– Supplemental Material –
Cross-Modal Transferable Adversarial Attacks from Images to Videos

Zhipeng Wei[1,2], Jingjing Chen[1,2†], Zuxuan Wu[1,2], Yu-Gang Jiang[1,2]
[1]Shanghai Key Lab of Intelligent Information Processing,
School of Computer Science, Fudan University
[2]Shanghai Collaborative Innovation Center on Intelligent Visual Computing
zpwei21@m.fudan.edu.cn, {chenjingjing, zxwu, ygj}@fudan.edu.cn

Attack	White-box Model		
	NL-101	SlowFast-101	TPN-101
FGSM	29.70	40.35	52.97
BIM	7.67	9.16	17.07
MI	22.03	30.45	48.27
DI	6.19	11.63	14.36
TI	9.41	9.41	13.36
SIM	7.67	11.88	16.33
SGM	6.68	9.40	15.59
TAP	33.42	42.57	57.67
ATA	5.20	8.91	15.84
TT	13.61	18.81	16.33

Table 1. AASR (%) against video recognition models on UCF-101. The three columns on the right use the NL-101, SlowFast-101 and TPN-101 models as white-box models separately. AASR is calculated by averaging ASR over black-box video models that have a different architecture from the white-box model.

Attack	White-box Model		
	NL-101	SlowFast-101	TPN-101
FGSM	41.31	46.93	58.87
BIM	10.93	11.62	18.37
MI	33.43	36.25	52.12
DI	10.68	11.50	21.93
TI	13.50	11.81	20.12
SIM	12.31	11.31	22.18
SGM	11.18	13.87	22.25
TAP	43.81	50.87	62.62
ATA	7.43	7.00	11.12
TT	44.62	40.56	20.31

Table 2. AASR (%) against video recognition models on Kinetics-400. The three columns on the right use the NL-101, SlowFast-101 and TPN-101 models as white-box models separately. AASR is calculated by averaging ASR over black-box video models that have a different architecture from the white-box model.

1. More Cosine Similarity Analysis

We provide more analysis about the cosine similarity of intermediate features between image and video models. Figure 1 shows the results. It can be seen that the intermediate features between images models and video models are similar to a certain extent regardless of the selected intermediate layer of video models. Therefore, the results also support our assumption that the intermediate feature space between images and video frames is somehow similar.

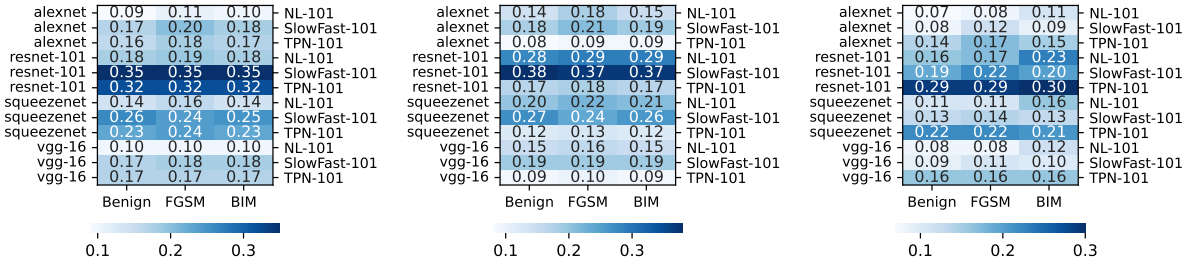
2. Results of Transfer-based Attacks

We provide the results of transfer-based attacks without the fine-tuning method. The results of attacking UCF-101 and Kinetics-400 datasets are shown in Table 1 and Table 2, respectively. From the results, we observe that compared to the results of ILAF, transfer-based attacks achieve much lower AASR. It indicates that ILAF can further improve transferability based on the generated adversarial examples from transfer-based attacks. Besides, the one-step attack, FGSM, achieves better results than iterative attacks for most cases, which indicates that transfer-based attack methods on the image domain are not applicable to the video domain. In general, the performance comparison between the proposed ENS-I2V and ILAF in Section 4.4 "Comparing against stronger baselines" can prove the effectiveness of ENS-I2V.

3. More Discussion about Stronger Baselines

From the results of Figure 6 in Section 4.4 "Comparing against stronger baselines", we observe that the proposed ENS-I2V attack performs worse than ILAF when they use TPN-101 as the white-box model on Kinetics-400. Com-

†Correspondence to: Jingjing Chen.



(a) The second block (b) The third block (c) The fourth block

Figure 1. Cosine similarity analysis of feature maps between image models and video models on benign examples and adversarial examples. The three columns use the second, third, fourth 3D-Resnet block to conduct analysis, separately.

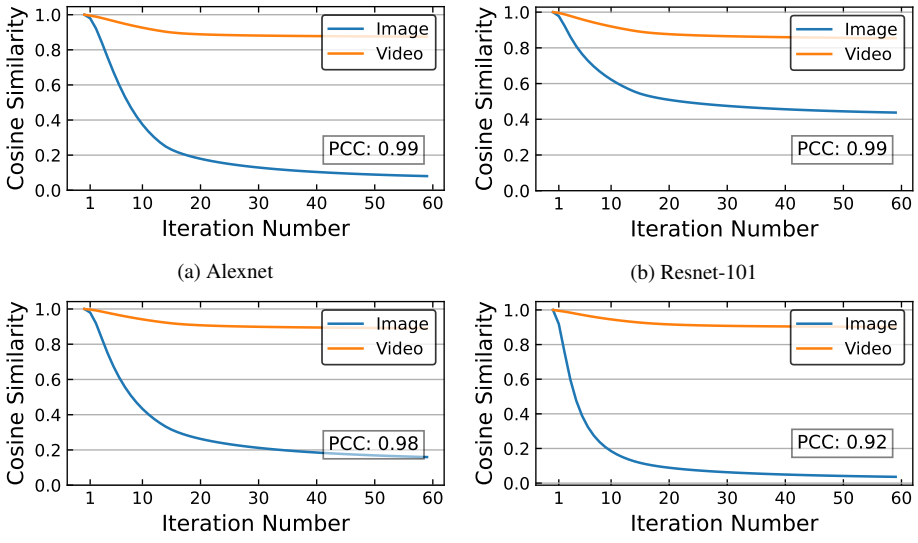
pared to ENS-I2V, ILAF can mine spatial and temporal adversarial information by accessing the white-box model TPN-101. Thus, the lack of temporal information may be the cause of the performance degradation in the ENS-I2V attack. However, the ENS-I2V attack achieves better AASR than ILAF for other cases. We attribute the used dataset and white-box model for the high performance of ILAF when using TPN-101 as the white-box model on Kinetics-400. First, Kinetics-400 contains richer motion information than UCF-101, thus a video model trained on the Kinetics-400 can capture more motion information than that trained on the UCF-101. It indicates that disrupting temporal information allows better performance on video models trained on Kinetics-400. That may be why ILAF achieves higher AASR on Kinetics-400 than on UCF-101. Second, TPN-101 captures the visual tempos through fusing multi-layer features and achieves a higher top-1 validation accuracy than NL and SlowFast models. This suggests that generating adversarial examples on TPN-101 can better disrupt temporal information. That may be why ILAF performs better using TPN-101 as the white-box model than using other models as the white-box model. As a result, attacking NL and SlowFast models in the black-box setting requires temporal information to further improve AASR. And TPN-101 provides better temporal information for ILAF. Therefore, in the future, we will combine temporal information of videos into image models to further boost transferability.

4. More PCC Analysis

Figure 2 and 3 shows the PCC analysis of cosine similarity trends when using SlowFast-101 model and TPN-101 model, respectively. We can observe similar trends as stated in Section 4.5 “Discussion”, which suggests that the stable positive linear relationship between the directional changes of image and video intermediate features. It experimentally supports that minimizing the cosine similarity between features from benign examples and adversarial examples on image models can lead to decrease ones on video models and also demonstrates the effectiveness of the optimized object function.

5. Visualization of Adversarial Examples

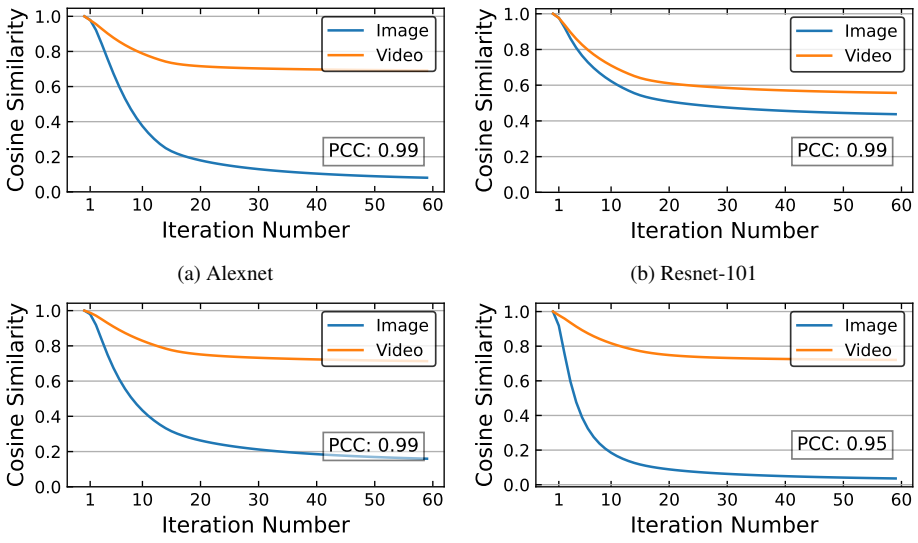
We further visualize 4 randomly selected benign video clips and their corresponding adversarial clips in Figure 4. These adversarial examples are generated on the ensemble of ImageNet-preatrained models (Alexnet, Resnet-101, Squeezenet, Vgg-16) by the proposed ENS-I2V attack. As can be seen, these adversarial examples do not affect human decision-making but fool video models into wrong predictions.



(a) Alexnet (b) Resnet-101

(c) Squeezenet (d) Vgg-16

Figure 2. Pearson correlation coefficient (PCC) analysis between cosine similarity trends computed from image models and the SlowFast-101 video model.



(a) Alexnet (b) Resnet-101

(c) Squeezenet (d) Vgg-16

Figure 3. Pearson correlation coefficient (PCC) analysis between cosine similarity trends computed from image models and the TPN-101 video model.

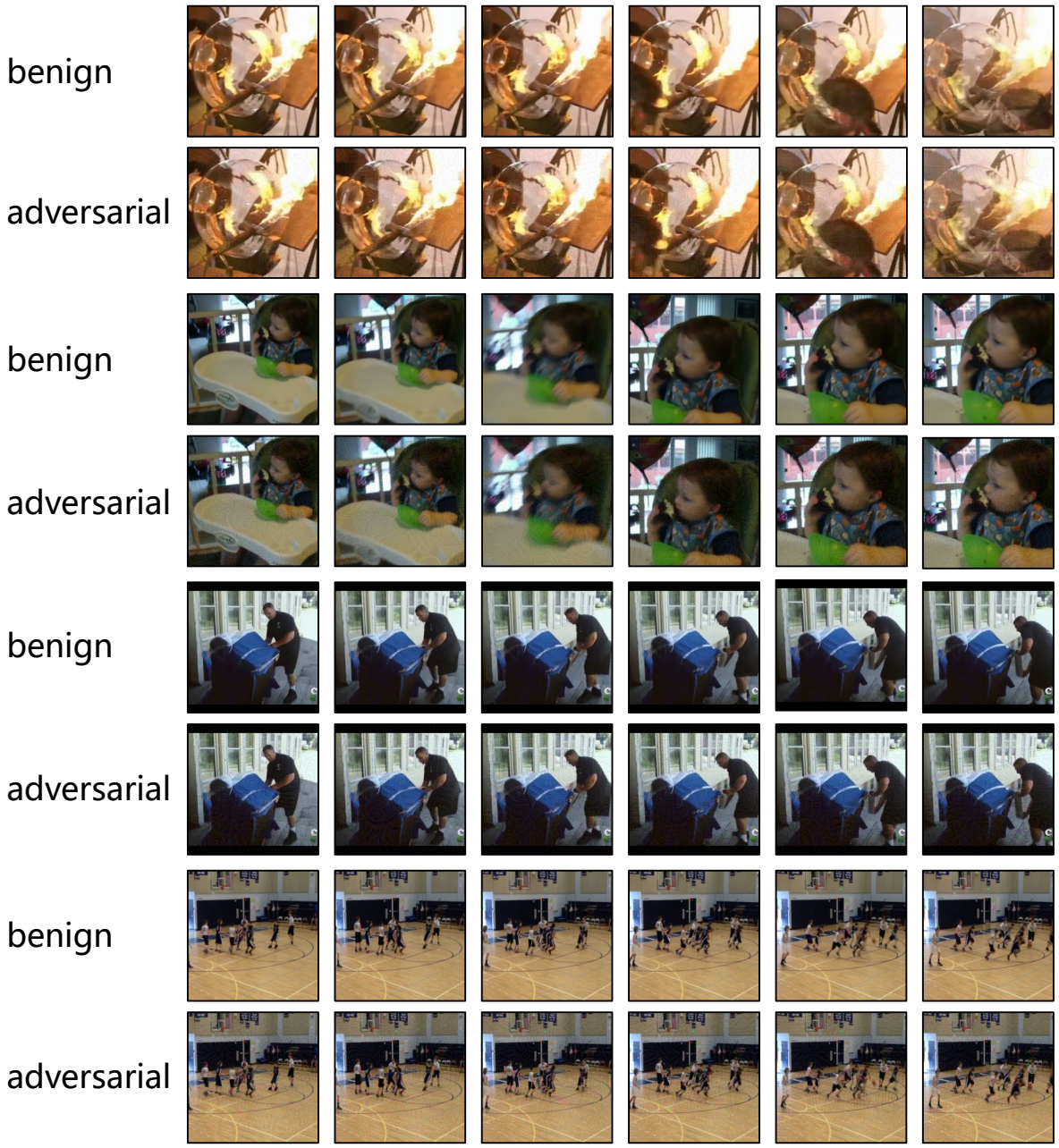


Figure 4. Visualization of randomly picked benign video clips and their corresponding adversarial clips, crafted by the proposed ENS-I2V. Labels from top row to bottom row are ”blowing glass”, ”eating cake”, ”moving furniture”, and ” playing basketball” separately