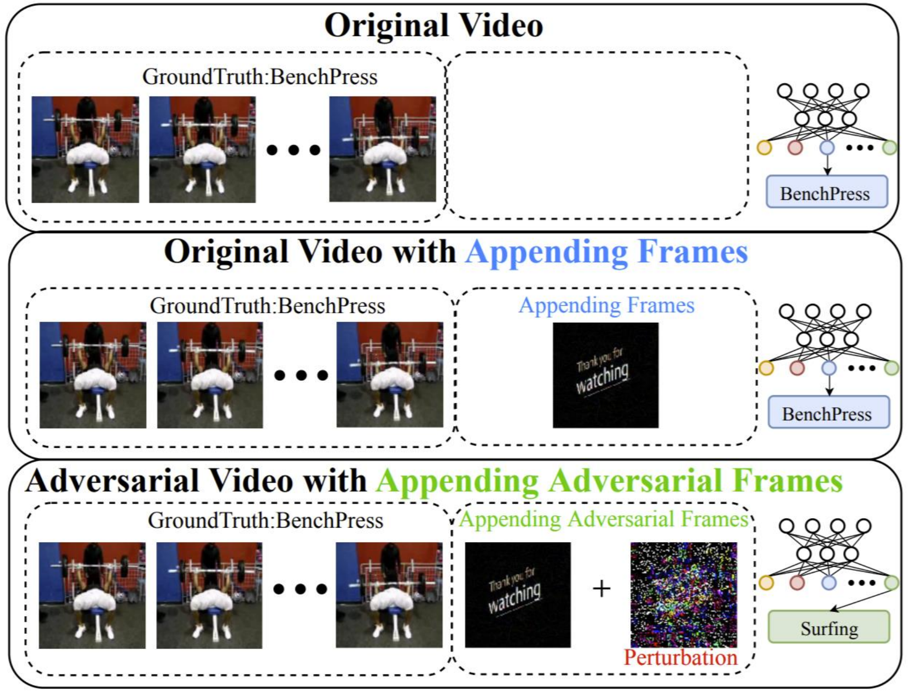
Appending Adversarial Frames for Universal Video Attack

AAF


Original Video
GroundTruth:BenchPress
BenchPress
Original Video with Appending Frames
GroundTruth:BenchPress
Appending Frames
Thank you for watching
BenchPress
Adversarial Video with Appending Adversarial Frames
GroundTruth:BenchPress
Appending Adversarial Frames
Thank you for watching
+
Perturbation
Surfing

Basic Attack Methods

- **Model**:

$$X = \{f_1, f_2 \dots, f_T\}$$
$$E = \{p_1, p_2 \dots p_T\}$$
$$\hat{X} = \{f_1 \oplus p_1, f_2 \oplus p_2 \dots, f_T \oplus p_T\}$$

- **Weakness**：

High authority needed.

Unsafe.(f_1, f_2…are related while p_1, p_2 are not)

High perturbation rate.(every frame)

Weak transferability.

Appending Adversarial Frames Method

- **Model**:

$X \in R^{T \times W \times H \times C}$ (T: number of frames W, H, C: width, height, and channel of each frame)

$\Delta \in R^{\Delta T \times W \times H \times C}$ (adversarial frames without perturbations)

$\hat{\Delta} \in R^{\Delta T \times W \times H \times C}$ (adversarial frames with perturbations)

$\hat{X} \in R^{(T+\Delta T) \times W \times H \times C}$ (adversarial video)

- **Optimization Function**

$$\arg\min_{\mathbf{E}} \lambda ||\mathbf{E}||_p - \ell(\mathbf{1}_y, \mathbb{J}(\hat{\mathbf{X}}; \boldsymbol{\theta}))$$

$$\arg\min_{\mathbf{E}} \lambda ||\mathbf{E}||_p + \ell(\mathbf{1}_{y^*}, \mathbb{J}(\hat{\mathbf{X}}; \boldsymbol{\theta}))$$

Variants of AAFM

- **Across Videos：**

$$\arg\min_{\mathbf{E}} \lambda||\mathbf{E}||_p - \sum_{n=1}^{N} \alpha_n \ell(\mathbf{1}_{y_n}, \mathbb{J}(\hat{\mathbf{X}}_n; \boldsymbol{\theta}))$$

- **Across Models：**

$$\arg\min_{\mathbf{E}} \lambda||\mathbf{E}||_p - \sum_{k=1}^{K} \beta_k \ell(\mathbf{1}_{y_k}, \mathbb{J}(\hat{\mathbf{X}}; \boldsymbol{\theta}_k))$$

- **Feature Similarity：**

$$\arg\min_{\mathbf{E}} \lambda||\mathbf{E}||_p - \ell(\mathbf{1}_y, \mathbb{J}(\hat{\mathbf{X}}; \boldsymbol{\theta})) + \lambda_l||\phi_l(\boldsymbol{\Delta}_{\mathbf{s}}) - \phi_l(\hat{\boldsymbol{\Delta}})||_p$$

Experiments

Table 2. Comparison of BAM and A^{2}FM with different video classification models.

Target Model	Methods	UCF-101		HMDB-51	
		FR (%)	AAP	FR (%)	AAP
I3D-ResNet	BAM	**100**	0.22	**100**	0.31
	A^{2}FM	**100**	**0.05**	**100**	**0.06**
I3D-Inception	BAM	**99.5**	0.20	**100**	0.28
	A^{2}FM	**99.5**	**0.08**	**100**	**0.07**
CNN+LSTM	BAM	**100**	0.20	**100**	0.28
	A^{2}FM	**100**	**0.02**	**100**	**0.02**
C3D	BAM	**99.5**	0.24	**100**	0.30
	A^{2}FM	97.3	**0.14**	96.8	**0.16**
ResNet3D	BAM	**97.8**	0.25	**100**	0.30
	A^{2}FM	95.1	**0.09**	**100**	**0.07**
P3D	BAM	**100**	0.20	**100**	0.28
	A^{2}FM	**100**	**0.02**	**100**	**0.02**

fewer perturbations

Table 3. Comparison of BAM and A^{2}FM-AV in transferability across different videos.

Target Model	Methods	UCF-101		HMDB-51	
		FR (%)	AAP	FR (%)	AAP
I3D-ResNet	BAM	95.4	0.62	93.0	0.70
	A^{2}FM-AV	**98.1**	**0.52**	**97.8**	**0.60**
I3D-Inception	BAM	2.6	**0.34**	2.0	**0.25**
	A^{2}FM-AV	**69.3**	1.25	**2.3**	0.84
CNN+LSTM	BAM	18.1	**0.09**	**69.6**	**0.13**
	A^{2}FM-AV	**47.1**	0.16	45.7	0.21
C3D	BAM	97.9	**0.75**	**98.0**	**0.68**
	A^{2}FM-AV	**98.1**	1.21	96.9	1.75
ResNet3D	BAM	45.2	**0.65**	58.6	**0.49**
	A^{2}FM-AV	**96.6**	1.21	**94.1**	0.79
P3D	BAM	20.7	**0.11**	46.9	0.16
	A^{2}FM-AV	**98.4**	0.25	**97.4**	**0.15**

Better transferability
AAP is not guaranteed

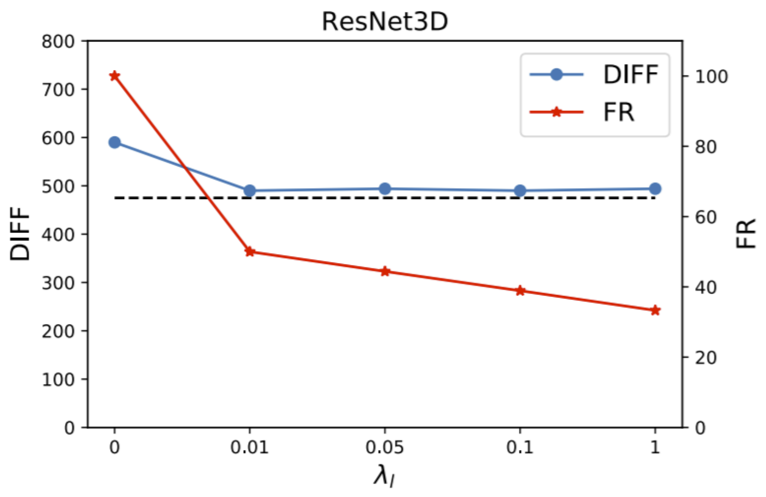
Experiments

Table 4. Comparison of BAM and A^2FM-AM in transferability across models on UCF-101 dataset. The first column indicates we use the Leave-One-Out ensemble method that excludes one model to produce perturbations. For instance,'−I3D-ResNet' means the corresponding ensemble model excludes I3D-ResNet. The numbers in the 3-8 columns are the fooling rates (%) for each attacked model.

Models	Method	I3D-ResNet	ResNet3D	P3D	I3D-Inception	C3D	CNN+LSTM
−I3D-ResNet	BAM	0	78.7	84.6	87.8	70.8	56.2
	A^2FM-AM	**39.5**	68.1	97.4	42.9	85.4	81.6
−ResNet3D	BAM	100	0	84.6	87.8	70.8	38.9
	A^2FM-AM	89.5	**6.4**	97.4	52.2	85.4	71.4
−P3D	BAM	100	80.9	15.4	87.8	72.9	58.8
	A^2FM-AM	86.8	74.5	**59.0**	50.0	85.4	83.7
−I3D-Inception	BAM	100	83.0	97.4	0	73.0	61.1
	A^2FM-AM	86.8	78.7	100	**2.0**	85.4	50.0
−C3D	BAM	100	83.0	100	90.0	0	64.7
	A^2FM-AM	92.1	80.9	100	60.0	**20.8**	79.6
−CNN+LSTM	BAM	100	80.9	97.4	97.8	72.9	35.7
	A^2FM-AM	89.5	74.5	100	55.6	85.4	**77.6**

Better transferability

Experiments



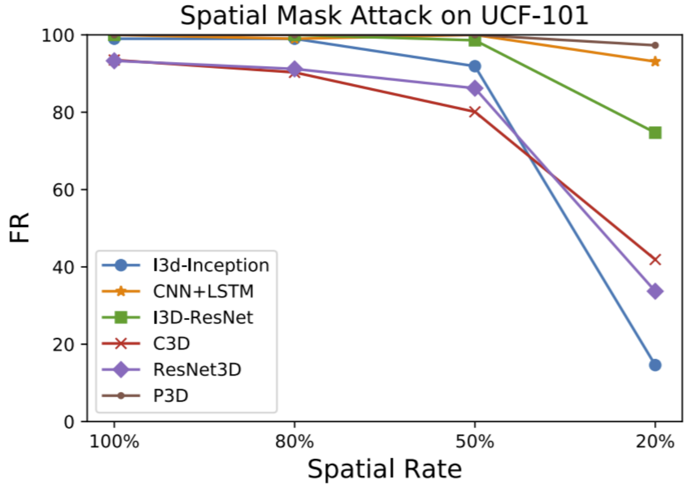


Table 5. Comparison of BAM and A^{2}FM for targeted attack.

Target Model	Methods	UCF-101		HMDB-51	
		FR (%)	AAP	FR (%)	AAP
I3D-ResNet	BAM	97.6	0.29	**97.8**	0.31
	A^{2}FM	**97.7**	**0.17**	**97.8**	**0.14**
I3D-Inception	BAM	**84.6**	0.23	**96.8**	0.27
	A^{2}FM	27.4	**0.08**	40.2	**0.08**
CNN+LSTM	BAM	**61.6**	0.23	**55.8**	0.27
	A^{2}FM	53.2	**0.07**	42.4	**0.07**
C3D	BAM	**97.9**	0.30	**97.8**	0.31
	A^{2}FM	83.8	**0.26**	95.0	**0.22**
Resnet3D	BAM	**98.1**	0.28	**98.0**	0.30
	A^{2}FM	**98.1**	**0.15**	**98.0**	**0.13**
P3D	BAM	**98.0**	0.22	**97.8**	0.26
	A^{2}FM	97.8	**0.07**	**97.8**	**0.08**

Bad performance on targeted attack

Conclusion

The paper present an adversarial video attack method, which appends a few dummy frames with adversarial perturbations to the original video.

Comparing with the basic adversarial video attack method, AAF has a better performance on perturbation rate and transferability between diff-erent videos and classifier modules, while a worse performance on targeted attack.