.0	45.6	53.5	7.9	48.8	55.6	6.8	57.7	65.6	7.9
	BSR	CVPR'24	100	100	0.0	79.6	84.4	4.8	81.7	86.1	4.4	88.7	91.7	3.0
RN101	MI-FGSM	CVPR'18	44.4	55.2	10.8	99.6	100	0.4	44.7	55.9	11.2	44.8	53.5	8.7
	AdMix	ICCV'21	75.7	76.2	0.5	99.7	99.7	0.0	73.0	73.6	0.6	75.5	76.1	0.6
	PAM	CVPR'23	59.3	68.6	9.3	99.3	98.9	-0.4	55.4	63.2	7.8	58.2	67.9	9.7
	BSR	CVPR'24	87.3	91.3	4.0	99.9	99.9	0.0	87.4	91.7	4.3	86.1	90.4	4.3
RNX50	MI-FGSM	CVPR'18	38.8	48.5	9.7	38.4	49.4	11.0	99.0	100	1.0	41.9	51.3	9.4
	AdMix	ICCV'21	67.9	68.3	0.4	63.1	63.8	0.7	98.0	99.1	1.1	69.7	72.3	2.6
	PAM	CVPR'23	52.1	58.5	6.4	45.4	53.7	8.3	98.2	97.6	-0.6	56.1	64.2	8.1
	BSR	CVPR'24	81.3	87.1	5.8	76.6	84.8	8.2	99.5	100	0.5	83.8	90.2	6.4
DN121	MI-FGSM	CVPR'18	51.2	59.3	8.1	42.5	50.7	8.2	47.3	56.4	9.1	99.9	99.9	0.0
	AdMix	ICCV'21	78.4	76.9	-1.5	68.8	68.4	-0.4	74.4	73.5	-0.9	99.8	100	0.2
	PAM	CVPR'23	65.5	70.0	4.5	52.6	58.1	5.5	59.3	63.4	4.1	99.6	99.8	0.2
	BSR	CVPR'24	88.7	92.6	3.9	79.2	84.5	5.3	87.2	92.0	4.8	100	100	0.0

Table 6: Performance comparison on ImageNet.

use the other two video models as the surrogate models, *e.g.*, SF and TPN as surrogate and NL as target (group 1). The results are averaged over those surrogate models, which indicates the advantages of the proposed attack method.

Image Classification. The results of image classification models are reported in Table 6 on ImageNet. Here, "Ori" denotes the vanilla attack method (column 2). Note that the temporal consistency reward and background-induced loss fail in image domain. Even though only using the adverse award and the transfer award to select the background, ours still improves the attack performance of the vanilla ones. This once again validates the power of the adversarial mixup with the background frame.

4.4 Ablation Study

We report the average ASR over black-box attacks and the hyper-parameters keep still as in training unless specified.

Individual component. The results are shown in Table 7, where the baseline is vanilla PGD, which is very poor as it neglects temporal cues. When adding the background attack loss (row 2), the performance is doubled; when using the Background Adversarial Mixup (BAM) module (row 3) or Background-induced Temporal Gradient enhancement (BTG) module (row 4), the performance is largely improved and the former is better. When using both of BAM and BTG (Ours), the performance achieves the best. When abandoning the reward R_{tbc} (row 5), the performance degrades,

Attack Method	SF/TPN→NL		NL/TPN→SF		NL/SF→TPN	
	RN101	RN50	RN101	RN50	RN101	RN50
Baseline	14.85	17.33	10.89	15.35	9.41	12.87
+$\mathcal{L}_{back}$	24.50	33.42	30.69	32.43	24.75	25.24
+BAM	53.96	60.40	49.01	52.97	35.15	47.52
+BTG	46.53	54.95	43.56	48.02	30.69	50.00
Ours w/o R_{tbc}	65.84	73.26	62.87	60.40	51.98	53.96
Ours w/o $\mathcal{L}_{tgc}$	59.90	65.35	55.94	57.43	45.05	49.50
Ours	**71.29**	**77.23**	**67.82**	**68.32**	**52.97**	**60.40**

Table 7: Ablation of components on UCF101. "w/o" is without.

Surrogate	Target	N_{iter}				ϵ			
		1	5	10*	20	4	8	16*	32
NL-101	SF-101	30.69	60.40	71.28	73.27	2.97	33.66	71.28	98.02
	SF-50	26.73	56.44	69.32	68.32	3.96	32.67	69.32	98.02
	TPN-101	30.69	46.53	53.47	56.44	0.00	17.82	53.47	91.09
	TPN-50	27.72	48.51	54.45	54.46	0.00	15.84	54.45	94.06
SF-101	NL-101	32.67	53.47	73.27	72.28	4.95	35.64	73.27	95.05
	NL-50	34.65	54.46	78.22	81.19	10.89	36.63	78.22	100.00
	TPN-101	27.72	40.59	52.48	54.46	2.97	24.75	52.48	84.16
	TPN-50	30.69	49.50	66.34	67.33	6.93	29.70	66.34	92.08
TPN-101	NL-101	29.70	65.35	69.31	72.28	6.93	39.60	69.31	91.09
	NL-50	34.65	64.36	76.24	77.23	13.86	36.63	76.24	97.03
	SF-101	34.65	51.49	64.36	65.35	5.94	33.66	64.36	93.07
	SF-50	36.63	55.45	67.33	68.32	4.95	37.62	67.33	94.06

Table 8: Ablation of N_{iter} and ϵ on UCF101.

which demonstrates the background consistency along the temporal dimension affects the total reward that decides the selected background frame. When abandoning $\mathcal{L}_{tgc}$ (row 6), the attack performances deteriorate significantly by about 7% to 12%, which validates the importance of gradient consistency between two nearby adversarial frames.

N_{iter} and ϵ. The results are shown in Table 8, which shows that the attack performance is naturally improved with the increasing iteration number N_{iter} and the maximum perturbation ϵ. However, we observe that when N_{iter} rises from 10 to 20, the performance becomes stable and even suffers from the bottleneck (row 2/4 in group 1) at much larger computational cost. So we choose 10 for N_{iter}. Besides, while the performance is greatly boosted by using larger perturbations, *e.g.*, ϵ=32, the adversarial examples are easily found by human, which violates the attack rule that requires small human-imperceptible noise. So we use the trade-off 16 for ϵ.

γ and β. The results are shown in Table 9, where the attack performance rises up when γ in the adversarial mixup increases from 0.1 to 0.2 and then decreases when γ is over 0.2. This demonstrates that the background frame should not dominate the mixup. Similar observations are found for the hyper-parameter β of the temporal gradient consistency loss $\mathcal{L}_{tgc}$, when it starts from 0.01 to 0.1. This suggests that neither too small nor large values are taken for β.

4.5 Computational Efficiency

We report the parameter size, the computational cost (GFLOPs), the inference speed (fps), and the ASR score on UCF101 and Kinetics-400 in Table 10. From the table, our BMTC method enjoys a more satisfying tradeoff between performance and efficiency. For example, compared to Temporal Translation (TT) [Wei *et al.*, 2022], ours have much higher ASR score (69.31 vs 52.47 on UCF101, 75.69 vs 62.75 on Kinetics-400) with only nearly one-tenth GFLOPs of TT

Surrogate	Target	γ				β				
		0.1	0.2*	0.4	0.6	0.01	0.05	0.1*	0.2	0.5
NL-101	SF-101	59.41	71.28	41.58	37.62	58.42	63.37	71.28	59.41	48.51
	SF-50	61.39	69.32	39.60	30.69	57.43	59.41	69.32	57.43	51.49
	TPN-101	43.56	53.47	32.67	32.67	43.56	46.53	53.47	46.53	32.67
	TPN-50	41.58	54.45	29.70	29.70	43.56	42.57	54.45	40.59	33.66
SF-101	NL-101	64.36	73.27	45.54	37.62	58.42	60.40	73.27	55.45	43.56
	NL-50	70.30	78.22	48.51	38.61	61.39	65.35	78.22	62.38	48.51
	TPN-101	42.57	52.48	34.65	31.68	39.60	47.52	52.48	43.56	35.64
	TPN-50	50.50	66.34	39.60	34.65	51.49	58.42	66.34	53.47	41.58
TPN-101	NL-101	51.49	69.31	37.62	36.63	55.45	63.37	69.31	51.49	43.56
	NL-50	57.43	76.24	47.52	40.59	64.36	70.30	76.24	66.34	56.44
	SF-101	53.47	64.36	40.59	32.67	53.47	62.38	64.36	50.50	42.57
	SF-50	51.49	67.33	42.57	38.61	48.51	59.41	67.33	44.55	35.64

Table 9: Ablation of γ (adversarial mixup) and β (temporal gradient consistency loss) on UCF101.

Attack Method	Params	FLOPs	FPS	ASR↑	
	(M)↓	(G)↓	↑	UCF101	Kinetics-400
PGD [Madry *et al.*, 2018]	99.7	217.4	357.2	11.88	15.54
AA [Croce and Hein, 2020]	99.7	295.3	290.4	15.35	22.96
TT [Wei *et al.*, 2022]	99.7	3634.1	17.8	52.47	62.75
Ours	123.4	391.6	288.9	**69.31**	**75.69**

Table 10: Computational cost comparison.

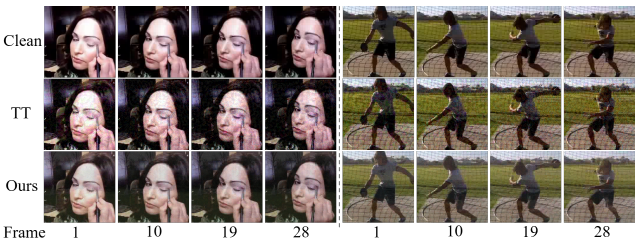

Figure 4: Examples of UCF101 (left) and Kinetics-400 (right).

at a 16 times faster inference speed.

4.6 Visualization of Adversarial Examples

To visualize the performance of our attack, we randomly chose one video from UCF101 and Kinetics-400, respectively, and show their adversarial examples in Fig. 4 (zoom in for better view). Compared to the baseline Temporal Translation (TT) [Wei *et al.*, 2022], the perturbations on the adversarial examples of ours are almost imperceivable by human but the model can be fooled to make wrong predictions.

5 Conclusion

This work presents a transferable black-box adversarial attack method for action recognition by considering both the temporal background consistency and the temporal gradient consistency. In particular, we adopt the adversarial mixup strategy to mix the clean sample with the background frame from other categories. To ensure the attack ability of the background frame, we design a reward function that considers the temporal consistency among nearby frames and the transferability across different models. Moreover, we strengthen the transferability of adversarial example by adopting the background-induced temporal consistency on the gradients of sample across frames. Empirical studies on both video and image datasets validate the effectiveness of the proposed attack on several models with different architectures.

Acknowledgments

This work was supported in part by Zhejiang Provincial Natural Science Foundation of China under Grant LR23F020002, and in part by Hangzhou Key Research and Development Program under Grant 2024SZD1A12.

References

[Arnab *et al.*, 2021] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lucic, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 6816–6826, 2021.

[Bertasius *et al.*, 2021] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139, pages 813–824, 2021.

[Carreira and Zisserman, 2017] João Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733, 2017.

[Chen *et al.*, 2023] Kai Chen, Zhipeng Wei andJingjing Chen, Zuxuan Wu, and Yu-Gang Jiang. GCMA: generative cross-modal transferable adversarial attacks from images to videos. In *Proceedings of the 31st ACM International Conference on Multimedia (ACM MM)*, pages 698–708, 2023.

[Croce and Hein, 2020] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119, pages 2206–2216, 2020.

[Deng *et al.*, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009.

[Dong *et al.*, 2018] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9185–9193, 2018.

[Dong *et al.*, 2019] Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 4312–4321, 2019.

[Feichtenhofer *et al.*, 2019] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 6201–6210, 2019.

[Goodfellow *et al.*, 2015] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[Huang *et al.*, 2017] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269, 2017.

[Huang *et al.*, 2019] Qian Huang, Isay Katsman, Zeqi Gu, Horace He, Serge J. Belongie, and Ser-Nam Lim. Enhancing adversarial example transferability with an intermediate level attack. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 4732–4741, 2019.

[Iandola *et al.*, 2016] Forrest N. Iandola, Matthew W. Moskewicz, Khalid Ashraf, Song Han, William J. Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and <1mb model size. *arXiv preprint*, arXiv:1602.07360, 2016.

[Krizhevsky *et al.*, 2012] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1106–1114, 2012.

[Lin *et al.*, 2020] Jiadong Lin, Chuanbiao Song, Kun He, Liwei Wang, and John E. Hopcroft. Nesterov accelerated gradient and scale invariance for adversarial attacks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.

[Liu *et al.*, 2021] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 9992–10002, 2021.

[Liu *et al.*, 2022] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR*, pages 3202–3211, 2022.

[Madry *et al.*, 2018] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018.

[Neimark *et al.*, 2021] Daniel Neimark, Omri Bar, Maya Zohar, and Dotan Asselmann. Video transformer network. In *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 3163–3172, 2021.

[Patrick *et al.*, 2021] Mandela Patrick, Dylan Campbell, Yuki M. Asano, Ishan Misra, Florian Metze, Christoph Feichtenhofer, Andrea Vedaldi, and João F. Henriques. Keeping your eye on the ball: Trajectory attention in video transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 12493–12506, 2021.

[Simonyan and Zisserman, 2015] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

[Soomro *et al.*, 2012] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint*, arXiv:1212.0402, 2012.

[Wang and He, 2021] Xiaosen Wang and Kun He. Enhancing the transferability of adversarial attacks through variance tuning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1924–1933, 2021.

[Wang *et al.*, 2018] Xiaolong Wang, Ross B. Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7794–7803, 2018.

[Wang *et al.*, 2021a] Xiaosen Wang, Xuanran He, Jingdong Wang, and Kun He. Admix: Enhancing the transferability of adversarial attacks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 16138–16147, 2021.

[Wang *et al.*, 2021b] Zhibo Wang, Hengchang Guo, Zhifei Zhang, Wenxin Liu, Zhan Qin, and Kui Ren. Feature importance-aware transferable adversarial attacks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 7619–7628, 2021.

[Wang *et al.*, 2023] Ruikui Wang, Yuanfang Guo, and Yunhong Wang. Global-local characteristic excited cross-modal attacks from images to videos. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence (AAAI)*, pages 2635–2643, 2023.

[Wei *et al.*, 2022] Zhipeng Wei, Jingjing Chen, Zuxuan Wu, and Yu-Gang Jiang. Boosting the transferability of video adversarial examples via temporaltranslation. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence (AAAI)*, pages 2659–2667, 2022.

[Wei *et al.*, 2024] Zhipeng Wei, Jingjing Chen, Zuxuan Wu, and Yu-Gang Jiang. Adaptive cross-modal transferable adversarial attacks from images to videos. *IEEE Transactions on Pattern Analysis and MachineIntelligence (TPAMI)*, 46(5):3772–3783, 2024.

[Wu *et al.*, 2020] Dongxian Wu, Yisen Wang, Shu-Tao Xia, James Bailey, and Xingjun Ma. Skip connections matter: On the transferability of adversarial examples generated with resnets. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.

[Xie *et al.*, 2017] Saining Xie, Ross B. Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5987–5995, 2017.

[Xie *et al.*, 2019] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 2730–2739, 2019.

[Yan *et al.*, 2023] Chiu Wai Yan, Tsz-Him Cheung, and Dit-Yan Yeung. ILA-DA: improving transferability of intermediate level attack with data augmentation. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

[Yang *et al.*, 2020] Ceyuan Yang, Yinghao Xu, Jianping Shi, Bo Dai, and Bolei Zhou. Temporal pyramid network for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 588–597, 2020.

[Yuan *et al.*, 2023] Yichen Yuan, Yifan Wang, Lijun Wang, Xiaoqi Zhao, Huchuan Lu, Yu Wang, Weibo Su, and Lei Zhang. Isomer: Isomerous transformer for zero-shot video object segmentation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 966–976, 2023.

[Zhang *et al.*, 2018] Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. Mixup: Beyond empirical risk minimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018.

[Zhang *et al.*, 2023] Jianping Zhang, Jen-tse Huang, Wenxuan Wang, Yichen Li, Weibin Wu, Xiaosen Wang andYuxin Su, and Michael R. Lyu. Improving the transferability of adversarial samples by path-augmented method. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8173–8182, 2023.

[Zhu *et al.*, 2023] Hegui Zhu, Yuchen Ren, Xiaoyan Sui, Lianping Yang, and Wuming Jiang. Boosting adversarial transferability via gradient relevance attack. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 4718–4727, 2023.