Adversarial Attacks on Black Box Video Classifiers: Leveraging the Power of Geometric Transformations







Shasha Li[1] , Abhishek Aich[1], Shitong Zhu, M. Salman Asif, Chengyu Song,
Amit K. Roy-Chowdhury, Srikanth V. Krishnamurthy

[1]joint first authors

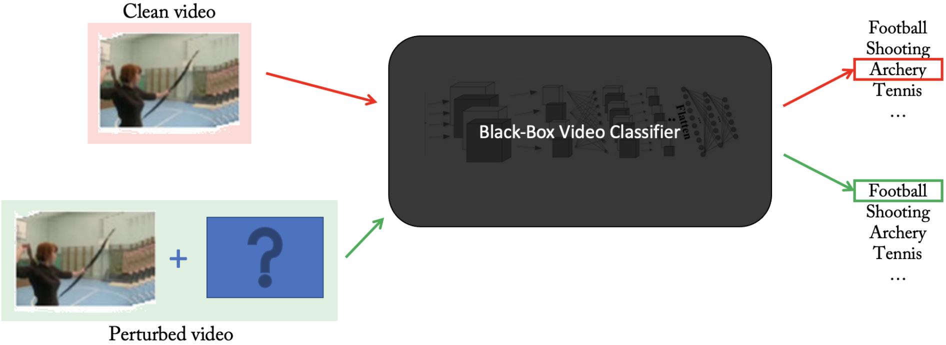
Introduction

Black-box Adversarial Attack on Video Classifiers

- **Problem Statement:** How to create imperceptible video perturbation, so that the perturbed video is misclassified by the black-box model?

Black-box Adversarial Attack on Video Classifiers

- **Problem Statement:** How to create imperceptible video perturbation, so that the perturbed video is misclassified by the black-box model?



Black-box Adversarial Attack on Video Classifiers

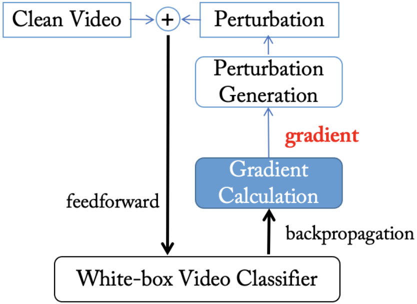
- **Effective attacks:** Better gradient estimation is the key to query-based black-box attack.

(a) Illustration of White-box attack

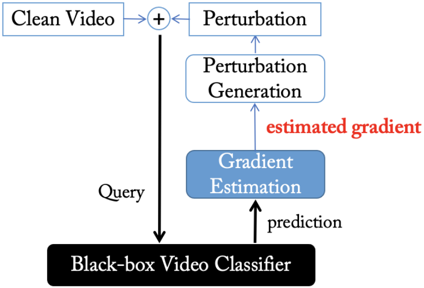
(b) Illustration of Black-box attack

Black-box Adversarial Attack on Video Classifiers

- **Effective attacks:** Better gradient estimation is the key to query-based black-box attack.



(a) Illustration of White-box attack



(b) Illustration of Black-box attack

Gradient Estimation

Gradient Estimation: Sampling Directions

- A simplified algorithm.
- How to sample π is important!

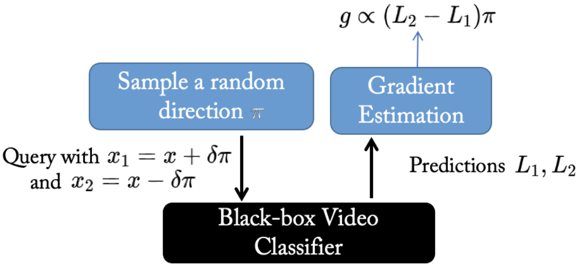


Figure 2: Gradient estimation for high dimensional function

Gradient Estimation: Sampling Directions

- A simplified algorithm.
- How to sample π is important!

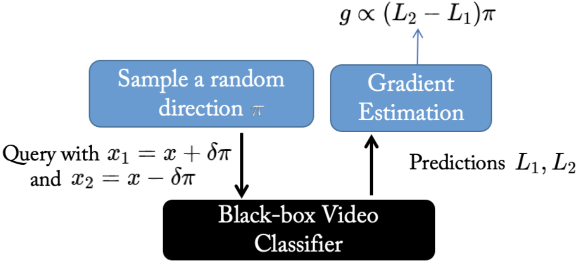


Figure 2: Gradient estimation for high dimensional function

Gradient Estimation: Query-efficiency

- π is in high dimensional space $D = T \times H \times W \times C$, where T is the number of frames, H and W are the height and width of the frames, C is the number of channels.
- Higher dimensionality leads to more number of queries $\rightarrow$ becomes worse compared to query-based image attacks.
- **Goal:** Query-efficient query-based video attack!

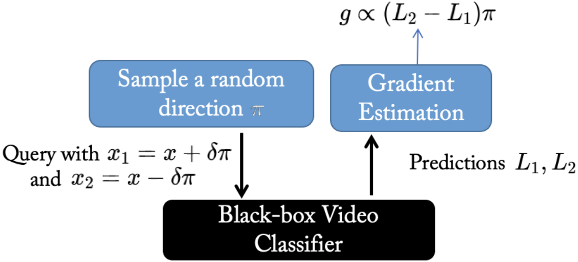


Figure 2: Gradient estimation for high dimensional function

Gradient Estimation: Query-efficiency

- π is in high dimensional space $D = T \times H \times W \times C$, where T is the number of frames, H and W are the height and width of the frames, C is the number of channels.
- Higher dimensionality leads to more number of queries $\rightarrow$ becomes worse compared to query-based image attacks.
- **Goal:** Query-efficient query-based video attack!

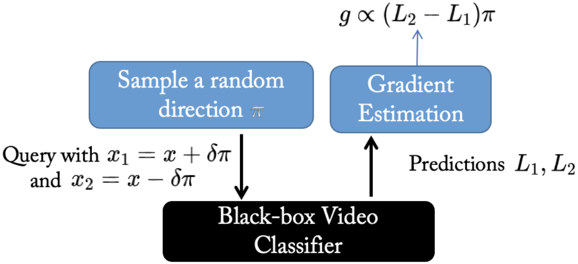


Figure 2: Gradient estimation for high dimensional function

Gradient Estimation: Query-efficiency

- π is in high dimensional space $D = T \times H \times W \times C$, where T is the number of frames, H and W are the height and width of the frames, C is the number of channels.
- Higher dimensionality leads to more number of queries $\rightarrow$ becomes worse compared to query-based image attacks.
- **Goal:** Query-efficient query-based video attack!

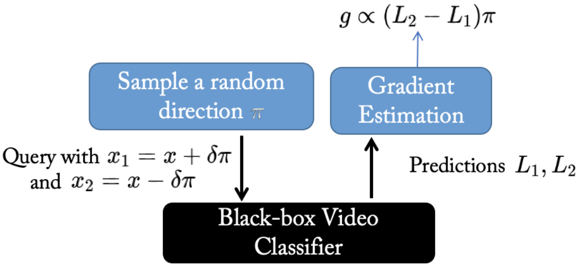


Figure 2: Gradient estimation for high dimensional function

Motivation of Proposed Work

Motivation: Reduce the Search Space

- To estimate better gradient g.
- Sample π in a subspace (dimensionality reduction), which contains more effective π.
- Consider the intrinsic different between images and videos, i.e., the temporal dimension and aim to disrupt the motion context of videos.

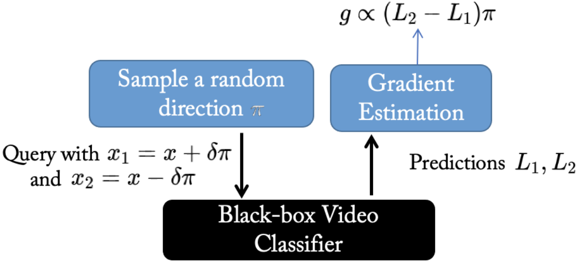


Figure 2: Gradient estimation for high dimensional function

Motivation: Reduce the Search Space

- To estimate better gradient g.
- Sample π in a subspace (dimensionality reduction), which contains more effective π.
- Consider the intrinsic different between images and videos, i.e., the temporal dimension and aim to disrupt the motion context of videos.

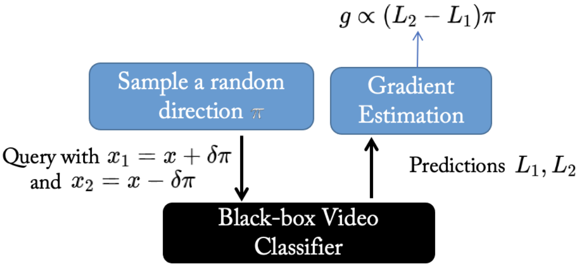


Figure 2: Gradient estimation for high dimensional function

Motivation: Reduce the Search Space

- To estimate better gradient $\boldsymbol{g}$.
- Sample $\boldsymbol{\pi}$ in a subspace (dimensionality reduction), which contains more effective $\boldsymbol{\pi}$.
- Consider the intrinsic different between images and videos, i.e., the temporal dimension and aim to disrupt the motion context of videos.

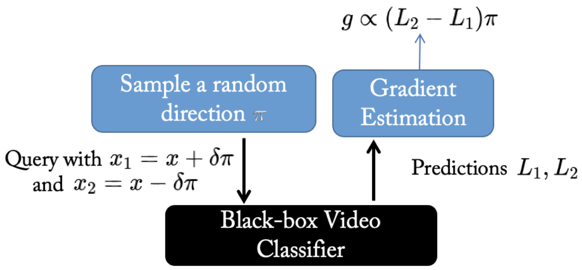


Figure 2: Gradient estimation for high dimensional function

Proposed Method:

GEOmetrically TRAnsformed Perturbations (GEO-TRAP)

Proposed Method: GEO-TRAP

- Randomly sample $\boldsymbol{r}_{\text{frame}} \in \mathbb{R}^{H\times W\times C}$, then warp $\boldsymbol{r}_{\text{frame}}$ with T random geometric transformations to get $\boldsymbol{\pi} \in \mathbb{R}^{T\times H\times W\times C}$

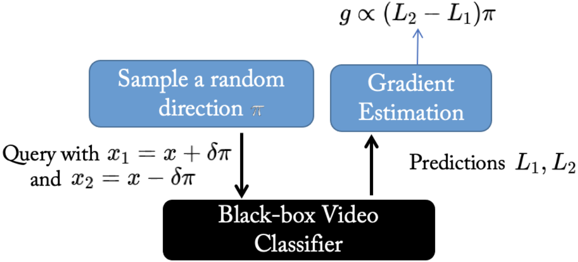
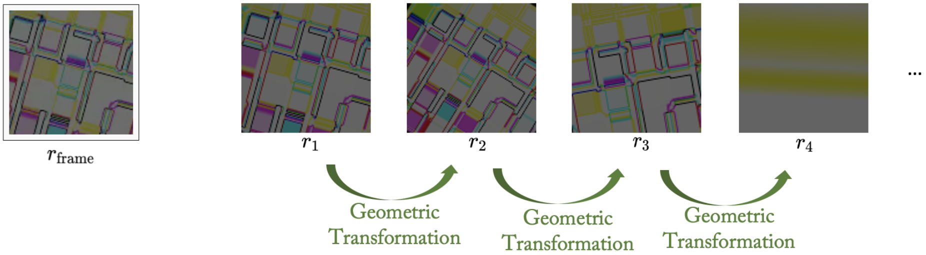


Figure 2: Gradient estimation for high dimensional function

Proposed Method: GEO-TRAP

- **Dummy Illustration:** Warping random noise r_{frame} to create search directions for gradients



Proposed Method: GEO-TRAP

- Randomly sample $\boldsymbol{r}_{\text{frame}} \in \mathbb{R}^{H \times W \times C}$, then warp $\boldsymbol{r}_{\text{frame}}$ with T random geometric transformations to get $\boldsymbol{\pi} \in \mathbb{R}^{T \times H \times W \times C}$.

Why does it work?

- Temporally structured perturbations.
  - Geometric progression in the temporal dimension.
- Assume the degrees of freedom of the geometric transformation is F, the dimensionality D is then reduced from $(T \times H \times W \times C)$ to $(H \times W \times C) + (T \times F)$ where, $F << T \times H \times W \times C$.
  - e.g. $F = 6$ for affine transformation.

Why GEO-TRAP works?

- Cosine similarity between the estimated $\boldsymbol{g}$ and the ground truth $\boldsymbol{g}^\star$, averaged over 1000 randomly chosen samples.
- **Takeaway:** GEO-TRAP estimates better gradients compared to baselines.

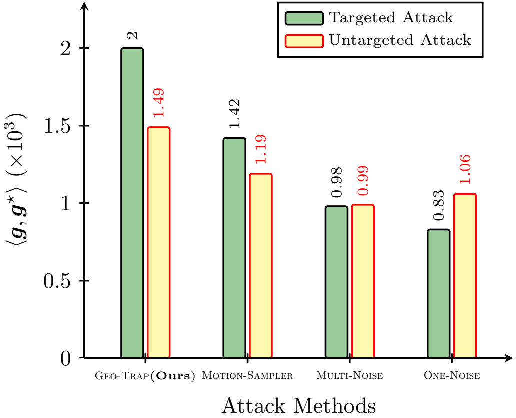
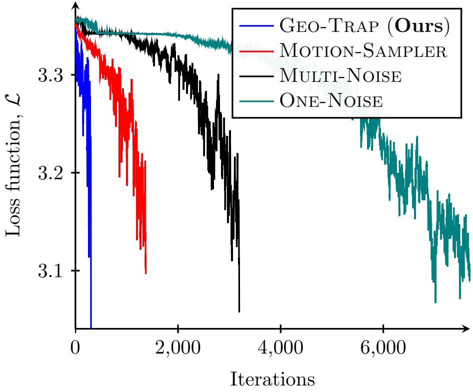


Figure 2: Measure the quality of estimated $\boldsymbol{g}$

Why GEO-TRAP works?

- **Better gradients** leads to quicker convergence, thus **fewer number of queries** required.



Experimental Result

Evaluation Metric:

- **Success Rates (SR):** total success rate of attack within query and perturbation budgets.
- **Average Number of Queries (ANQ):** the average total queries from attacks for all videos (including failed ones).

Datasets:

- **UCF-101**[1]: UCF-101 includes 13320 videos from 101 human action categories (e.g., applying lipstick, biking).
- **20BN-JESTER (Jester)**[2]: Jester includes 27 kinds of gesture videos recorded by crowd-sourced workers (e.g., sliding hand left, sliding two fingers right).

[1] Khurram Soomro et al. "UCF101: A Dataset of 101 Human Actions Classes from Videos in the Wild". *arXiv:1212.0402* (2012).
[2] Joanna Materzynska et al. "The Jester Dataset: A Large-scale Video Dataset of Human Gestures". *ICCV Workshops*. 2019.

Experimental Result

Evaluation Metric:

- **Success Rates (SR):** total success rate of attack within query and perturbation budgets.
- **Average Number of Queries (ANQ):** the average total queries from attacks for all videos (including failed ones).

Datasets:

- **UCF-101**[1]: UCF-101 includes 13320 videos from 101 human action categories (e.g., applying lipstick, biking).
- **20BN-JESTER (Jester)**[2]: Jester includes 27 kinds of gesture videos recorded by crowd-sourced workers (e.g., sliding hand left, sliding two fingers right).

[1] Khurram Soomro et al. "UCF101: A Dataset of 101 Human Actions Classes from Videos in the Wild". *arXiv:1212.0402* (2012).

[2] Joanna Materzynska et al. "The Jester Dataset: A Large-scale Video Dataset of Human Gestures". *ICCV Workshops*. 2019.

Experimental Result

- **Takeaway:** GEO-TRAP achieves the same or higher attack Success Rates (SR) compared to other methods, and requires fewer Average Number of Queries (ANQ).
- More results and analysis in the paper.

Table 1: GEO-TRAP demonstrates highly successful untargeted attacks with fewer queries.

Datasets	Methods	Black-box Video Classifiers							
		C3D		SlowFast		TPN		I3D	
		ANQ (↓)	SR (↑)	ANQ (↓)	SR (↑)	ANQ (↓)	SR (↑)	ANQ (↓)	SR (↑)
Jester	HEURISTICATTACK[3]	4699	99.0%	3572	98.1%	4679	82.0%	4248	98.1%
	MOTION-SAMPLER ATTACK[4]	4549	99.0%	1906	100%	6269	91.3%	3029	99.4%
	GEO-TRAP (**Ours**)	**1602**	**100%**	**521**	**100%**	**3315**	**92.4%**	**1599**	**100%**
UCF-101	HEURISTICATTACK	**5206**	70.2%	3507	87.2%	**6539**	71.8%	6949	84.7%
	MOTION-SAMPLER ATTACK	14336	81.6%	4673	97.2%	20369	75.8%	7400	94.4%
	GEO-TRAP (**Ours**)	11490	**86.2%**	**1547**	**98.8%**	17716	**76.1%**	**4887**	**97.4%**

[3] Zhipeng Wei et al. "Heuristic black-box adversarial attacks on video recognition models". *AAAI*. 2020.

[4] Hu Zhang et al. "Motion-Excited Sampler: Video Adversarial Attack with Sparked Prior". *ECCV*. 2020.

Conclusion

- We propose a new black-box video attack method, which parameterizes the video search space into an image search space and a geometric transformation parameter search space.
- With the reduced and temporally structured search space, we are able to achieve higher attack success rate with fewer queries.

Thank You!

- **Acknowledgement** → The authors would like to thank Dr. Cliff Wang of US Army Research Office for his extensive comments and input on this work. This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Agreement No. HR00112090096.
- **Paper ID: 27235** → Adversarial Attacks on Black Box Video Classifiers: Leveraging the Power of Geometric Transformations
- **Paper link** → `https://arxiv.org/pdf/2110.01823.pdf`