Global-Local Characteristic Excited Cross-Modal Attacks from Images to Videos

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

[1]School of Computer Science and Engineering, Beihang University, China
[2]Zhongguancun Laboratory, Beijing, China
{rkwang, andyguo, yhwang}@buaa.edu.cn

Abstract

The transferability of adversarial examples is the key property in practical black-box scenarios. Currently, numerous methods improve the transferability across different models trained on the same modality of data. The investigation of generating video adversarial examples with image-based substitute models to attack the target video models, i.e., cross-modal transferability of adversarial examples, is rarely explored. A few works on cross-modal transferability directly apply image attack methods for each frame and no factors especial for video data are considered, which limits the cross-modal transferability of adversarial examples. In this paper, we propose an effective cross-modal attack method which considers both the global and local characteristics of video data. Firstly, from the global perspective, we introduce inter-frame interaction into attack process to induce more diverse and stronger gradients rather than perturb each frame separately. Secondly, from the local perspective, we disrupt the inherently local correlation of frames within a video, which prevents black-box video model from capturing valuable temporal clues. Extensive experiments on the UCF-101 and Kinetics-400 validate the proposed method significantly improves cross-modal transferability and even surpasses stronger baseline using video models as substitute model. Our source codes are available at https://github.com/lwmming/Cross-Modal-Attack.

Introduction

Deep Neural Networks (DNNs) have achieved great success in a wide range of computer vision tasks (He et al. 2016; Ren et al. 2015; He et al. 2017). Unfortunately, these DNNs are highly vulnerable to adversarial examples, i.e., the inconspicuous perturbations injected to the input can easily deceive the model, leading to incorrect decisions. This phenomenon implies the existence of potential safety risks in current deployed DNN-based systems, such as autonomous driving, face recognition payment. Thus, researches on adversarial example are necessary and helpful to the development of reliable DNNs.

Analogous to adversarial attacks for images (Goodfellow, Shlens, and Szegedy 2014; Kurakin et al. 2017; Madry et al. 2017; Carlini and Wagner 2017; Dong et al. 2018; Xie et al. 2019), video attacks have also attracted more and more attentions from researchers, because of the applications of DNNs in various video related tasks (Wu et al. 2019; Wang et al. 2019; Kim et al. 2019; Feichtenhofer et al. 2019; Yang et al. 2020). Recently, a couple of works (Wei et al. 2019; Jiang et al. 2019; Wei et al. 2020; Zhang et al. 2020; Wang, Sha, and Su 2021; Wei et al. 2022c; Wei, Yan, and Li 2022) have been proposed to perform video attacks and achieved promising performances. However, most of them are query-based methods, which seem infeasible in reality due to considerable budgets of queries. To ameliorate this problem, a transfer-based video attack named TT (Wei et al. 2022a,b) is proposed. TT adopts temporal translations to augment video data to improve the transferability of adversarial examples. In detail, TT is essentially a transfer-based video attack on homomodal models, i.e., both the substitute and target models are video models. Nevertheless, cross-modal attack is rarely explored, i.e., generating video adversarial examples using image-based substitute models to attack the target video models.

Comparing with the query-based attacks and transfer-based video attack on homomodal models, cross-modal attack setting has two typical distinctions. Firstly, comparing with the query-based video attacks, no information is required about the target model, which avoids the hassle of frequent queries. Secondly, comparing with the transfer-based video attacks with video-based substitute models, image-based substitute models are easier to collect and the attack process is more simplified. If the attack performance of cross-modal attack is comparable to or even better than transfer-based video attacks with video-based substitute models, it will be more practical and economical. To the best of our knowledge, I2V (Wei et al. 2022b) is the first work to tackle this new challenging attack setting. In detail, it feeds each frame into the model and craft video adversarial examples by minimizing the cosine similarity of intermediate feature between clean frames and adversarial frames then concatenating each adversarial frame according to the temporal order of original video. Unfortunately, it treats a video as an orderless image set and ignores the inherent temporal information in videos. Consequently, the cross-modal transferability of I2V (Wei et al. 2022b) is slightly moderate.

In this paper, by considering the characteristic of video data from both global and local perspectives, we propose

*Corresponding author.

an effective cross-modal transfer attack framework from images to videos. From the globality perspective, videos usually consist of a large number of semantic-consistent frames. An intuitive idea is to perform inter-frame interactions in video attack process. In other words, the gradient for current frame fuses information from associated frames, which has the potential to generate more diverse attack direction thus more likely to fool various black-box video models. To this end, we extend a video clip to be a convex hull spanned by isolated frames and minimize the cosine similarities between the intermediate features of the benign points lied on convex hull and their adversarial counterparts. The advantage of introducing convex hull lies in it contains interaction information from isolate frames naturally.

From the locality perspective, the temporal local correlations are vital characteristics for videos. Specifically, there exists inherent correlations among adjacent frames (Tong et al. 2022), which are highly relevant with optical flow and demonstrate effectiveness in video action recognition (Wang et al. 2016; Zhao, Xiong, and Lin 2018; Wang et al. 2021a). This motivates us to eliminate the correlations among adjacent frames when optimizing adversarial perturbation for videos. Consequently, original temporal cues can be disrupted and the target video model fails to recognize videos from temporal patterns. Concretely, we optimize video adversarial examples by decreasing the similarities among intermediate features of adjacent adversarial frames.

Under the joint optimization of the above two constraints, video adversarial perturbation is finally optimized and can be directly injected onto benign video samples. The main contributions of this paper are summarized as follows.

- We focus on cross-modal attack from images to videos and propose an attack framework by fully considering the characteristics of video data from both global and local perspectives.
- We perform global inter-frame interactions in the attack process by extending a video clip to a convex hull and then stronger gradients are induced for effective adversarial perturbation optimization.
- We propose local correlation disturbance to disrupt the inherent temporal patterns in video adversarial examples, from which target black-box video models cannot capture valuable temporal clues.

Related Work

In this section, we firstly review several representative transfer-based attack methods in image domain. Then, the recent transfer-based attack methods in video domain is briefly introduced. At last, we present several video recognition and representation works which are closely related to our method and partially inspire us.

Transfer-based Image Attacks. Current adversarial attack methods usually improve the black-box transferability of adversarial examples from three aspects. Firstly, gradient-based methods achieve high transferability by introducing gradient enhancement strategies, such as momentum, variance tuning, which leads to stronger and more stable gradients for effective optimization of adversarial perturbation (Dong et al. 2018; Lin et al. 2020; Zhao, Liu, and Larson 2021; Wang and He 2021; Yuan et al. 2021; Xiong et al. 2022). Secondly, since the mid-level features extracted by different DNNs are similar to a certain extent, a couple of works aim to perturb the mid-level features of substitute model to fool various black-box models (Huang et al. 2019; Inkawhich et al. 2020; Salzmann et al. 2021). Thirdly, input augmentation can also facilitate the adversarial transferability (Xie et al. 2019; Lin et al. 2020).

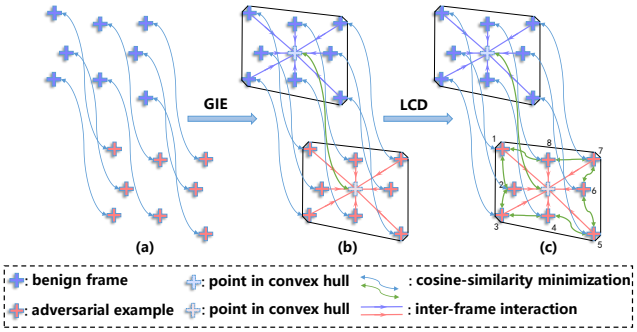


Figure 1: An illustration of proposed GIE and LCD. (a) our baseline, in which the cosine similarities among the intermediate features of the benign frames and adversarial frames are minimized separately. (b) GIE, in which a convex hull is spanned by isolated frames and each point in convex hull is pushed away from its adversarial counterpart. (c) GIE+LCD, in which the temporal local correlations of video adversarial example are disrupted. The numbers next to the marks indicate the temporal order.

Transfer-based Video Attacks. The study of the transfer-based attack methods on video recognition models are far less than transfer-based attack methods on image models. Temporal Translation (TT) (Wei et al. 2022a) is a representative method, which optimizes video adversarial examples over augmented video clips by temporal translation transformation. Most recently, based on the observation that intermediate feature space between images and video frames are somewhat similar, I2V (Wei et al. 2022b) uses image-based model as substitute model and achieve even better performance than methods with video-based substitute model, which is a promising finding.

Video Recognition and Representation. With the development of DNNs, video action recognition has made remarkable progress in recent years. In this task, temporal local correlations plays a key role in video discrimination and a series of works (Wang et al. 2016; Zhao, Xiong, and Lin 2018; Wang et al. 2021a) utilize this characteristic to improve video recognition. Specifically, (Wang et al. 2016; Zhao, Xiong, and Lin 2018) use RGB differences, which is closely related to temporal local correlations to represent motion information in videos. (Wang et al. 2021a) introduces a temporal difference module to capture both the short-term and long-term motion information. These methods motivate us to disrupt original video temporal differ-

ence when performing video attacks. Besides, our method is partial motivated by (Cevikalp and Triggs 2010; Zhu et al. 2013), in which an image set is modeled as an affine-hull or convex-hull. Different from these two works which focus on the distance metrics between two image sets represented by convex hulls to achieve an exact image set recognition, we focus on optimizing the distance between points lied on convex hull to generate adversarial examples with high transferability.

Methodology

Preliminary

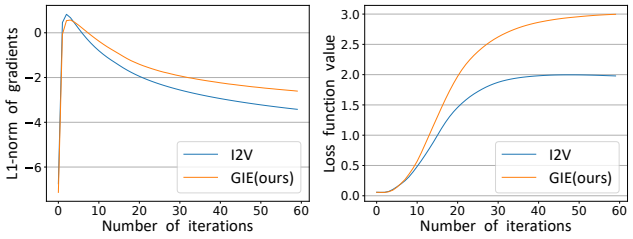
Given a substitute model g and a clean video clip $X = [x_1, x_2, ...x_T] \in \mathbb{R}^{T\times H\times W\times C}$ coupled with its label $y \in \{1, 2, ...K\}$, where T, H, W, C, K denote the number of frames, image height, image width, the number of channels and classes, respectively, we aim to optimize an adversarial perturbation $\Delta = [\delta_1, \delta_2, ..., \delta_T] \in \mathbb{R}^{T\times H\times W\times C}$ for X to generate a video adversarial example $X^{adv} = X + \Delta$, to fool the target video model $\mathcal{F}_t$ to make incorrect decisions, i.e., $\mathcal{F}_t(X^{adv}) \neq y$. Meanwhile, to ensure the imperceptibility of adversarial perturbation Δ, we constrain the ℓ_∞-norm of Δ to satisfy the inequality, i.e., $\|\Delta\|_\infty < \epsilon$, where ϵ is the maximum threshold, following most existing works (Xie et al. 2019; Dong et al. 2019; Wu et al. 2020; Wei et al. 2022a,b). In this paper, we consider the case that the target video model $\mathcal{F}_t$ is unknown and we can only use a substitute image model g pre-trained on ImageNet (Russakovsky et al. 2015), such as ResNet, VGG, to generate video adversarial examples.

Global Interaction Enhancement (GIE)

For video adversarial example generation, a straightforward scheme is to regard the input video as an image set and separately perturb each benign frame. Obviously, such way induces monotonous gradients because the interactions among the video frames has not been fully exploited. An intuitive physical implication of the inter-frame interactions is image blending. Taking image blending as an example, a natural idea is to endow the blended adversarial perturbations with the ability of disrupting the blended frames. In such way, adversarial perturbation for each frame fuses information from other frames then more robust and diverse gradients for optimizing may be induced and various black-box video models are likely to be fooled. To this end, we firstly extend a video clip to be a convex hull spanned by isolated frames and then focus on the adversarial perturbation optimization of each point lying on convex hull. In this way, inter-frame interaction is naturally introduced, which is vividly shown in Fig. 1. Formally, the convex hull of the video clip can be written as

$$H(X) = \left\{ \sum_{i=1}^{T} \alpha_i x_i \middle| \sum_{i=1}^{T} \alpha_i = 1, 0 \leq \alpha_i \leq 1 \right\}, \quad (1)$$

where α_i denotes the coefficient of the convex combination, x_i denotes the i-th frame in the clean video X. From Eq. 1, it can be seen that convex hull contains rich virtual



(a) (b)

Figure 2: (a) Gradient trends for I2V and our GIE. (b) Loss (trends are shown in log scale) on NL-101. Numerical values are computed on 400 randomly selected videos in Kinetics-400. The adversarial video examples are generated using VGG-16 as the substitute model and evaluated using NL-101 as the target model.

frames, i.e., blended frames, after fully interactions of isolated frames, which is more likely to creating diverse gradients. For the sake of the following formalization, we define a benign-adversarial pair set:

$$H(X, X^{adv}) = \left\{ \left(\sum_{i=1}^{T} \alpha_i x_i, \sum_{i=1}^{T} \alpha_i (x_i + \delta_i) \right) \middle| \sum_{i=1}^{T} \alpha_i = 1, 0 \leq \alpha_i \leq 1 \right\}, \quad (2)$$

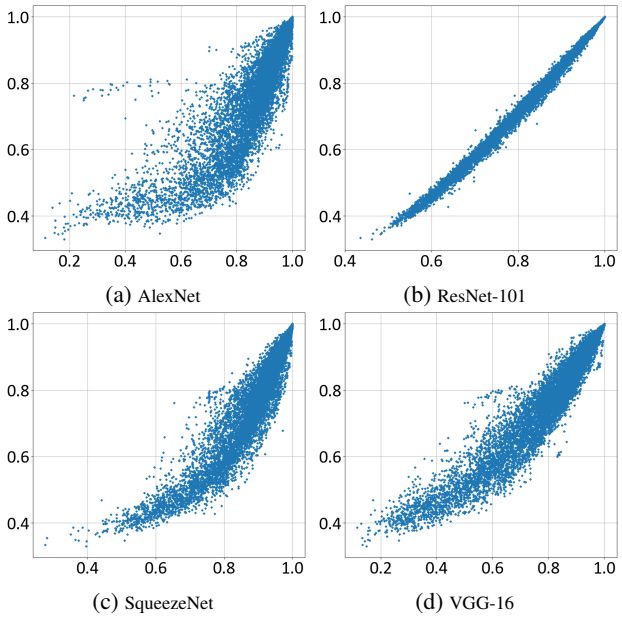
where δ_i denotes the i-th frame of the adversarial perturbation Δ. Then, the entire video adversarial perturbation is optimized by minimizing the cosine similarities among the intermediate features of the benign points lying on convex hull and their adversarial counterparts, which can be formulated as

$$\mathcal{L}_G = \mathbb{E}_{(u,v)\in H(X,X^{adv})} \left[\frac{g_l(u)^T g_l(v)}{||g_l(u)|| \cdot ||g_l(v)||} \right], \quad (3)$$

where $\mathbb{E}[\cdot]$ is the expectation w.r.t. the set of benign-adversarial pairs, $g_l(u)$ denotes the intermediate feature map w.r.t. u of l-th layer in model g, and $\|\cdot\|$ stands for the ℓ_2-norm of a vector. In practice, we treat adversarial perturbation of a video as integrity and optimized it entirely using above loss function.

To verify its effectiveness, a numerical analysis is conducted on the impact of the inter-frame interactions on the black-box transferability. Specifically, we present the evolution of ℓ_1-norm of the gradients and loss values on target video model over optimization process, in Fig. 2. It can observed that our method produces larger gradient magnitudes than I2V, resulting in efficient optimization and stronger adversarial examples. Besides, our method makes black-box target video model yield larger losses, which indicates a better attack performance and no over-fitting has happened.

Note that the existing I2V (Wei et al. 2022b) is a special case of Eq. 3, when the vector $[\alpha_1, \alpha_2, ..., \alpha_T]$ is in the form of one-hot. Besides, different from a trivially mixing within a batch of images (Zhang et al. 2018; Wang et al. 2021b), our method performs inter-frame interactions in the input



(a) AlexNet (b) ResNet-101 (c) SqueezeNet (d) VGG-16

Figure 3: Correlation scatter diagram of adjacent frame similarity based on four image models and one video model. The x axis and the y axis are adjacent similarity computed by image and video model, respectively. A narrower and longer ellipse indicates a higher degree of linear correlations. Numerical values are computed on 400 randomly selected videos from Kinetics-400.

video and aims to facilitate the adversarial perturbation optimization in a video-centric manner. In addition, Eq. 3 can be viewed as a kind of self-supervised signal especially developed for video-based adversarial perturbation generation.

Local Correlation Disturbance (LCD)

GIE boosts the transferability of video adversarial examples from the globality perspective. Here, our Local Correlation Disturbance aims to enhance the attack strength from the locality perspective of video data. Actually, a video can be viewed as a smooth extension of still images in the temporal dimension, and there exists inherent high correlations among the adjacent frames (Tong et al. 2022). It is such temporal local correlation that forms distinctive manifold structure of videos and differentiates videos from orderless image sets. Therefore, this characteristic of video data plays an essential role in different video understanding models (Wang et al. 2016; Zhao, Xiong, and Lin 2018; Ng and Davis 2018; Wang et al. 2021a). This observation raises an intuitive idea, i.e., can we disrupt the temporal local correlations of the video adversarial examples thus target video models fail to capture valuable temporal cues?

Unfortunately, in cross-modal attack, we have no access to video models. A question arises, is the temporal local correlation computed by image model and video model consistent with each other? Therefore, it is necessary to analyze the similarities of the temporal local correlations computed by the image and video models. Here, a numerical verification is conducted. For convenience, we simply use feature similarities of adjacent frames to measure the temporal local correlations numerically. The image models are four commonly used models, i.e., AlexNet, ResNet-101, SqueezeNet and VGG-16, which are trained on ImageNet (Russakovsky et al. 2015). The target video model is TPN-101 (Yang et al. 2020). It is selected because the frame number will not decrease in its inference process. Fig. 3 gives the adjacent similarity correlation scatter diagram based on image and video models. As can be observed, there exists a linearly proportional relationship between the local correlations computed from the image and video models. The averaged Pearson Correlation Coefficient (PCC)(Anderson 1962) value of adjacent similarities over 400 randomly selected videos in Kinetics-400 between four image models and video model are 0.9127, 0.9970, 0.9328 and 0.9412, which implies that the temporal local correlations computed by image and video models are somewhat similar. Based on the above observation, we can decrease the similarity among the adjacent frames by generating adversarial perturbation with respect to image-based substitute model g via the following loss function,

$$\mathcal{L}_L = \frac{1}{T-1}\sum_{i=1}^{T-1}\frac{g_l(x_i+\delta_i)^T g_l(x_{i+1}+\delta_{i+1})}{||g_l(x_i+\delta_i)||\cdot||g_l(x_{i+1}+\delta_{i+1})||}. \quad (4)$$

At last, we generate the adversarial perturbation for cross-modal attack via

$$\min_{\Delta} \mathcal{L}_T = \mathcal{L}_G + \lambda \mathcal{L}_L, \quad (5)$$

where λ controls the contribution of each term to the final objective. Following (Wei et al. 2019, 2022b), we initialize the video adversarial perturbation Δ with a small float

---

Algorithm 1: Global-Local Characteristic Excited Cross-Modal Attack.

Input: Video clip X, image model g, perturbation budget ϵ, number of iterations I, step size α, balanced factor λ, index of feature layer l

Output: Video adversarial example X^{adv}.

Initialize video adversarial perturbation: $\Delta_0 = (0.01/255)^{T\times H\times W\times C}$
for t $\in \{1, \cdots I\}$ **do**
  Construct benign-adversarial pair set $H(X, X^{adv})$ as Eq. 2
  Forward each tuple in $H(X, X^{adv})$ to obtain intermediate feature pair through g
  Calculate the GIE loss $\mathcal{L}_G$ as Eq. 3
  Calculate the LCD loss $\mathcal{L}_L$ as Eq. 4
  Update Δ_t over total loss $\mathcal{L}_T$ with Adam optimizer: $\Delta_t = ADAM(\Delta_{t-1}, \alpha, \mathcal{L}_T)$
  Project X_t^{adv} to the ϵ-ball of X^t: $X_t^{adv} = clip_{X,\epsilon}(X_{t-1}^{adv} + \Delta_t)$
end for
return X_I^{adv}

---

vector composed by value $0.01/255$ and leverage the Adam optimizer (Kingma and Ba 2014) to optimize Eq. 5.

Algorithm 1 summarizes the complete procedures of generation of video adversarial examples. Unlike I2V, our algorithm directly outputs a video adversarial example without temporal combination.

Experiments

Experimental Settings

Datasets. Two video recognition datasets, UCF-101 (Soomro, Zamir, and Shah 2012) and Kinetics-400 (Carreira and Zisserman 2017), are used for evaluations. UCF-101 contains 13,320 videos spanned over 101 actions. Kinetics-400 is a relatively large dataset which includes about 240,000 videos in 400 classes.

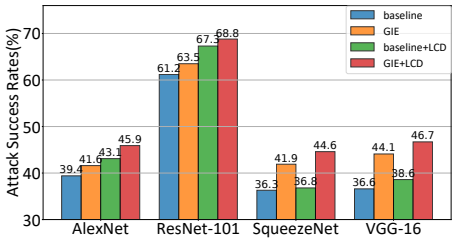
ImageNet-pretrained image models. Video adversarial examples are generated on four commonly used ImageNet-pretrained image models: AlexNet (Krizhevsky, Sutskever, and Hinton 2012), ResNet-101 (He et al. 2016), SqueezeNet 1.1 (Iandola et al. 2016) and VGG-16 (Simonyan and Zisserman 2015), which are widely used in the image related tasks.

Video recognition models. We choose three different video action recognition architectures, i.e., Non-local (NL) (Wang et al. 2018), SlowFast (Feichtenhofer et al. 2019), TPN (Yang et al. 2020), as our black-box models. NL, SlowFast and TPN all employ 3D ResNet-50/101 as the backbone. For the training strategy and input clip constructing scheme of video models, we use the same experimental settings as I2V (Wei et al. 2022b).

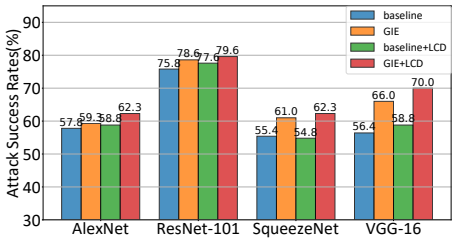
Attack setting. By following I2V, we randomly select one video per class, which is correctly classified by all black-box models, to conduct experiments. The attack performance is evaluated by the commonly used Attack Success Rate (ASR). A higher ASR indicates a better attack performance. For optimization strategy, we set the maximum perturbations ϵ as 16.0, step size α as 0.005, number of iterations I as 60, λ in Eq. 5 as 0.01. For the intermediate layer l in Eq. 3, we select $feature.7$ for AlexNet, $layer2$ for ResNet-101, $features.6.expand3 \times 3activation$ for SqueezeNet and $features.20$ for VGG-16, which is consistent with I2V.

Implementation Details

In our method, a convex hull is introduced for inter-frame interactions. Unfortunately, the points in Eq. 1 form a continuous region, which cannot be totally enumerated. In practice, we sample points uniformly from this contiguous region to simply this issue. The scale of sampling can be controlled by two key factors, i.e., the number of frames for interaction, n_1, and number of interaction combinations, n_2. $n_1 < T$ means that the convex combination coefficients of the rest $T - n_1$ frames are all zero. Besides, for the selected n_1 frames, we set the convex combination coefficients all equal to $1/n_1$ and do not consider the influence of coefficients in this paper. For number of interaction combinations n_2, we randomly sample them from the total $C_T^{n_1}$. In such way, the we can flexibly control the degree of frame interaction by setting these two hyper-parameters. In particular, our method degenerates to I2V when n_1=1. Furthermore, the intermediate features of the points in the benign convex hull are calculated in advance and saved into a lookup-table. When optimizing adversarial examples, we only need compute the intermediate features for the points in the adversarial convex hull online and query intermediate features for the points in the benign convex hull from lookup-table, which saves much of computational cost. In practice, we set n_1 as 2 and n_2 as 3.



(a) UCF-101



(b) Kinetics-400

Figure 4: The average attack success rates (%) against six black-box video recognition models.

Attack	UCF-101	Kinetics-400
baseline	36.63	56.42
GIE (n_1=2,n_2=2)	**45.05**	65.67
GIE (n_1=2,n_2=3)	44.06	**66.04**
GIE (n_1=2,n_2=4)	43.89	66.00
GIE (n_1=3,n_2=2)	42.90	64.46

Table 1: The average attack success rates (%) against six black-box video recognition models for various number of frames interacted, n_1 and number of interaction combinations, n_2. The adversarial video examples are crafted using VGG-16 as substitute model.

Attack	UCF-101	Kinetics-400
baseline	36.63	56.42
baseline+LCD (λ=0.1)	35.14	56.33
baseline+LCD (λ=0.01)	**38.61**	**58.75**
baseline+LCD (λ=0.001)	37.29	56.33

Table 2: The average attack success rates (%) against six black-box video recognition models for various λ. The adversarial video examples are crafted using VGG-16 as the substitute model.

Image Model	Attack	Black-box Video Model						
		NL-101	NL-50	SlowFast-101	Slow-Fast-50	TPN-101	TPN-50	Average
AlexNet	DR	29.70	26.73	19.80	25.74	8.91	13.86	20.79
	I2V	50.49	53.46	35.64	43.56	28.71	44.50	42.73
	I2V*(baseline)	48.51	53.47	31.68	38.61	24.72	39.60	39.43
	Ours	**53.47**	**62.38**	**37.62**	**46.53**	**29.70**	**45.54**	**45.87**
ResNet-101	DR	14.85	23.76	18.81	27.72	15.84	20.79	20.30
	I2V	71.28	60.39	50.49	57.42	61.38	71.28	62.04
	I2V*(baseline)	64.36	58.42	55.45	63.37	59.41	66.34	61.23
	Ours	**71.29**	**68.32**	**58.42**	**72.28**	**64.36**	**78.22**	**68.82**
SqueezeNet	DR	12.87	24.75	12.87	15.84	4.95	13.86	14.19
	I2V	43.56	54.45	36.63	37.62	23.76	35.64	38.61
	I2V*(baseline)	38.61	52.48	36.63	32.67	**27.72**	29.70	36.30
	Ours	**52.48**	**59.41**	**43.56**	**43.56**	26.73	**41.58**	**44.55**
VGG-16	DR	14.85	29.70	13.86	22.77	6.93	15.84	17.33
	I2V	43.56	48.51	28.71	39.60	21.78	32.67	35.81
	I2V*(baseline)	35.64	48.51	31.68	45.54	23.76	34.65	36.63
	Ours	**47.52**	**65.35**	**39.60**	**47.52**	**34.65**	**45.54**	**46.70**
Ensemble	I2V	71.28	76.23	56.43	62.37	52.47	75.24	65.67
	I2V*(baseline)	73.27	78.22	53.47	69.31	61.39	73.27	68.16
	Ours	**79.21**	**79.21**	**67.33**	**72.28**	**63.37**	**74.26**	**72.61**

Table 3: The attack success rates (%) against video recognition models on UCF-101. I2V*: results re-implemented by us.

Image Model	Attack	Black-box Video Model						
		NL-101	NL-50	SlowFast-101	Slow-Fast-50	TPN-101	TPN-50	Average
AlexNet	DR	22.00	31.50	43.00	41.75	31.00	39.00	34.71
	I2V	44.00	54.75	61.50	59.50	59.75	69.50	58.17
	I2V*(baseline)	44.50	54.75	60.50	60.25	58.00	69.00	57.83
	Ours	**51.00**	**61.00**	**65.75**	**62.25**	**62.00**	**71.50**	**62.25**
ResNet-101	DR	25.50	37.25	49.00	52.25	41.50	42.75	41.38
	I2V	56.25	64.50	74.75	77.00	87.25	90.25	75.00
	I2V*(baseline)	58.25	63.75	76.50	75.75	89.25	91.25	75.79
	Ours	**65.00**	**70.25**	**79.50**	**81.25**	**89.50**	**92.00**	**79.58**
SqueezeNet	DR	17.00	25.00	37.00	36.50	24.25	29.50	28.21
	I2V	37.75	51.00	62.50	60.25	55.50	58.50	54.25
	I2V*(baseline)	38.25	49.25	64.00	65.25	56.75	59.00	55.42
	Ours	**47.75**	**57.75**	**70.25**	**68.50**	**63.25**	**66.25**	**62.29**
VGG-16	DR	16.75	23.00	36.75	35.75	23.75	29.00	27.50
	I2V	39.00	46.25	57.75	59.00	59.00	70.50	55.25
	I2V*(baseline)	40.25	47.25	57.25	59.75	63.00	71.00	56.42
	Ours	**54.50**	**66.50**	**72.00**	**70.50**	**74.00**	**79.50**	**69.50**
Ensemble	I2V	65.00	72.25	79.75	76.50	85.75	88.00	77.88
	I2V*(baseline)	66.50	73.50	80.25	76.25	**87.75**	88.75	78.83
	Ours	**73.25**	**79.75**	**83.75**	**80.50**	86.50	**90.00**	**82.29**

Table 4: The attack success rates (%) against video recognition models on Kinetics-400. I2V*: results re-implemented by us.

Ablation Study

We firstly study the effectiveness of our proposed GIE and LCD for improving the cross-modal attack. Then a hyper-parameter sensitivity analysis is performed for each module. Our baseline is I2V (Wei et al. 2022b).

Evaluation of GIE. Using various substitute models, we evaluate our proposed GIE in terms of average ASR over six black-box video models. The results in Fig. 4 reveal that the average ASR is respectively improved by 2.2% $\sim$ 8.5% and 1.5% $\sim$ 9.6% on UCF-101 and Kinetics-400 with GIE. These remarkable improvements demonstrate that the inter-frame interaction can indeed boost the cross-modal transferability of adversarial examples. According to the results, our GIE prefers VGG and SqueezeNet, on which the gains are about 5.6% $\sim$ 8.5% and 5.6% $\sim$ 9.6% on UCF-101 and Kinetics-400, respectively.

Tab. 1 gives the average ASR of GIE with different number of frames in the interaction, n_1 and number of interaction combinations, n_2. It can observed that our GIE, i.e., $n_1 > 1$, clearly outperforms the baseline, i.e., $n_1 = 1$. Besides, increasing n_1 and n_2 can slightly boost the performance and then show a tendency to saturate. Since larger n_1 and n_2 will introduce heavy computational cost, we set $n_1 = 2$, $n_2 = 3$ to balance the computational overheads and ASR in the latter experiments.

Evaluation of LCD. To evaluate LCD, we add it to

the baseline for a better comparison. As shown in Fig. 4, LCD gives a positive effect on cross-modal transferability in most cases. In particular, using ResNet-101 as the substitute model, LCD outperforms the baseline with a margin of 6.1% on UCF-101, which suggests the vulnerability of video models against disrupting temporal local correlations. In the meantime, it gives unsatisfactory results when using SqueezeNet as the substitute model on Kinetics-400. This may result from that adjacent frame similarity computed by 2D ResNet-101 and 3D ResNet-101, backbone of video models, seems more linear correlated. On the contrary, there is a non-negligible discrepancy between SqueezeNet and 3D ResNet in terms of adjacent frame similarity as shown in Fig. 3. As mentioned in the previous evaluation, the proposed GIE works well on SqueezeNet, which can compensate for overall performance.

Tab. 2 shows the results of parameter sensitivity analysis of λ. LCD gives better performances when λ is slightly small and it performs less satisfactory when λ is relatively large. This is probably due to that too large λ may interfere with functionality of the baseline.

Assembling. Based on the above evaluations, GIE and LCD can both improve the cross-modal transferability. Moreover, from Fig. 4, it can be observed that the assembled method, GIE+LCD, improves baseline by a large margin of 6.5% $\sim$ 10.1% and 3.8% $\sim$ 13.6% on UCF-101 and Kinetics-400, which suggests that GIE and LCD can work compatibly and play a complementary role to each other. Therefore, we combine GIE with LCD to form our final framework to compare with the state-of-the-art methods.

Performance Comparisons

We compare our method with two state-of-the-art methods, i.e., DR (Lu et al. 2020) and I2V (Wei et al. 2022b). For fair comparison, we re-implement I2V with the released code provided by the authors. The results on UCF-101 and Kinetics-400 are shown in Tabs. 3 and 4 respectively where the performance of I2V reported in original paper is presented together with our implementation.

Tabs. 3 and 4 demonstrate that our method consistently outperforms the state-of-the-art methods. Especially for VGG-16, our method surpasses I2V by a margin of 10.07% and 13.08% on UCF-101 and Kinetics-400, respectively. These remarkable gains mainly come from the organic integration of GIE and LCD. Specifically, GIE boosts the cross-modal transferability by yielding more diverse and stronger gradients, while LCD disrupts the temporal local correlations of videos to expose the vulnerability of video models. They play a role from the global and local perspective respectively. Moreover, our method can obtain a better performance by the model ensemble technique and finally achieve an ASR of 82.29% on Kinetics-400, which demonstrates that our approach can be further enhanced via model ensemble.

Comparing against Stronger Baselines

We further compare our method with stronger baselines using video recognition models as substitute models. Concretely, we choose two representative competitors, i.e., MI-FGSM (Dong et al. 2018) and TAP (Zhou et al. 2018), which perform better according to the results in I2V (Wei et al. 2022b). For these competitors, TPN-101 is employed as the substitute model due to the superior performance shown in (Wei et al. 2022b). Besides, we leverage ILAF (Huang et al. 2019) to fine-tune the generated adversarial examples for them and re-implement all the competitors with the released codes provided by their authors. For our method, we newly introduce DenseNet-121 (Huang et al. 2017) as a candidate substitute model.

Attack	substitute model	Kinetics-400
MI-FGSM	TPN-101	52.50
MI-FGSM w/ fine-tuning	TPN-101	81.06
TAP	TPN-101	67.75
TAP w/ fine-tuning	TPN-101	82.63
I2V	DenseNet-121	81.69
I2V+LCD (ours)	DenseNet-121	82.00
GIE (ours)	DenseNet-121	82.63
GIE+LCD (ours)	DenseNet-121	84.00
GIE+LCD (ours)	DenseNet-121+ResNet-101	**85.50**

Table 5: The average ASR (%) comparison of our method and stronger baselines. The average ASR is calculated by averaging ASRs over black-box video models which have different architectures from the substitute model.

The comparison results are demonstrated in Tab. 5. We have two observations from the results. Firstly, the attack performance of our method, i.e., GIE+LCD, with single DenseNet-121 can approximate the attack performance of TAP with fine-tuning, which is the strongest competitor in the table and finally surpass it using an ensemble of DenseNet-121 and ResNet-101 as the substitute model on Kinetics-400. Secondly, the state-of-the-art method I2V is weaker than TAP with fine-tuning. Besides, the attack performance of I2V can be further improved by our proposed LCD, which verifies the effectiveness of LCD. In general, these comparison results verify the advantage of our method in video attack task, i.e., under the premise that no loss in attack performance, our method is more superior in economy because training image models is cheaper than training video models.

Conclusion

In this work, we propose a new cross-modal attack method from images to videos, in which global and local video characteristic is considered. Specifically, inter-frame interaction is naturally introduced by extending a video clip to be a convex hull thus stronger gradients are induced for enhancing transferability of video adversarial examples. For another, local correlation disturbance is proposed to make video recognition models fail to capture valid temporal cues. Extensive experiments show the effectiveness of proposed two modules respectively. More importantly, it is shown that security of video recognition models is seriously threatened by simple attack methods using even plain image models as substitute models. Therefore, in the future, we plan to investigate the robustness of video recognition models and develop defense algorithms for evading common attacks.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62272020 and in part by the Fundamental Research Funds for Central Universities.

References

Anderson, T. W. 1962. An Introduction to Multivariate Statistical Analysis. Technical report, Wiley New York.

Carlini, N.; and Wagner, D. 2017. Towards Evaluating the Robustness of Neural Networks. In *IEEE Symposium on Security and Privacy*, 39–57.

Carreira, J.; and Zisserman, A. 2017. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6299–6308.

Cevikalp, H.; and Triggs, B. 2010. Face Recognition Based on Image Sets. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2567–2573.

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

Dong, Y.; Pang, T.; Su, H.; and Zhu, J. 2019. Evading Defenses to Transferable Adversarial Examples by Translation-invariant Attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4312–4321.

Feichtenhofer, C.; Fan, H.; Malik, J.; and He, K. 2019. Slowfast Networks for Video Recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6202–6211.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2014. Explaining and Harnessing Adversarial Examples. In *International Conference on Learning Representations*.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision*, 2961–2969.

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

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

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

Iandola, F. N.; Han, S.; Moskewicz, M. W.; Ashraf, K.; Dally, W. J.; and Keutzer, K. 2016. SqueezeNet: AlexNet-level Accuracy with 50x Fewer Parameters and¡ 0.5 MB Model Size. *arXiv preprint arXiv:1602.07360*.

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

Jiang, L.; Ma, X.; Chen, S.; Bailey, J.; and Jiang, Y.-G. 2019. Black-box Adversarial Attacks on Video Recognition Models. In *Proceedings of the ACM International Conference on Multimedia*, 864–872.

Kim, D.; Woo, S.; Lee, J.-Y.; and Kweon, I. S. 2019. Deep Video Inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5792–5801.

Kingma, D. P.; and Ba, J. 2014. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, 1097–1105.

Kurakin, A.; Goodfellow, I.; Bengio, S.; et al. 2017. Adversarial Examples in the Physical World. In *International Conference on Learning Representations (workshop)*.

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

Lu, Y.; Jia, Y.; Wang, J.; Li, B.; Chai, W.; Carin, L.; and Velipasalar, S. 2020. Enhancing Cross-Task Black-Box Transferability of Adversarial Examples with Dispersion Reduction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 940–949.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2017. Towards Deep Learning Models Resistant to Adversarial Attacks. *arXiv preprint arXiv:1706.06083*.

Ng, J. Y.-H.; and Davis, L. S. 2018. Temporal Difference Networks for Video Action Recognition. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1587–1596.

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

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; and Bernstein, M. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, (3): 211–252.

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

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

Soomro, K.; Zamir, A. R.; and Shah, M. 2012. UCF101: A Dataset of 101 Human Actions Classes from Videos in the Wild. *arXiv preprint arXiv:1212.0402*.

Tong, Z.; Song, Y.; Wang, J.; and Wang, L. 2022. VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training. *arXiv preprint arXiv:2203.12602*.

Wang, L.; Tong, Z.; Ji, B.; and Wu, G. 2021a. TDN: Temporal Difference Networks for Efficient Action Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1895–1904.

Wang, L.; Xiong, Y.; Wang, Z.; Qiao, Y.; Lin, D.; Tang, X.; and Gool, L. V. 2016. Temporal Segment Networks: Towards Good Practices for Deep Action Recognition. In *Proceedings of the European Conference on Computer Vision*, 20–36.

Wang, Q.; Zhang, L.; Bertinetto, L.; Hu, W.; and Torr, P. H. 2019. Fast Online Object Tracking and Segmentation: A Unifying Approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1328–1338.

Wang, X.; Girshick, R.; Gupta, A.; and He, K. 2018. Non-local Neural Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7794–7803.

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

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

Wang, Z.; Sha, C.; and Su, Y. 2021. Reinforcement Learning Based Sparse Black-box Adversarial Attack on Video Recognition Models. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 955–962.

Wei, X.; Yan, H.; and Li, B. 2022. Sparse Black-box Video Attack with Reinforcement Learning. *International Journal of Computer Vision*, (6): 1459–1473.

Wei, X.; Zhu, J.; Yuan, S.; and Su, H. 2019. Sparse Adversarial Perturbations for Videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 01, 8973–8980.

Wei, Z.; Chen, J.; Wei, X.; Jiang, L.; Chua, T.-S.; Zhou, F.; and Jiang, Y.-G. 2020. Heuristic Black-box Adversarial Attacks on Video Recognition Models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 07, 12338–12345.

Wei, Z.; Chen, J.; Wu, Z.; and Jiang, Y.-G. 2022a. Boosting the Transferability of Video Adversarial Examples via Temporal Translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 3, 2659–2667.

Wei, Z.; Chen, J.; Wu, Z.; and Jiang, Y.-G. 2022b. Cross-Modal Transferable Adversarial Attacks from Images to Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15064–15073.

Wei, Z.; Chen, J.; Zhang, H.; Jiang, L.; and Jiang, Y.-G. 2022c. Adaptive Temporal Grouping for Black-box Adversarial Attacks on Videos. In *Proceedings of the International Conference on Multimedia Retrieval*, 587–593.

Wu, H.; Chen, Y.; Wang, N.; and Zhang, Z. 2019. Sequence Level Semantics Aggregation for Video Object Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9217–9225.

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

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

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14983–14992.

Yang, C.; Xu, Y.; Shi, J.; Dai, B.; and Zhou, B. 2020. Temporal Pyramid Network for Action Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 591–600.

Yuan, Z.; Zhang, J.; Jia, Y.; Tan, C.; Xue, T.; and Shan, S. 2021. Meta Gradient Adversarial Attack. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7748–7757.

Zhang, H.; Cisse, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2018. Mixup: Beyond Empirical Risk Minimization. In *International Conference on Learning Representations*.

Zhang, H.; Zhu, L.; Zhu, Y.; and Yang, Y. 2020. Motion-excited Sampler: Video Adversarial Attack with Sparked Prior. In *Proceedings of the European Conference on Computer Vision*, 240–256.

Zhao, Y.; Xiong, Y.; and Lin, D. 2018. Recognize Actions by Disentangling Components of Dynamics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6566–6575.

Zhao, Z.; Liu, Z.; and Larson, M. 2021. On Success and Simplicity: A Second Look at Transferable Targeted Attacks. In *Advances in Neural Information Processing Systems*, 6115–6128.

Zhou, W.; Hou, X.; Chen, Y.; Tang, M.; Huang, X.; Gan, X.; and Yang, Y. 2018. Transferable Adversarial Perturbations. In *Proceedings of the European Conference on Computer Vision*, 452–467.

Zhu, P.; Zhang, L.; Zuo, W.; and Zhang, D. 2013. From Point to Set: Extend the Learning of Distance Metrics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2664–2671.