Adaptive Cross-Modal Transferable Adversarial Attacks from Images to Videos

Zhipeng Wei, Jingjing Chen, Zuxuan Wu, Yu-Gang Jiang

Abstract—The cross-model transferability of adversarial examples makes black-box attacks to be practical. However, it typically requires access to the input of the same modality as black-box models to attain reliable transferability. Unfortunately, the collection of datasets may be difficult in security-critical scenarios. Hence, developing cross-modal attacks for fooling models with different modalities of inputs would highly threaten real-world DNNs applications. The above considerations motivate us to investigate cross-modal transferability of adversarial examples. In particular, we aim to generate video adversarial examples from white-box image models to attack video CNN and ViT models. We introduce the Image To Video (I2V) attack based on the observation that image and video models share similar low-level features. For each video frame, I2V optimizes perturbations by reducing the similarity of intermediate features between benign and adversarial frames on image models. Then I2V combines adversarial frames together to generate video adversarial examples. I2V can be easily extended to simultaneously perturb multi-layer features extracted from an ensemble of image models. To efficiently integrate various features, we introduce an adaptive approach to re-weight the contributions of each layer based on its cosine similarity values of the previous attack step. Experimental results demonstrate the effectiveness of the proposed method.

Index Terms—Cross-modal attack, Transferable attack.

1 Introduction

Adversarial examples have been shown to be security vulnerabilities in deep neural networks (DNNs) [1], [2]. It raises numerous research attention to the security of DNN applications over the past few years, especially for security-critical scenarios, such as autonomous driving [3], face recognition [4], video analysis [5], etc. To attain a high attack success rate, they typically require prior knowledge about the attacked model, such as the structure and parameters. In this way, they optimize perturbations guided by the gradients of the loss function with respect to the inputs. However, prior knowledge may be unavailable in the real-world applications. To address this problem, recent studies resort to the transferability of adversarial examples, which means that an adversarial example crafted from one white-box surrogate model has the ability to attack other black-box models with unknown structures and parameters [6]. Current works aim to mitigate the over-fitting of generated adversarial examples to the surrogate model. Specifically, they either optimize perturbations on a set of augmented images [7], [8], [9], incorporate the momentum term or nesterov accelerated gradient into the gradient calculations [8], [10], or corrupt the critical features shared among different models, such as the low-level features [11], the critical features [12], [13]. However, the above methods suffer from the restriction that the surrogate models and attacked models are trained on the same modality of inputs. In contrast, rather less attention has been paid to the transferability between hetero-modal models.

The cross-modal transferability enables the collection and annotation of datasets for other modalities to be discarded, making it easy to attack black-box models trained on different input modalities. However, existing transferable attacks proposed in homo-modal models are unsuitable in the cross-modal scenario. This is because these attacks typically require the label information to calculate the classification loss so as to optimize perturbations, and there are no shared labels between datasets with different modalities. Besides, there exists a domain gap between different datasets, especially for datasets with different modalities. Hence, the modality difference prevents existing attacks from attaining satisfactory performance in cross-modal transferability. Above analysis are especially true for images and videos, which are widely used data modalities. There exists a domain gap and additional temporal information between image and video data, which yields differences in learned features between image and video models. Therefore, in this paper, we mainly investigate the cross-modal transferability of adversarial examples from image models to video models, with the purpose of utilizing image models pre-trained on ImageNet to generate video adversarial examples, which can fool video Convolutional Neural Networks (CNN) and Vision Transformer (ViT) models with high probability.

In light of the fact that video models benefit from bootstraping parameters from the ImageNet-pretrained image models [14], [15], we conduct an empirical analysis to find that the intermediate features of video frames between image and video models are similar to a certain extent. Motivated by this observation, we propose a simple yet effective cross-modal attack method, named Image To Video (I2V) attack. Specifically, I2V optimizes perturbations of

• *Z. Wei, J. Chen, Z. Wu and Y. Jiang are with the Shanghai Key Lab of Intelligent Information Processing, School of Computer Science, Fudan University and the Shanghai Collaborative Innovation Center on Intelligent Visual Computing.*
E-mail: zpwei21@m.fudan.edu.cn, {chenjingjing, zxwu, ygj}@fudan.edu.cn

Manuscript received xxxxx xx, xxxx; revised xxxxxx xx, xxxx.

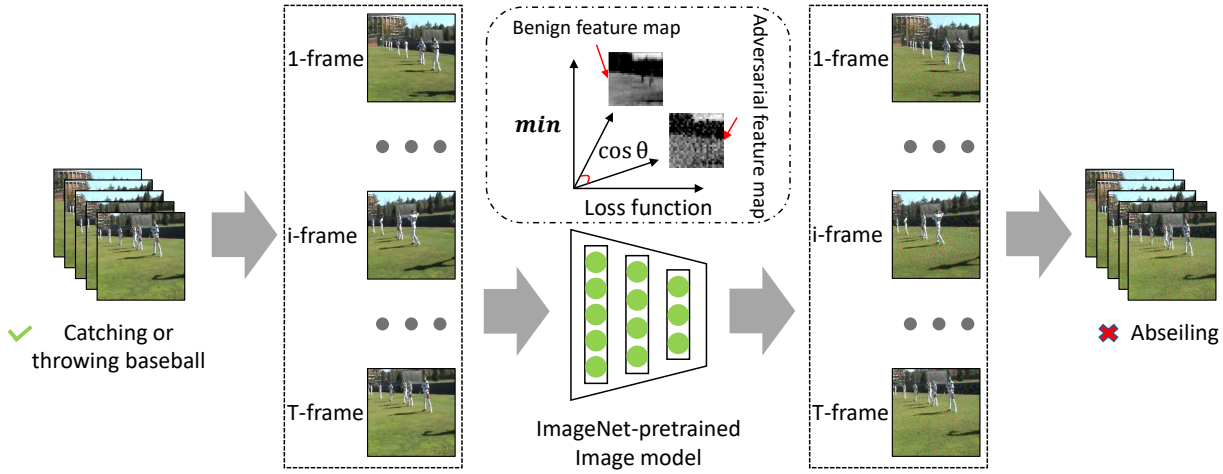


Figure 1: Overview of the proposed I2V attack. Given a video clip with a true label of "Catching or throwing baseball", where each frame is input into the ImageNet-pretrained image model separately. Then the image model generates adversarial frames by minimizing the cosine similarity between features from adversarial and benign examples. As the image model and the video model share similar feature space, the generated video adversarial example can fool the video recognition models, and be misclassified as "Abseiling".

each frame by minimizing the cosine similarity of the intermediate features extracted from image models between benign and adversarial frames. Afterwards, I2V includes all crafted adversarial frames together to generate video adversarial examples for attacking video models. Figure 1 gives an overview of I2V. Given a video clip, I2V first split the clip into multiple frames, and then adopts each frame as input separately. Afterwards, I2V disturbs intermediate features to generate adversarial frames, which are finally combined together into video adversarial examples.

To further improve adversarial transferability, we extend the I2V attack into perturbing multi-layer features of the ensemble image models. This is because, multi-layer features can better represent the feature space of the model compared to single-layer features [16], and the adversarial examples of fooling the ensemble models may transfer to attack other black-box models with high probability [10]. In addition, we further propose an adaptive approach that assigns weights to each layer instead of treating them equally. Our paper shows the feasibility of cross-modal adversarial attacks, which induce security concerns in video models. We briefly summarize our primary contributions as follows:

- We propose a novel attack, named Image to Video (I2V), to improve the cross-modal transferability of adversarial examples from image models to video models. Through building bridges between image and video models, I2V enables the image model to serve as the surrogate model to evaluate video model robustness.
- We discover that image and video models share similar intermediate features. Hence, perturbing features of image models would highly affect video features.
- To further boost the cross-modal transferability, we propose several variants of I2V by simultaneously perturbing multi-layer features on an ensemble of models and devise an adaptive method to re-weight the contribution of each layer.
- We provide extensive experiments to demonstrate the effectiveness of the proposed I2V attack and the adaptive method. It suggests that both video CNN and ViT models suffer from adversarial examples crafted by image CNNs.

A preliminary version of this paper appeared in [17]. The present paper includes a complete review of literatures on video recognition ViTs; a new adaptive method, of assigning weights to each layer in multi-layers attacks so as to further improve the attack success rate; new comparisons of the adaptive version of I2V attack on multi-layers; new comparisons of existing transfer-based attacks with the proposed method for video ViTs; a new visualization of weight changes for the adaptive method, demonstrating the effectiveness of the proposed method.

2 Related Work

2.1 Transfer-based Attacks on Image Models

Prior works in generating adversarial examples with high transferability are based on white-box attacks, such as Fast Gradient Sign Method (FGSM) [2] and Basic Iterative Method (BIM) [18]. FGSM linearizes the loss function around the current parameters and performs a one-step update along with the gradient sign of the loss function with respect to inputs. BIM is the iterative version of FGSM and overfits the white-box model to generate stronger adversarial examples in attacking the white-box model. To further improve the transferability of adversarial examples in attacking black-box models, several approaches are proposed recently. In general, there are three ways to improve transferability, which include data augmentation, gradient modification and common property disruption of discriminative features among different models. The main idea of data augmentation is to improve the generalization of adversarial examples and avoid overfitting the white-box model. For example, Diversity Input (DI) [7] attack

conducts random resizing and padding to the input. Scale-invariant method (SIM) [8] applies the scale transformation to the input. Translation-invariant (TI) [9] attack performs horizontal and vertical shifts with a short distance to the input. The second way modifies gradients used for updating adversarial perturbations. For example, Momentum Iterative (MI) [10] attack integrates the momentum into the iterative process for stabilizing update directions. As an improved momentum method, Nesterov Accelerated Gradient (NAG) [8] can be also integrated into the BIM. Skip Gradient Method (SGM) [11] uses more gradients from the skip connections and emphasizes the gradients of shallow layers. The main idea of the third way focuses on disrupting the common property of classification among different models. For example, Attention-guided Transfer Attack (ATA) [12] prioritizes the corruption of critical features that are likely to be adopted by diverse architectures. Other transfer-based attacks such as Dispersion Reduction (DR) [19], Intermediate Level Attack (ILA) [20] have improved the transferability of adversarial examples in different tasks through perturbing feature maps. In contrast, the proposed I2V attack implements a cross-modal transfer-based attack through a correlation between the spatial features encoded between the image model and the video model.

2.2 Transfer-based Attacks on Video Recognition Models

There is much less work about transfer-based attacks on video models compared with transfer-based attacks on image models. Temporal Translation (TT) attack method [21] optimizes the adversarial perturbations over a set of temporal translated video clips for avoiding overfitting to the white-box model being attacked. Although TT achieves better results than transfer-based image attack methods, it increases the computational cost. Different from it, the proposed I2V attack achieves better performance without trained video models and is easy to perform.

2.3 Video Recognition CNNs

Video action recognition CNNs have made significant progress in recent years. Previous studies [22], [23] adopt a 2D + 1D paradigm, where 2D CNNs are applied over per-frame input to extract features, followed by a 1D module (*e.g.*, RNNs) that integrates per-frame features. Current studies use 3D CNNs to jointly capture the dynamic semantics of videos. For example, I3D [14] leverages ImageNet architecture designs and their parameters to encode spatio-temporal features by inflating the 2D convolution kernels into 3D. Non-local (NL) [15] network inserts a non-local operation into I3D for encoding long-range temporal dependencies between video frames. SlowFast [24] contrasts the visual tempos along the temporal axis, which involves a slow pathway and a fast pathway to capture spatial semantics and motion at fine temporal resolution respectively. Temporal Pyramid Network (TPN) [25] capture action instances at various tempos through a feature hierarchy architecture. In this paper, we use six representative video action recognition CNNs for experiments, including NL, SlowFast, TPN with 3D Resnet-50 and Resnet-101 as backbones.

2.4 Video Recognition ViTs

The vision transformer [26] is the first work that adopts the transformer architecture in computer vision. This considerable success of ViTs on images [27], [28] inspire researches to develop video ViTs [29], [30], [31], [32], [33], [34]. Video Transformer Network (VTN) [29] utilizes Longformer [35] to encode temporal features that are extracted by any given 2D spatial backbone. With the local-context self-attention and task-specific global attention of Longformer, VTN can perform whole video prediction. TimeSformer [33] replaces the convolution operation with the proposed 3D self-attention mechanism on space and time, in order to eliminate the inductive bias of convolution operation and capture the long-range dependencies. ViViT [30] is also a pure-transformer based model and consists of four variants of factorising the transformer encoder, which can improve the effectiveness of processing numerous tokens. MViT [31] fuses multiscale features by the proposed pooling attention operation. Video Swin [32] extends the windows of Swin [28] from 2D to 3D for globally encoding patches over space and time. Moreover, Motionformer [34] proposes the trajectory attention for connecting the temporal information among videos. In this paper, we use VTN, TimeSformer, Motionformer, and Video Swin to conduct experiments.

3 METHODOLOGY

3.1 Preliminary

Given a video sample $x \in \mathcal{X} \subset \mathbf{R}^{T\times H\times W\times C}$ with the true label $y \in \mathcal{Y} = \{1, 2, ..., K\}$, where T, H, W, C denote the number of frames, height, width and channels respectively. K represents the number of classes. Let g denote the ImageNet-pretrained image model (*e.g.*, ResNet, VGG), f denote the video recognition model. We use $f(x) : \mathcal{X} \to \mathcal{Y}$ to denote the prediction of the video recognition model for an input video. Thus, the proposed I2V attack aims to generate the adversarial example $x_{adv} = x + \delta$ by g, which can fool the video model f into $f(x_{adv}) \neq y$ without knowledge about f, where δ denotes the adversarial perturbation. To ensure that the adversarial perturbation δ is imperceptible, we restrict it by $||\delta||_p \leq \epsilon$, where $||\cdot||_p$ denotes the L_p norm, and ϵ is a constant of the norm constraint. We adopt L_∞ norm and untargeted adversarial attacks, which are commonly used in [7], [8], [9], [11], [12]. In a white-box setting, the objective of untargeted adversarial attacks can be formulated as follows:

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

where J is the loss function (*e.g.*, cross-entropy loss) of the video model f. However, in this paper, the adversary cannot access knowledge about f. The proposed I2V attack leverages adversarial examples generated from g to attack f in the black-box setting.

3.2 Correlation analysis between image and video models

Before introducing the proposed method, we firstly give an empirical analysis of the correlation between image and video models. It has been demonstrated in the prior work

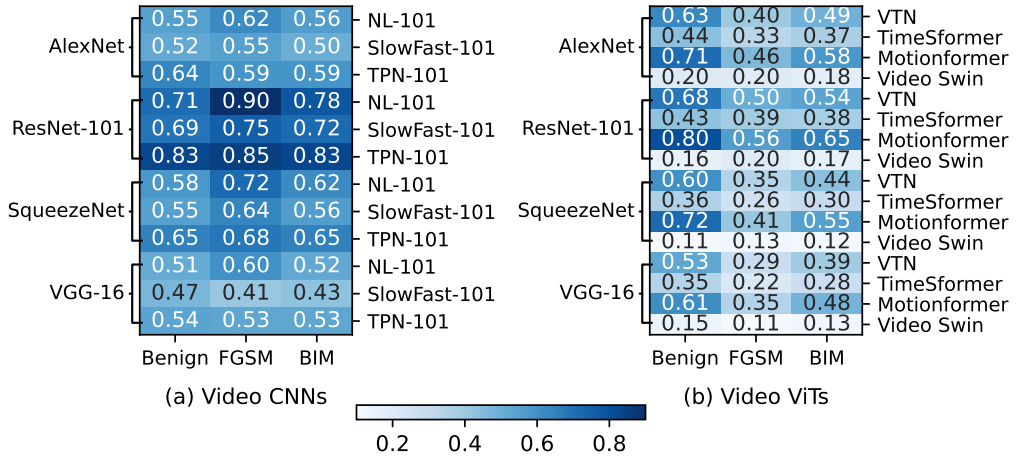
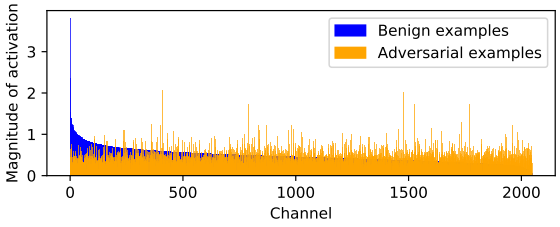


Figure 2: Centered Kernel Alignment (CKA) analysis on intermediate features extracted from two scenarios: (a) image and video CNNs and (b) image CNNs and video ViTs. Adversarial video examples are generated by FGSM and BIM, using video CNNs and ViTs as white-box models. Darker color represents a higher feature similarity. CKA is calculated based on minibatches [38]. The batch size is set to 12.

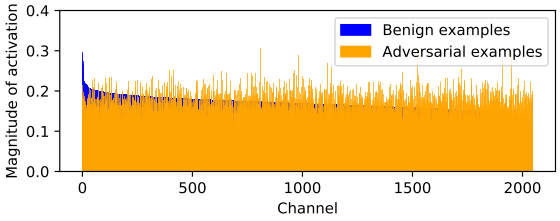
[36] that utilizing ImageNet-pretrained image models to generate tentative perturbations assumes fewer queries for attacking black-box video recognition models. This basically suggests that *the intermediate features between image models and video models may be similar to a certain extent*. Hence perturbing intermediate feature maps of image models could affect that of video models. To verify this assumption, we analyze the feature similarity of both benign and adversarial frames between image and video models with Centered Kernel Alignment (CKA) [37], which is used to compare feature similarity between different models [38], [39].

Figure 2 shows CKAs of intermediate features between image and video models. For all video models, the intermediate features are extracted from the first block, while for different image models, the features are extracted from different intermediate layers, which are summarized in Table 1 (marked in **bold**). Here we choose different intermediate layers for different image models for the purpose of maximizing the similarity between image features and video features. From Figure 2, we observe that the average CKAs of benign examples and adversarial examples crafted by FGSM and BIM are $[0.60, 0.65, 0.60]$ and $[0.46, 0.32, 0.37]$ for video CNNs and Video ViTs, respectively. It suggests that for both benign and adversarial samples, their intermediate layer features extracted from image and video models are similar to a certain extent (CKA>0.30). This is primarily attributable to the intrinsic resemblance in spatial information inherent to images and videos. Besides, the average standard deviation of benign and adversarial examples for video CNNs is 0.03, while for video ViTs, it obtains a slightly higher value at 0.06. It indicates the stability of feature similarity in the presence of adversarial perturbations.

To demonstrate that the adversarial perturbations on the feature maps are transferable between video and image models, we further compare the magnitude changes in the channel-wise activation of image and video models before and after adding the same adversarial perturbations to the video frames. As shown in Figure 3, the adversarial



(a) Resnet-101



(b) NL-101

Figure 3: The magnitudes (y-axis) of channel-wise activation at the penultimate layer (2048 channels at x-axis) for both image and video models. The magnitude is calculated by the global average pooling at each channel. In each plot, the channel-wise magnitudes are averaged from randomly selected 400 videos of Kinetics-400 and displayed separately for benign and adversarial examples (generated by BIM). The 2048 channels are sorted in descending order of magnitude for benign examples.

examples generated on NL-101 perturb the channel-wise activation magnitude not only in NL-101 but also in Resnet-101. Since each channel of features captures a specific pattern of the object and contributes differently to the final classification, the magnitude changes of image and video models are likely to result in wrong predictions, demonstrating possibilities in transferring adversarial perturbations between image and video models. Moreover, we visualize

Figure 4: The visualization of benign and adversarial feature maps on image models. This adversarial example is generated by NL-101.

Model	Layer 1	Layer 2	Layer 3	Layer 4
Alexnet	ReLU-1	ReLU-2	**ReLU-3**	ReLU-5
Resnet-101	Block-1	**Block-2**	Block-3	Block-4
Squeezenet	Fire-1	**Fire-3**	Fire-5	Fire-8
Vgg-16	ReLU-1	ReLU-5	**ReLU-9**	ReLU-13

Table 1: Intermediate layer selection. For each image model, the selected intermediate layer used for crafting adversarial perturbations in the proposed I2V attack is in **bold**.

the spatial feature maps of image models in Figure 4 before and after adding perturbations crafted by video models. This reveals that the spatial features become corrupted with noise when combined with the perturbations, suggesting that the perturbations generated by the video model have the potential to impair spatial information.

3.3 Image To Video (I2V) Attack

Based on the above observations, we propose the Image To Video (I2V) attack, which generates video adversarial examples from an ImageNet-pretrained image model, to boost the transferability of hetero-modal models and attack video models in the black-box setting. By perturbing intermediate features of image models, I2V generates adversarial examples to disturb intermediate features of black-box video models with high probability. In particular, I2V optimizes the i-th adversarial frame by:

$$\arg\min_{\delta} CosSim(g_l(x^i+\delta), g_l(x^i)), s.t.\ ||\delta||_\infty \leq \epsilon, \quad (2)$$

where $g_l(x^i)$ denotes the intermediate feature map of the l-th layer with respect to x^i in the image model, $x^i \in \mathbf{R}^{H\times W\times C}$ denotes the i-th frame of x, the function $CosSim$ calculates the cosine similarity between $g_l(x^i+\delta)$ and $g_l(x^i)$.

In this way, the minimization of the cosine similarity makes it possible to optimize adversarial examples with features that are orthogonal to ones of benign examples. Consider that $g_l(x^i)$ is the output of the penultimate layer and let $W = (W_1, ..., W_y, ..., W_K)$ denote the weight of the classification layer, thus W_y and $g_l(x^i)$ are highly aligned for making a true prediction. With minimizing $CosSim(g_l(x^i+\delta), g_l(x^i)) = \frac{g_l(x^i+\delta)^T g_l(x^i)}{||g_l(x^i+\delta)||\cdot||g_l(x^i)||}$, we can get minimizing $g_l(x^i+\delta)^T g_l(x^i)$ if $g_l(x^i+\delta)$ and $g_l(x^i)$ have unit length. Due to the high alignment between W_y and $g_l(x^i)$, the minimization of the cosine similarity induces that the value of $W_y \cdot g_l(x^i+\delta)$ decreases a lot to fool the image model g into making error predictions. Based on the similarity of feature space between image and video models, the generated adversarial examples $x_{adv} = (x^1_{adv}, ..., x^i_{adv}, ..., x^T_{adv})$ may fool video models with high probability by perturbing video intermediate features.

Algorithm 1 Image to Video (I2V) attack.

Input: A video example x, image model g.
Parameter: Perturbation budget ϵ, iteration number I, step size α, the number of layer l.
Output: The adversarial example x_{adv}.

for $i = 1$ to T **do**
  x^i = the i-th frame of x
  $\delta^i_0 = (\frac{\mathbf{0.01}}{\mathbf{255}})^{H\times W\times C}$
  for $j = 0$ to $I-1$ **do**
    Update δ^i_j by Adam optimizer:
    $\delta^i_{j+1} = ADAM(\delta^i_j, \alpha, CosSim(g_l(x^i+\delta^i_j), g_l(x^i)))$
    Project x^i_{adv} to the ϵ-ball of x^i:
    $x^i_{adv} = clip_{x^i,\epsilon}(x^i + \delta^i_I)$
  end for
end for
return $x_{adv} = (x^1_{adv}, ..., x^i_{adv}, ..., x^T_{adv})$

Following [40], we initialize the adversarial perturbations δ with a small constant value $\frac{0.01}{255}$ and use the Adam optimizer [41] to solve the Equation 2 and updates δ^i_j. Algorithm 1 illustrates the generation of adversarial examples of the proposed I2V attack. Where I denotes the iteration number of Adam optimizer, $clip_{x^i,\epsilon}$ denote to project $x^i+\delta^i_I$ to the vicinity of x^i for meeting $||\delta^i_I||_\infty \leq \epsilon$. In the end, I2V attack combines all generated adversarial frames x^i_{adv} into a video adversarial examples x_{adv}.

3.4 Multi-layer Features Attack

Fusing multi-layer features shows significant improvement on several computer vision tasks [25], [42]. This method motivates us to develop the I2V Multi-layer Features (I2V-MF) attack to perturb multi-layer features of image models simultaneously. Thus, rather than minimizing the cosine similarity with a single-layer feature as the Equation 2, the I2V-MF attack optimizes adversarial examples with multi-layer features as

$$\arg\min_{\delta} \sum_{l\in L} CosSim(g_l(x^i+\delta), g_l(x^i)), s.t.\ ||\delta||_\infty \leq \epsilon, \quad (3)$$

where L is the perturbed layers. With this method, the generated adversarial examples impair the information of shallow and deep layers to further improve transferability.

3.5 Ensemble models Attack

MIFGSM [10] shows that attacking an ensemble of models can boost the transferability of generated adversarial examples. When a generated example remains adversarial over an ensemble of models, it may transfer to attack other models. Based on this, we propose to use multiple ImageNet-pretrained image models to perform the I2V attack, named ENS-I2V, which optimizes i-th adversarial frame by:

$$\arg\min_{\delta} \sum_{n=1}^{N} CosSim(g^n(x^i+\delta), g^n(x^i)), s.t.\ ||\delta||_\infty \leq \epsilon, \quad (4)$$

where N is the number of used image models, $g^n(\cdot)$ returns the intermediate features of the l-th layer in the n-th image

model. The intermediate features of the adversarial frames generated by ENS-I2V are orthogonal to the ensemble of features from benign examples, thus ENS-I2V allows the generation of highly transferable adversarial examples. In addition, we can easily combine the ENS-I2V attack and the multi-layer features attack, named ENS-I2V-MF. Its objective function is defined as

$$\arg\min_{\delta} \sum_{n=1}^{N} \sum_{l \in L} CosSim(g_l^n(x^i + \delta), g_l^n(x^i)), s.t.\ ||\delta||_\infty \leq \epsilon, \tag{5}$$

3.6 Adaptive Ensemble models Attack

Each image model in the ENS-I2V attack equally contributes to the optimization of the Equation 4. However, due to the different feature representations of image models, the cosine similarity values of the adversarial image features to the benign image features decrease at different rates (as illustrated in Figure 11 and discussed in Section 4.7). This finding means that an image model with higher cosine similarity values should have relatively higher weights to get a more rapid decrease in its cosine similarity values. Hence, we consider an adaptive approach to generate weights for the ensemble models. The adaptive ensemble models attack (AENS-I2V) assigns a weight for each image model to optimize the adversarial example as

$$\arg\min_{\delta} \sum_{n=1}^{N} w_n CosSim(g^n(x^i + \delta), g^n(x^i)), s.t.\ ||\delta||_\infty \leq \epsilon, \tag{6}$$

where w_n is the weight for the n-th model. To emphasize the models with higher cosine similarity values, we use the cosine similarity values of the previous attack step to adaptively generate weights at the s-th attack step:

$$w_n = \frac{e^{CosSim_{s-1}^n}}{\sum_{n=1}^{N} e^{CosSim_{s-1}^n}}, \tag{7}$$

where $CosSim_{s-1}^n$ denotes the cosine similarity value of the n-th model at the $(s-1)$-th attack step. In this way, the AENS-I2V attack assigns higher weights to low optimized image models in each attack step. To further improve the adversarial transferability, we can also combine the AENS-I2V attack and the multi-layer features attack, named AENS-I2V-MF. It generates weights for each layer at s-th step:

$$w_n^l = \frac{e^{CosSim_{s-1}^{n,l}}}{\sum_{n=1}^{N} \sum_{l \in L} e^{CosSim_{s-1}^{n,l}}}, \tag{8}$$

where w_n^l is the weight for the layer l of the n-th model, $CosSim_{s-1}^{n,l}$ denotes the cosine similarity value of the layer l of the n-th model at the $(s-1)$-th attack step. Thus, the objective function of AENS-I2V-MF is defined as:

$$\arg\min_{\delta} \sum_{n=1}^{N} \sum_{l \in L} w_n^l CosSim(g_l^n(x^i + \delta), g_l^n(x^i)), \quad s.t.\ ||\delta||_\infty \leq \epsilon. \tag{9}$$

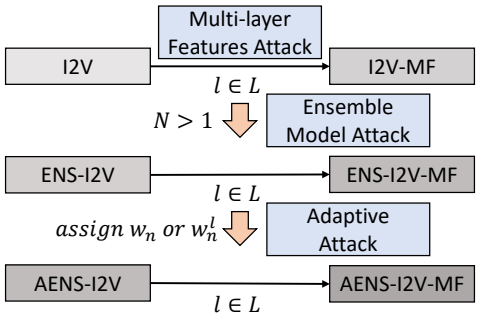


Figure 5: Connections between I2V attacks. By introducing the multi-layer features attack module, ensemble model attack module and adaptive attack module, we can obtain various versions of the I2V attack.

Model	UCF-101	Kinetics-400
NL-50	81.26	75.17
NL-101	82.21	75.81
SlowFast-50	85.25	76.66
SlowFast-101	86.10	76.95
TPN-50	87.13	78.90
TPN-101	90.28	79.70

Table 2: Top-1 validation accuracy(%) of video CNNs on UCF-101 and Kinetics-400.

3.7 Connections between I2V attacks

Figure 5 summarizes connections between I2V attacks:

- If the perturbed layer l belongs to the set of perturbed layers L, I2V will evolve into I2V-MF, ENS-I2V will evolve into ENS-I2V-MF, and AENS-I2V will evolve into AENS-I2V-MF.
- If the number of image models N is larger to 1, I2V will evolve into ENS-I2V, and I2V-MF will evolve into ENS-I2V-MF.
- If weights w_n are assigned to each model, ENS-I2V will evolve into AENS-I2V. If weights w_n^l are assigned to each layer per model, ENS-I2V-MF will evolve into AENS-I2V-MF.

4 Experiments

4.1 Experimental Settings

4.1.1 Dataset

We evaluate our approach using UCF-101 [43] and Kinetics-400 [14] datasets, which are widely used datasets for video recognition. UCF-101 consists of 13,320 videos from 101 actions. Kinetics-400 contains approximately 240,000 videos from 400 human actions.

4.1.2 ImageNet-pretrained Image Models

We perform our proposed methods on four ImageNet-pretrained image models: Alexnet [44], Resnet-101 [45], Squeezenet 1.1 [46] and Vgg-16 [47]. Where Squeezenet 1.1 has 2.4x less computation and slightly fewer parameters than SqueezeNet 1.0, without sacrificing accuracy. These four models are commonly used in the image classification.

Model	Temporal Stride		
	4	8	12
VTN	67.07	72.62	75.63
TimeSformer	56.69	64.03	67.78
Motionformer	70.95	75.09	77.29
Video Swin	69.44	72.49	74.34

Table 3: Top-1 validation accuracy(%) of video vision transformers on Kinetics-400 with 4/8/12 temporal strides.

4.1.3 Video Recognition Models

CNN Models. Our proposed methods are evaluated on three different architectures of video recognition models: Non-local (NL) [15], SlowFast [24], TPN [25]. NL, SlowFast and TPN use 3D Resnet-50/101 as the backbone. We train these video models from scratch with Kinetics-400 and fine-tune them on UCF-101. For Kinetics-400, we skip every other frame from randomly selected 64 consecutive frames into constructing input clips. For UCF-101, we use 32 consecutive frames as input clips. The spatial size of the input is 224×224. Table 2 summarizes top-1 validation accuracy of these six models on UCF-101 and Kinetics-400.

Video Vision Transformers. We evaluate adversarial examples generated from image CNNs on four video vision transformers: VTN [29], TimeSformer [33], Motionformer [34], Video Swin [32]. In the testing phase, we sample one clip per video and each clip consists of 16 frames with $4/8/12$ temporal strides. The spatial size of clips is 224×224. Table 3 shows top-1 validation accuracy of video ViTs on Kinetics-400. We observe that introducing more dynamic cues (a larger temporal stride) leads to higher validation accuracy. However, these accuracies are lower than the ones reported in their works. This is because they average predictions from multiple clips per video. In addition, the temporal strides of inputs among these models are different. To make these ViTs share the same inputs, we set the temporal stride to $4/8/12$ in this paper. These differences together lead to degraded performance.

4.1.4 Attack Setting

In our experiments, we use the Attack Success Rate (ASR) to evaluate the attack performance, which is the rate of adversarial examples that are successfully misclassified by the black-box video recognition model. Thus higher ASR means better adversarial transferability. If not specifically stated, average ASR (AASR) is the average ASR over all black-box video models. Following [7], [9], we randomly sample one video, which is correctly classified by video CNNs or ViTs, from each class to conduct our experiments, and set the norm constraint $\epsilon = 16$.

4.2 Ablation study

We first investigate the effects of step size α, iteration number I and different attacked layers l in the I2V attack. The evaluations are conducted on video CNNs trained on Kinetics-400. These optimized parameters are directly used to attack other video models.

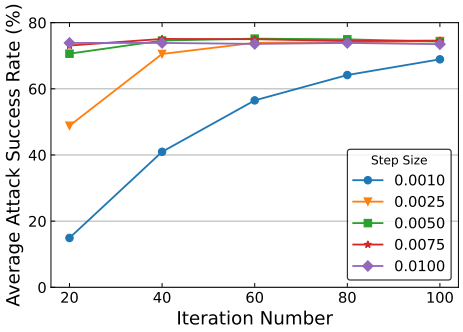


Figure 6: AASR (%) of the I2V attack with various step sizes and iteration numbers.

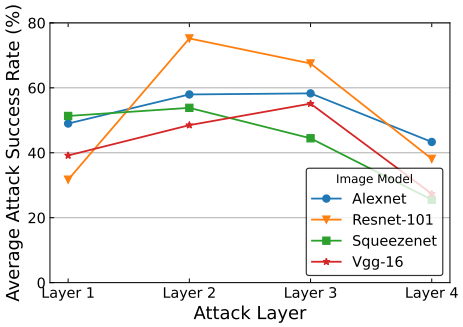


Figure 7: AASR (%) of the I2V attack with various attack layers l in different image models.

4.2.1 Step Size and Iteration Number

Equation 2 is solved by the Adam optimizer, which can be affected by the step size α and iteration number I. Figure 6 shows the results of using Block-2 of Resnet-101 as the perturbed layer of the image model with different step sizes and iteration numbers. It can be seen that smaller α and I have poorer AASR because of under-fitting. While larger α can achieve better AASR with a smaller I. To achieve the best performance, we adopt $\alpha = 0.005$ and $I = 60$ in subsequent experiments.

4.2.2 Intermediate Layer Selection

For each image model, we select four layers from bottom to top layer (as shown in Table 1) to craft the adversarial perturbations. Figure 7 shows the results of attacking different layers. Attacking the middle layers (layer 2 or layer 3) of image models is better than attacking the bottom or top layers. Based on the results, we attack the middle layers, which are marked in **bold** in Table 1 for each model. Besides, due to the higher performance of middle layers, we attack layer 2 and 3 for the multi-layer features attack.

4.3 Performance comparison

Since transferability between hetero-modal models has never been explored, we compare our proposed I2V and its variants with DR [19], which is originally proposed to enhance cross-task transferability. DR minimizes the standard deviation of intermediate features to degrade the recognizability of images. We extend DR to optimize adversarial

Attack	Image Model	Black-box Video Model					
		NL-101	NL-50	SlowFast-101	SlowFast-50	TPN-101	TPN-50
DR [19]	Alexnet	29.70	26.73	19.80	25.74	8.91	13.86
	Resnet-101	14.85	23.76	18.81	27.72	15.84	20.79
	Squeezenet	12.87	24.75	12.87	15.84	4.95	13.86
	Vgg-16	14.85	29.70	13.86	22.77	6.93	15.84
I2V	Alexnet	50.49	53.46	35.64	43.56	28.71	44.50
	Resnet-101	71.28	60.39	50.49	57.42	61.38	71.28
	Squeezenet	43.56	54.45	36.63	37.62	23.76	35.64
	Vgg-16	43.56	48.51	28.71	39.60	21.78	32.67
I2V-MF	Alexnet	55.44	53.46	37.62	45.54	32.67	45.54
	Resnet-101	73.26	72.27	50.49	64.35	73.26	74.25
	Squeezenet	46.53	56.43	29.70	37.62	21.78	30.69
	Vgg-16	52.47	48.51	33.66	45.54	23.76	42.57
ENS-I2V	Ensemble	71.28	76.23	56.43	62.37	52.47	75.24
AENS-I2V	Ensemble	79.20	80.19	61.38	66.33	61.38	**78.21**
ENS-I2V-MF	Ensemble	84.15	**86.13**	61.38	71.28	70.29	77.22
AENS-I2V-MF	Ensemble	**86.13**	85.14	**64.35**	**75.24**	**71.28**	77.22

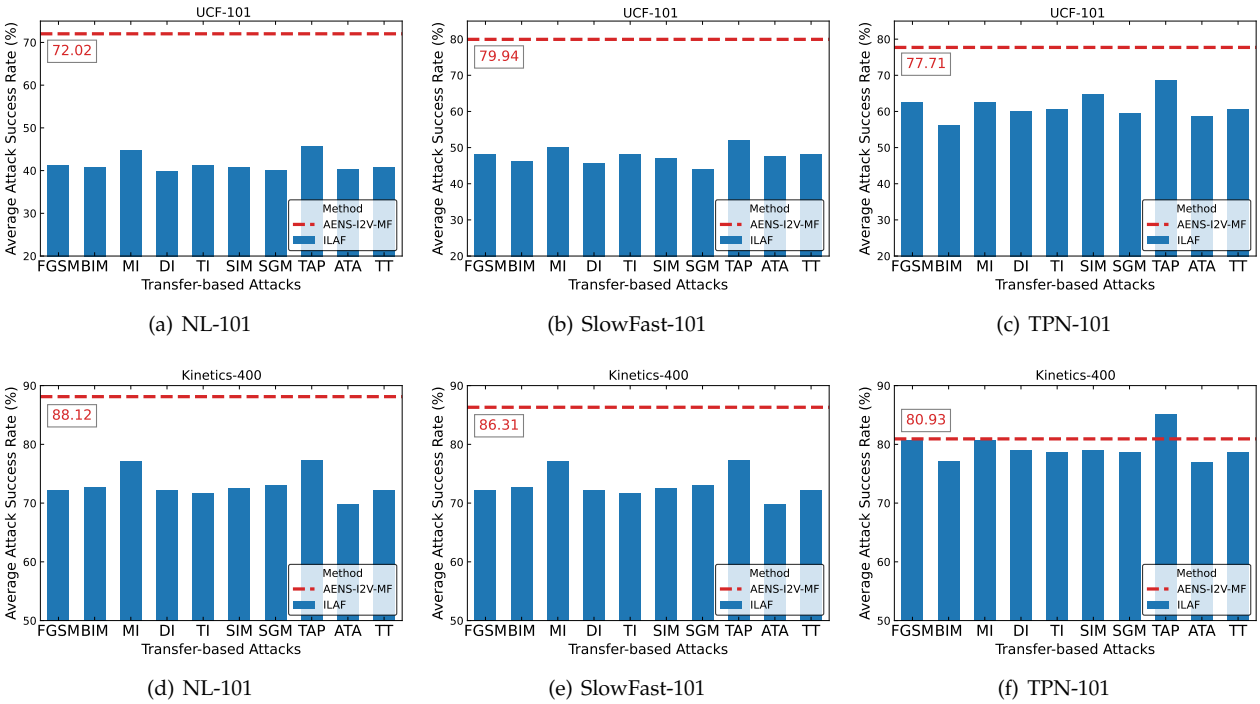
Table 4: ASR (%) against video recognition models on UCF-101.

Attack	Image Model	Black-box Video Model					
		NL-101	NL-50	SlowFast-101	SlowFast-50	TPN-101	TPN-50
DR [19]	Alexnet	22.00	31.50	43.00	41.75	31.00	39.00
	Resnet-101	25.50	37.25	49.00	52.25	41.50	42.75
	Squeezenet	17.00	25.00	37.00	36.50	24.25	29.50
	Vgg-16	16.75	23.00	36.75	35.75	23.75	29.00
I2V	Alexnet	44.00	54.75	61.50	59.50	59.75	69.50
	Resnet-101	56.25	64.50	74.75	77.00	87.25	90.25
	Squeezenet	37.75	51.00	62.50	60.25	55.50	58.50
	Vgg-16	39.00	46.25	57.75	59.00	59.00	70.50
I2V-MF	Alexnet	45.25	57.25	64.50	64.25	61.50	71.25
	Resnet-101	63.00	70.00	78.25	76.50	**94.50**	92.00
	Squeezenet	37.75	44.00	61.75	59.00	56.75	62.25
	Vgg-16	39.75	50.00	59.75	62.75	66.25	73.00
ENS-I2V	Ensemble	65.00	72.25	79.75	76.50	85.75	88.00
AENS-I2V	Ensemble	69.25	73.50	82.75	80.00	89.25	91.25
ENS-I2V-MF	Ensemble	76.25	79.00	83.00	80.50	90.50	92.75
AENS-I2V-MF	Ensemble	**79.00**	**79.25**	**84.50**	**81.00**	93.50	**93.50**

Table 5: ASR (%) against video recognition models on Kinetics-400.

Attack	Image Model	Black-box Video Model			
		VTN	TimeSformer	Motionformer	Video Swin
DR	Alexnet	19.75 / 15.00 / 13.50	19.75 / 16.00 / 12.75	11.25 / 8.75 / 10.50	19.25 / 15.75 / 16.75
	Resnet-101	19.75 / 15.00 / 13.75	26.75 / 19.50 / 18.75	9.00 / 8.25 / 9.00	20.75 / 17.00 / 17.00
	Squeezenet	11.75 / 7.50 / 6.50	13.25 / 8.25 / 8.00	6.25 / 4.75 / 5.25	18.00 / 15.00 / 16.00
	Vgg-16	11.50 / 7.25 / 7.00	12.25 / 11.75 / 7.75	6.00 / 5.00 / 5.75	19.75 / 15.25 / 14.75
I2V	Alexnet	43.50 / 36.00 / 34.25	36.25 / 27.75 / 26.25	25.75 / 20.75 / 21.75	36.50 / 33.00 / 33.75
	Resnet-101	32.75 / 28.50 / 26.25	26.75 / 20.50 / 19.50	17.75 / 14.50 / 15.75	34.75 / 31.25 / 32.50
	Squeezenet	17.25 / 13.75 / 10.25	15.50 / 11.25 / 10.75	9.50 / 7.50 / 9.00	30.00 / 26.75 / 31.00
	Vgg-16	31.25 / 26.50 / 23.25	24.25 / 16.75 / 18.50	17.75 / 12.25 / 14.00	47.25 / 40.75 / 40.75
I2V-MF	Alexnet	43.75 / 35.00 / 35.50	37.50 / 30.25 / 26.50	26.25 / 22.50 / 22.25	38.25 / 35.00 / 37.00
	Resnet-101	37.00 / 29.50 / 28.25	31.00 / 23.75 / 22.00	22.75 / 16.75 / 15.00	38.00 / 30.50 / 33.50
	Squeezenet	20.00 / 14.75 / 13.25	18.25 / 13.00 / 14.25	10.50 / 7.50 / 7.75	32.75 / 27.25 / 30.25
	Vgg-16	32.00 / 25.25 / 22.75	24.50 / 17.25 / 17.25	18.50 / 12.50 / 13.75	46.00 / 41.50 / 42.00
ENS-I2V	Ensemble	53.50 / 49.00 / 48.50	39.00 / 32.75 / 30.50	36.75 / 30.75 / 28.50	56.25 / 52.25 / 51.50
AENS-I2V	Ensemble	54.00 / 49.50 / 51.00	35.25 / 29.25 / 29.75	**39.50** / 33.50 / 31.75	55.50 / 51.50 / 50.25
ENS-I2V-MF	Ensemble	57.00 / **52.00** / 50.25	42.00 / 37.50 / **37.00**	38.25 / **35.75** / **33.50**	**61.00** / **59.00** / **56.25**
AENS-I2V-MF	Ensemble	**58.25** / 51.75 / **51.50**	**43.75** / **38.25** / 35.25	39.25 / **35.75** / **33.50**	60.50 / 58.00 / **56.25**

Table 6: ASR (%) against video vision transformers with 4/8/12 temporal strides on Kinetics-400.



(a) NL-101 (b) SlowFast-101 (c) TPN-101

(d) NL-101 (e) SlowFast-101 (f) TPN-101

Figure 8: AASR (%) against video recognition models with fine-tuning attack methods. The top and bottom rows are the results on UCF-101 and Kinetics-400 respectively. The three columns use the NL-101, SlowFast-101, and TPN-101 models as white-box models respectively. For each white-box model, we average ASR over black-box video CNNs that have a different architecture to obtain AASR. Red dashed lines denote the performance of our proposed AENS-I2V-MF.

examples on the ImageNet-pretrained image models and use the same Adam optimizer and other settings as I2V.

4.3.1 Attacking Video CNNs

The results of attacking UCF-101 and Kinetics-400 datasets on video CNNs are shown in Table 4 and 5, respectively. From the results, we have the following observations. First, the proposed attacks achieve much higher ASR than DR by a large margin. For example, compared with DR, I2V can boost AASR of more than 63.33% and 42.51% for UCF-101 and Kinetics-400 separately. Second, I2V and I2V-MF using Resnet-101 as the white-box image model outperforms all the other I2V attacks, which suggests that 2D Resnet101 and 3D Resnet-101 in the backbone of video models share more similar feature space than other 2D image models. Third, ENS-I2V further improves the average AASR to 65.68% against UCF101 and 77.88% against Kinetics-400. Which demonstrates the validity of attacking an ensemble of image models. Fourth, by introducing the multi-layer features attack, I2V-MF, ENS-I2V-MF and AENS-I2V-MF achieve higher performance than I2V, ENS-I2V and AENS-I2V, respectively. It indicates that perturbing multi-layer features helps to further destruct the visual information and improve adversarial transferability. Lastly, AENS-I2V and AENS-I2V-MF achieve better performance than ENS-I2V and ENS-I2V-MF, respectively. It suggests that assigning larger weights to low optimized layers can further improve performance. Finally, AENS-I2V-MF, which includes multi-layer features attack, ensemble model attack and adaptive attack, achieves almost the best results.

In general, our method, which considers minimizing the cosine similarity between features from adversarial and benign examples, consistently outperforms DR. These experiments validate the effectiveness of the proposed attacks.

4.3.2 Attacking Video ViTs

Table 6 exhibits the results of attacking the Kinetics-400 dataset on video ViTs. As can be seen, the results are gradually improved with the introduction of our proposed methods. In particular, AENS-I2V-MF incorporating multiple attack methods works best in attacking VTN, Motionformer and Video Swin. While for attacking TimeSformer, ENS-I2V-MF performs slightly better than AENS-I2V-MF. Despite these similar trends, there are some differences between attacking video CNNs and ViTs. First, it is easier to attack video CNNs than ViTs using image CNNs. The reason is that both video CNNs and image CNNs share the basic convolutional structure. However, the proposed AENS-I2V-MF still achieves an average attack success rate of 46.83% despite the structural difference. Second, the I2V attack using Resnet-101 as the white-box model in attacking video CNNs is superior to that in attacking video ViTs, because video CNNs use 3D Resnet-101 as the backbone network. Third, video ViTs with larger temporal strides exhibit lower ASR values. Specifically, when utilizing AENS-I2V-MF to attack VTN, the temporal stride of 12 yields a degradation of 6.75 in ASR compared to the temporal stride of 4. In general, our methods consistently outperform DR attacks. These results indicate the vulnerability of video ViTs against adversarial examples generated from image CNNs.

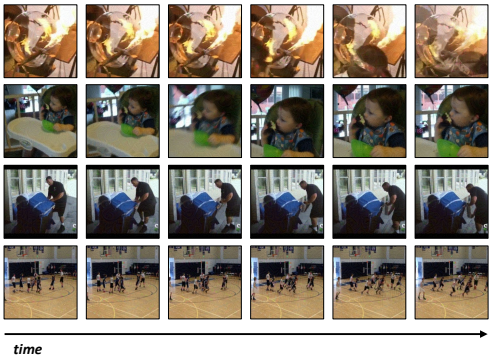


Figure 9: Visualization of randomly picked adversarial clips, crafted by the proposed ENS-I2V. Labels from top row to bottom row are "blowing glass", "eating cake", "moving furniture", and " playing basketball" separately.

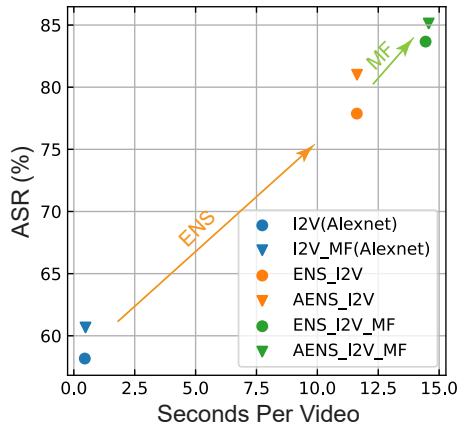


Figure 10: Seconds per video and attack success rate (ASR) on Kinetics-400 of our methods.

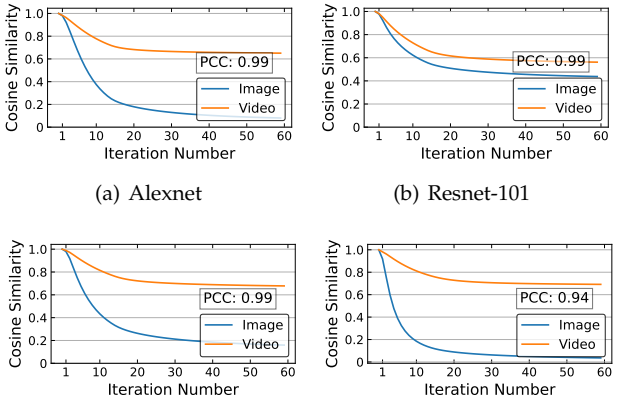
4.4 Comparing against stronger baselines

We further compare the proposed AENS-I2V-MF attack against several existing transfer-based attacks that are designed for homomodal models (e.g., images models or video models) on video CNNs. It's worthwhile to mention that the comparisons are unfair because the existing transfer-based attacks require the white-box video recognition models to generate the adversarial perturbations. For the comparison, several transfer-based attacks, such as FGSM [2], BIM [18], MI [10], DI [7], TI [9], SIM [8], SGM [11], TAP [16], ATA [12], and TT [21] are used as baselines. For these baselines, NL-101, SlowFast-101 and TPN-101 are used as the white-box models. It has been illustrated in ILA [20], the transferability of generated adversarial examples can be further enhanced through the proposed fine-tuning methods ILAP and ILAF. Compared to ILAP, ILAF achieves better performance by maintaining the existing adversarial direction and increases the magnitude of feature perturbations under L_2 norm [20]. Therefore, for the compared baseline methods, we use ILAF to fine-tune the generated adversarial examples.

Figure 8 shows the comparison results. We have the following observations. First, despite the comparisons being unfair since our method does not require any white-box video models, the proposed AENS-I2V-MF still performs much better than ILAF in most cases. As shown in Figure 8(a)(b)(d)(e), on both UCF-101 and Kinetics-400, AENS-I2V-MF exceeds ILAF by a large margin. Second, AENS-I2V-MF performs worse than TAP when using TPN-101 as the white-box model on Kinetics-400 (Figure 8(f)). This may be because that Kinetics-400 contains richer motion information than UCF-101 and such motion information is unlikely to be well captured by image models. On the contrary, by fusing multi-layer features, TPN-101 can better capture the motion information. As a result, disrupting motion information (Figure 8(f)) can achieve better performances.

4.5 Computational Efficiency

We report the seconds per video and attack success rate (%) on Kinetics-400 of our methods in Figure 10. The ensemble method (the yellow arrow) and the multi-layer (the green arrow) method result in a notable improvement in the attack success rate, however, they require more time to back-propagate gradients from multi-layers or multiple models. Besides, the proposed adaptive method does not introduce any additional back-propagation operations, thereby maintaining the same level of computational complexity.



(a) Alexnet (b) Resnet-101

(c) Squeezenet (d) Vgg-16

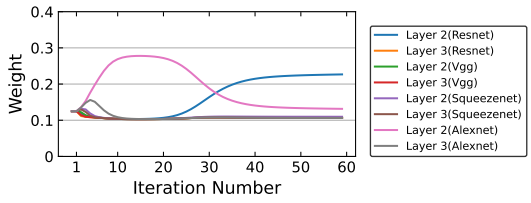
Figure 11: Pearson correlation coefficient (PCC) analysis between cosine similarity trends computed from image models and NL-101.

4.6 Visualization of Adversarial Examples

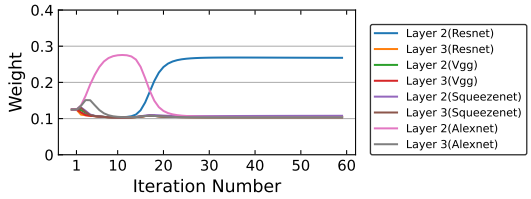
We further visualize 4 randomly selected adversarial clips in Figure 9. These adversarial examples are generated on the ensemble of ImageNet-pretrained models (Alexnet, Resnet, Squeezenet, Vgg) by the proposed ENS-I2V attack. These adversarial examples do not affect human decision-making but fool video models into wrong predictions.

4.7 Discussion

To experimentally demonstrate the effectiveness of the optimized object function (Equation 2), we investigate the changes in the cosine similarity of the adversarial image/video features to the benign image/video features by increasing the iteration number. Pearson Correlation Coefficient (PCC) [48] is used to measure the linear correlation



(a) UCF-101



(b) Kinetics-400

Figure 12: Weight change curves of AENS-I2V-MF.

of cosine similarity trends computed from image and video models. Figure 11 shows the PCC analysis of cosine similarity trends using 4 image models and a video model (NL-101). As can be seen, all PCCs are close to 1, which implies an exact positive linear relationship between the directional changes of image and video intermediate features. It suggests that minimizing the cosine similarity of image models can make intermediate features of video adversarial examples generated from ImageNet-pretrained models orthogonal to their benign video features. In addition, there are different cosine similarity gaps between image and video features in different image models. Specifically, the cosine similarity gap is smaller for Resnet-101 and larger for Vgg-16, which is caused by the similar architecture between Resnet-101 and 3D Resnet-101 used in NL-101. It suggests that image models may contribute differently to the perturbation optimization. Thus, we propose the adaptive method for assigning weights to each layer.

To further improve the comprehension of the adaptive method, we also visualize the weight change curves of each layer in Figure 12. From the figure, we observe that the weight of layer 2 of Alexnet increases first and then decreases with the increase of the iteration number. This observation suggests that the proposed adaptive method effectively minimizes the loss of this layer by assigning a larger weight to it. Besides, the weights of all layers plateau at the later stage. It demonstrates that the relative magnitude of losses in all layers is essentially constant.

5 Conclusion

In this paper, we identify the existence of a similar feature space between image and video models, which can be leveraged to generate adversarial examples from image models to attack black-box video models. More specifically, we proposed the Image To Video (I2V) attack, which optimizes adversarial frames on the ImageNet-pretrained image model by minimizing the cosine similarity between features from adversarial and benign examples for perturbing intermediate feature space. Besides, we proposed the multi-layer features attack, the ensemble models attack, and the adaptive method attack with an aim to further improve adversarial transferability. These attacks can be combined to form the more powerful attack method, named AENS-I2-MF. The results indicate that cross-modal adversarial transferability occurs even across image and video domains. In the future, we will combine temporal information of videos into image models to further boost transferability.

Acknowledgments

This project was supported by National Key R&D Program of China (No. 2021ZD0112804) and in part by NSFC project (#62032006), Science and Technology Commission of Shanghai Municipality Project (20511101000). Y.-G. Jiang was sponsored in part by "Shuguang Program" supported by Shanghai Education Development Foundation and Shanghai Municipal Education Commission (No. 20SG01).

References

[1] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.

[2] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[3] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter, "Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition," in *Proceedings of the 2016 acm sigsac conference on computer and communications security*, 2016, pp. 1528–1540.

[4] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song, "Robust physical-world attacks on deep learning visual classification," in *CVPR*, 2018, pp. 1625–1634.

[5] Z. Wei, J. Chen, X. Wei, L. Jiang, T.-S. Chua, F. Zhou, and Y.-G. Jiang, "Heuristic black-box adversarial attacks on video recognition models," in *AAAI*, vol. 34, no. 07, 2020, pp. 12 338–12 345.

[6] Y. Liu, X. Chen, C. Liu, and D. Song, "Delving into transferable adversarial examples and black-box attacks," *arXiv preprint arXiv:1611.02770*, 2016.

[7] C. Xie, Z. Zhang, Y. Zhou, S. Bai, J. Wang, Z. Ren, and A. L. Yuille, "Improving transferability of adversarial examples with input diversity," in *CVPR*, 2019, pp. 2730–2739.

[8] J. Lin, C. Song, K. He, L. Wang, and J. E. Hopcroft, "Nesterov accelerated gradient and scale invariance for adversarial attacks," *arXiv preprint arXiv:1908.06281*, 2019.

[9] Y. Dong, T. Pang, H. Su, and J. Zhu, "Evading defenses to transferable adversarial examples by translation-invariant attacks," in *CVPR*, 2019, pp. 4312–4321.

[10] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li, "Boosting adversarial attacks with momentum," in *CVPR*, 2018, pp. 9185–9193.

[11] D. Wu, Y. Wang, S.-T. Xia, J. Bailey, and X. Ma, "Skip connections matter: On the transferability of adversarial examples generated with resnets," *arXiv preprint arXiv:2002.05990*, 2020.

[12] W. Wu, Y. Su, X. Chen, S. Zhao, I. King, M. R. Lyu, and Y.-W. Tai, "Boosting the transferability of adversarial samples via attention," in *CVPR*, 2020, pp. 1161–1170.

[13] Z. Wang, H. Guo, Z. Zhang, W. Liu, Z. Qin, and K. Ren, "Feature importance-aware transferable adversarial attacks," in *ICCV*, 2021, pp. 7639–7648.

[14] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *CVPR*, 2017, pp. 6299–6308.

[15] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *CVPR*, 2018, pp. 7794–7803.

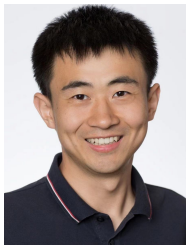
[16] W. Zhou, X. Hou, Y. Chen, M. Tang, X. Huang, X. Gan, and Y. Yang, "Transferable adversarial perturbations," in *ECCV*, 2018, pp. 452–467.

[17] Z. Wei, J. Chen, Z. Wu, and Y.-G. Jiang, "Cross-modal transferable adversarial attacks from images to videos," in *CVPR*, 2022, pp. 15 064–15 073.

[18] A. Kurakin, I. Goodfellow, S. Bengio *et al.*, "Adversarial examples in the physical world," 2016.
[19] Y. Lu, Y. Jia, J. Wang, B. Li, W. Chai, L. Carin, and S. Velipasalar, "Enhancing cross-task black-box transferability of adversarial examples with dispersion reduction," in *CVPR*, 2020, pp. 940–949.
[20] Q. Huang, I. Katsman, H. He, Z. Gu, S. Belongie, and S.-N. Lim, "Enhancing adversarial example transferability with an intermediate level attack," in *ICCV*, 2019, pp. 4733–4742.
[21] Z. Wei, J. Chen, Z. Wu, and Y.-G. Jiang, "Boosting the transferability of video adversarial examples via temporal translation," 2021.
[22] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," in *CVPR*, 2015, pp. 4694–4702.
[23] Y.-G. Jiang, Z. Wu, J. Wang, X. Xue, and S.-F. Chang, "Exploiting feature and class relationships in video categorization with regularized deep neural networks," *IEEE TPAMI*, vol. 40, no. 2, pp. 352–364, 2017.
[24] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *ICCV*, 2019, pp. 6202–6211.
[25] C. Yang, Y. Xu, J. Shi, B. Dai, and B. Zhou, "Temporal pyramid network for action recognition," in *CVPR*, 2020, pp. 591–600.
[26] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *ArXiv*, vol. abs/2010.11929, 2021.
[27] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. J'egou, "Training data-efficient image transformers & distillation through attention," in *ICML*, 2021.
[28] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," *ICCV*, pp. 9992–10 002, 2021.
[29] D. Neimark, O. Bar, M. Zohar, and D. Asselmann, "Video transformer network," *ICCV Workshops*, pp. 3156–3165, 2021.
[30] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lucic, and C. Schmid, "Vivit: A video vision transformer," *ICCV*, pp. 6816–6826, 2021.
[31] H. Fan, B. Xiong, K. Mangalam, Y. Li, Z. Yan, J. Malik, and C. Feichtenhofer, "Multiscale vision transformers," *ICCV*, pp. 6804–6815, 2021.
[32] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu, "Video swin transformer," *ArXiv*, vol. abs/2106.13230, 2021.
[33] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?" *ArXiv*, vol. abs/2102.05095, 2021.
[34] M. Patrick, D. Campbell, Y. M. Asano, I. M. F. Metze, C. Feichtenhofer, A. Vedaldi, and J. F. Henriques, "Keeping your eye on the ball: Trajectory attention in video transformers," *ArXiv*, vol. abs/2106.05392, 2021.
[35] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," *ArXiv*, vol. abs/2004.05150, 2020.
[36] L. Jiang, X. Ma, S. Chen, J. Bailey, and Y.-G. Jiang, "Black-box adversarial attacks on video recognition models," in *ACM Multimedia*, 2019, pp. 864–872.
[37] S. Kornblith, M. Norouzi, H. Lee, and G. Hinton, "Similarity of neural network representations revisited," in *ICML*. PMLR, 2019, pp. 3519–3529.
[38] T. Nguyen, M. Raghu, and S. Kornblith, "Do wide and deep networks learn the same things? uncovering how neural network representations vary with width and depth," *arXiv preprint arXiv:2010.15327*, 2020.
[39] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, and A. Dosovitskiy, "Do vision transformers see like convolutional neural networks?" *Advances in NeurIPS*, vol. 34, pp. 12 116–12 128, 2021.
[40] X. Wei, J. Zhu, S. Yuan, and H. Su, "Sparse adversarial perturbations for videos," in *AAAI*, vol. 33, no. 01, 2019, pp. 8973–8980.
[41] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
[42] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *2017 IEEE Conference on CVPR (CVPR)*, 2017, pp. 936–944.
[43] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.
[44] A. Krizhevsky, "One weird trick for parallelizing convolutional neural networks," *arXiv preprint arXiv:1404.5997*, 2014.
[45] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.
[46] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size," *arXiv preprint arXiv:1602.07360*, 2016.
[47] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
[48] T. W. Anderson, "An introduction to multivariate statistical analysis," Wiley New York, Tech. Rep., 1962.

Zhipeng Wei received the B.E. and M.Sc. degrees from Jilin University, Changchun, China in 2017 and 2020, respectively. He is currently working toward the Ph.D. degree in School of Computer Science at Fudan University, Shanghai, China. His research interests including computer vision and AI security.

Jingjing Chen is now an Associate Professor at the School of Computer Science, Fudan University. Before joining Fudan University, she was a postdoc research fellow at the School of Computing in the National University of Singapore. She received her Ph.D. degree in Computer Science from the City University of Hong Kong in 2018. Her research interest lies in the areas of robust AI, multimedia content analysis, and deep learning. She Won Best Student Paper Awards in ACM Multimedia 2016 and Multimedia Modeling 2017. In 2020, she was selected to the Shanghai Pujiang Talent Program.



Zuxuan Wu received his Ph.D. in Computer Science from the University of Maryland with Prof. Larry Davis in 2020. He is currently an Associate Professor in the School of Computer Science at Fudan University. His research interests are in computer vision and deep learning. His work has been recognized by an AI 2000 Most Influential Scholars Honoralble Mention in 2021, a Microsoft Research PhD Fellowship (10 people Worldwide) in 2019 and a Snap PhD Fellowship (10 people Worldwide) in 2017.

Yu-Gang Jiang received the PhD degree in Computer Science from City University of Hong Kong in 2009 and worked as a Postdoctoral Research Scientist at Columbia University, New York during 2009-2011. He is currently a Professor and Dean at School of Computer Science, Fudan University, Shanghai, China. His research lies in the areas of multimedia, computer vision and AI security. His work has led to many awards, including the inaugural ACM China Rising Star Award, the 2015 ACM SIGMM Rising Star Award, and the research award for excellent young scholars from NSF China.