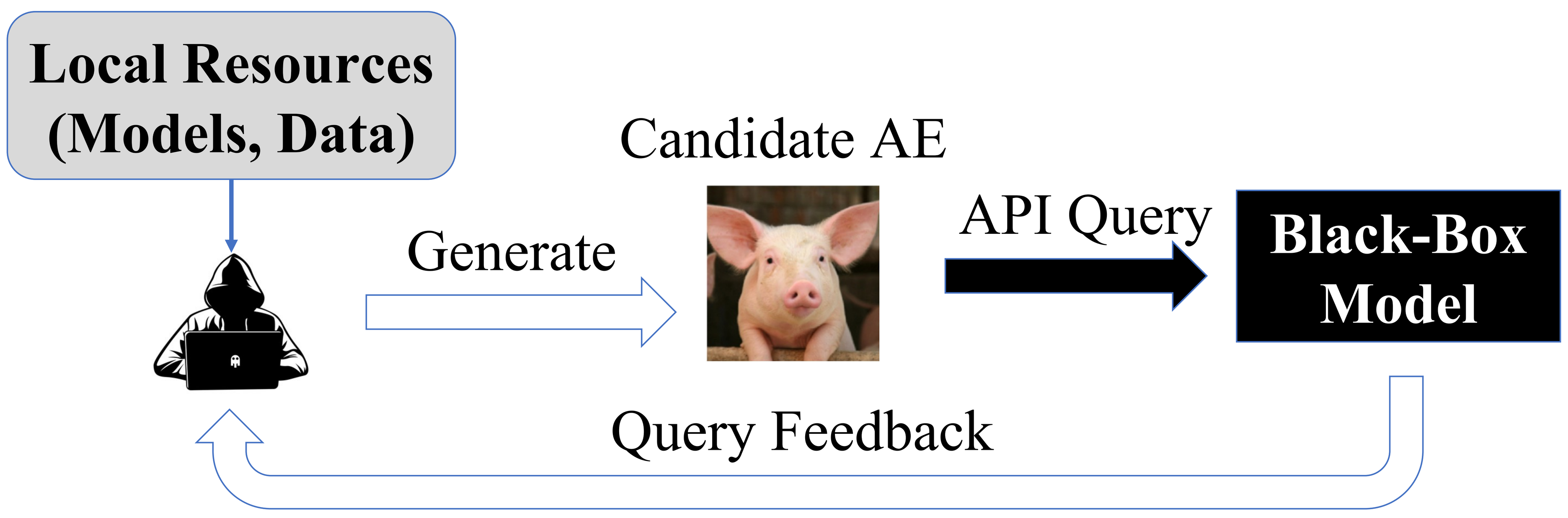
SoK: Pitfalls in Evaluating Black-Box Attacks

Fnu Suya, Anshuman Suri*, Tingwei Zhang, Scott Hong, Yuan Tian, David Evans*

Paper: https://arxiv.org/abs/2310.17534 (Link to code inside), accepted to SaTML 2024

NDSS SYMPOSIUM/2024

Black-box AEs



Taxonomy on Threat Model

- **Query Access:** with/without interactive access
- **API Feedback**: details of target model's API returns
- **Quality of Initial Auxiliary Data**: overlap between attacker's auxiliary data and target model's train data
- **Quantity of Initial Auxiliary Data**: if sufficient to train well-performing surrogate models

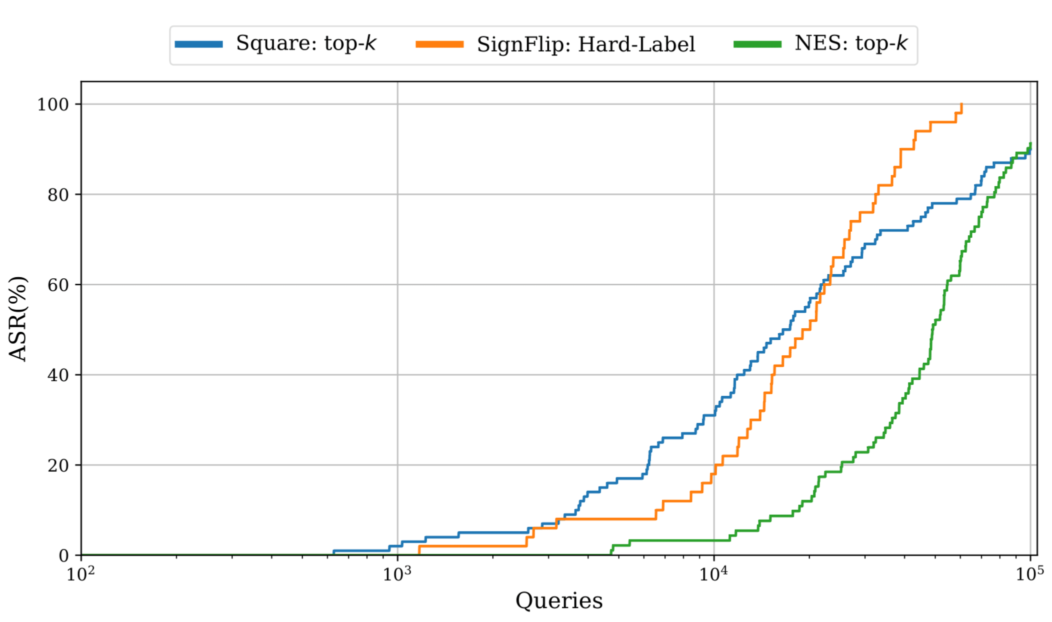
Quality	Quantity	No Interactive Access	With Interactive Access: Hard-Label	With Interactive Access: Top-K	With Interactive Access: Complete Confidence Vector
None	Insufficient	Frequency Manipulation [156] **w/ Pretrained Surrogate*:** Better Loss: [90–92, 155, 157–165] Better Loss for AE Generator: [90, 91, 162]	Random walk: [129–135] Gradient estimation: [98–100, 112–116] Other Gradient-free: [97, 136–139] Classic Black-box Opt.: [108, 166]	NES [3]	Gradient Estimation: [3, 4, 16, 101–111] Classic Black-box Opt.: [117–121] Efficient Random Search: [96, 117–119, 122–128]
	Sufficient	∅	∅	∅	∅
Partial	Insufficient	**w/ Pretrained Surrogate*:** Better Loss: [92, 155, 158, 163]	∅	∅	Boost Existing Methods w/ Trained Generator: [167]
	Sufficient	∅	∅	∅	∅
Complete	Insufficient	Train Shallow Surrogate: [168, 169] **w/ Pretrained Surrogate*:** (Basic) Gradient Sign: [2, 23] Input Augmentation: [32, 34, 37, 42–52, 170] Gradient Stabilization: [24–40] Better Loss: [31, 53–67, 165] Refine Surrogate: [32, 72–80, 84, 88]	Improve UAP w/ Feedback: [164] Train Surrogate w/ Synthetic Data: [171–174] Boost Existing Methods w/ Unlabeled Data [175]	∅	Boost Existing Methods: Trained Generator: [167, 176–179], Unlabeled Data [175] **w/ Pretrained Surrogate***: Save Queries with Surrogate: [140–149, 151] Refine Surrogate with Queries: [143, 150, 152]
	Sufficient	Train Better (Deep) Surrogate: [81–83, 85, 86] Train AE Generator: [89, 91, 93, 180–182] Input Transformation Network: [49, 50, 52] Train Simple Auxiliary Classifier: [58, 59, 91]	Improved Gradient Estimation w/ Trained Generator: [94, 95]	∅	Train AE Generator: [87, 183–185]

The symbol ∅ corresponds to areas in the threat space that, to the best of our knowledge, are not considered by any attacks in the literature.

Insights from Taxonomy

Insight 1: Many underexplored areas need research investigation

Insight 2: Stronger baselines exist under same threat model



Square top-k: our adapted attack. NES: top-k is current state-of-the-art.

Attacks	Square-Attack	ODS-RGF	Hybrid-Square
Attack Success (%)	100	97.7	100
Average Queries	2,317	1,242	117

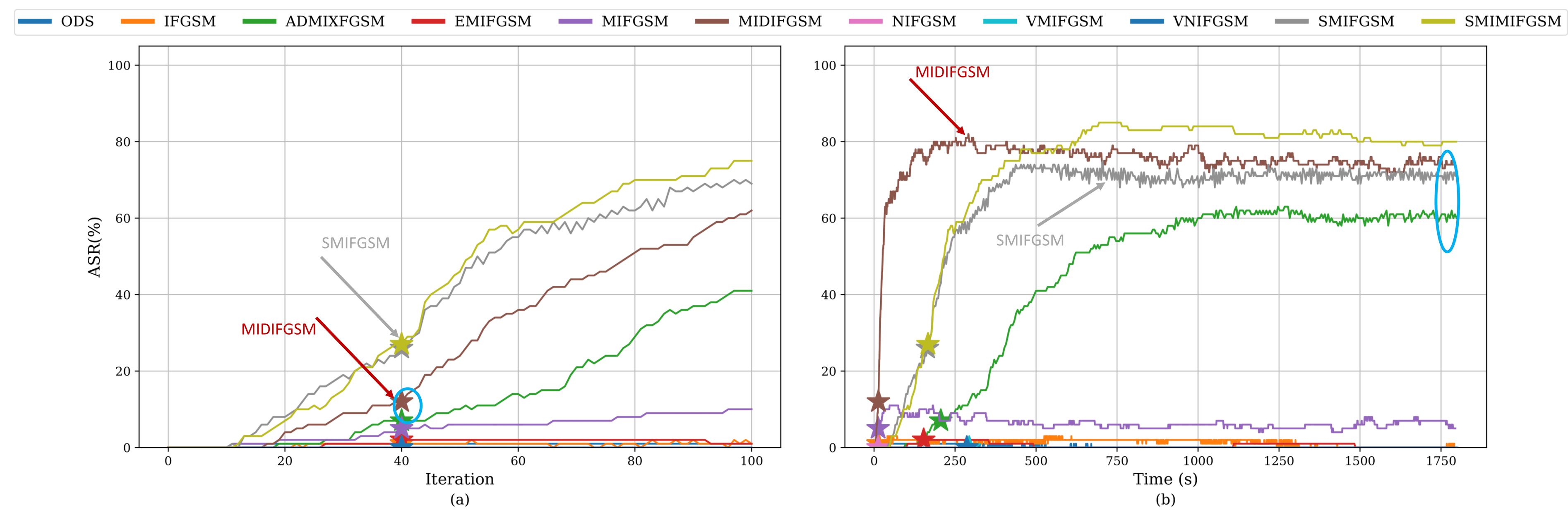
Square Attack is by Andriushchenko et al. (2019). ODS-RGF is by Tashiro et al. (2020). Hybrid Square is ours.

Model extraction attacks: better attacks provide better pretrained surrogate models

Model inversion attacks: better provide better (improved quality) auxiliary data

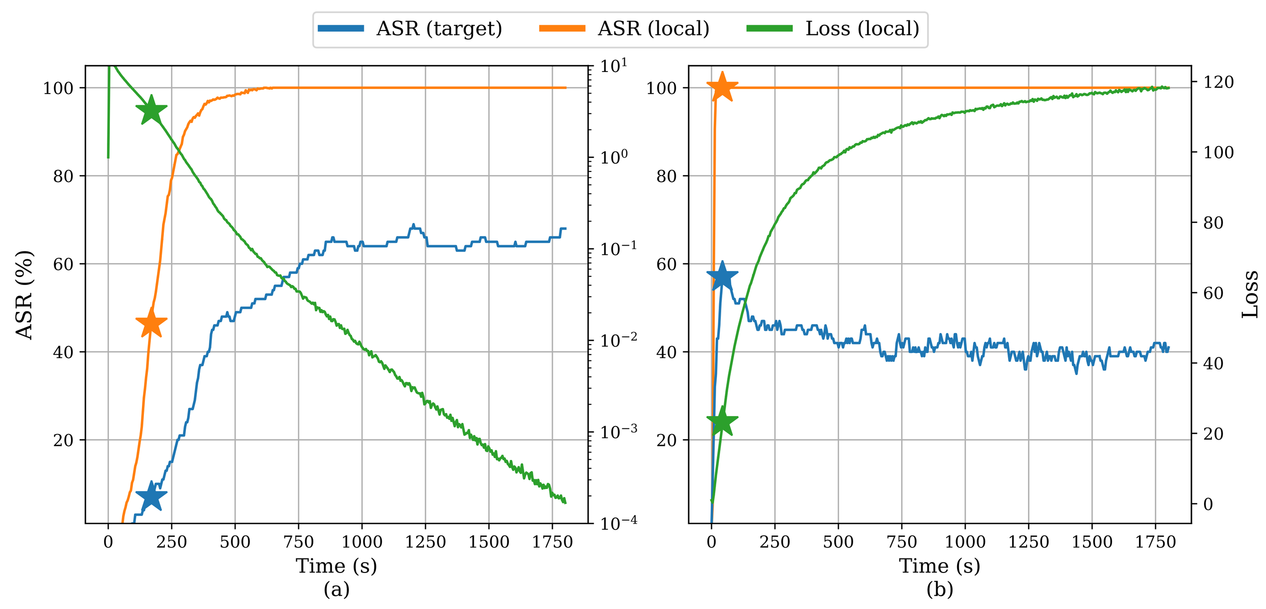
Dynamic combination of extraction and inversion attacks

Rethinking Baseline Comparisons

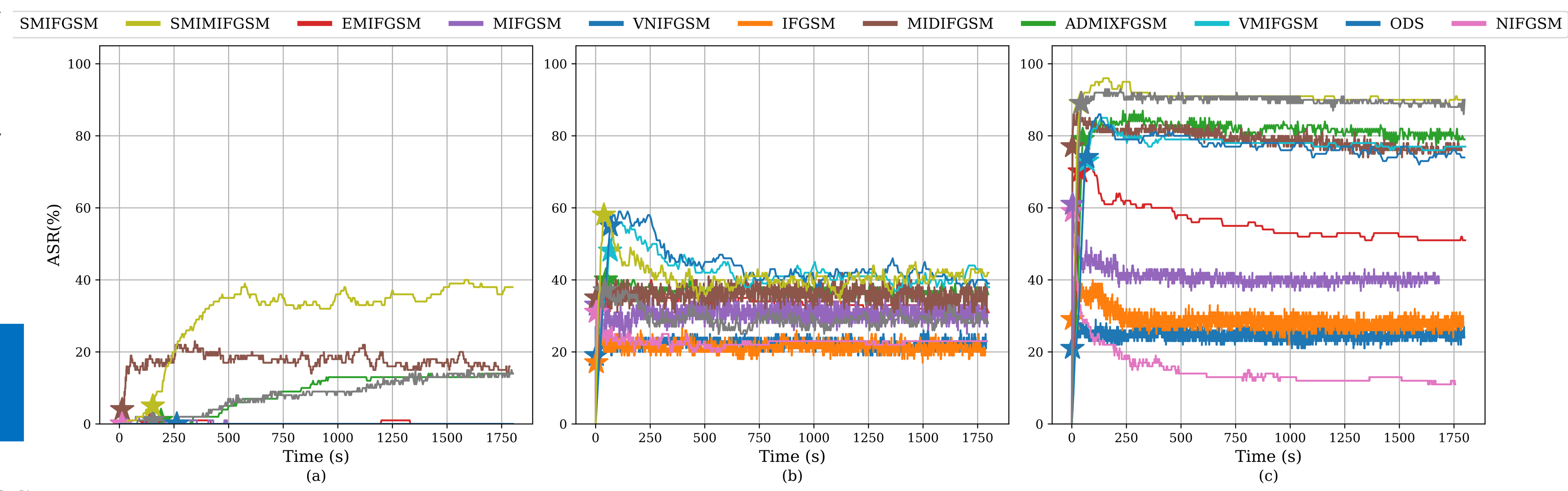


Against DenseNet201 model. (**Left**) current transfer attack evaluation at fixed # of iterations. (**Right**) evaluation of attacks with realistic metric of total local runtime.

Recommendation: run attacks for enough iterations until attack success rate plateau. Execution cost (e.g., local runtime) should be used as equalizing factor when comparing different attacks, not arbitrary number of iterations.



Recommendation: do not rely on local metrics such as attack success or model loss on local models. Develop better metrics that can predict optimal target success rates.



(**Left**) targeted attack with 16/255 perturbation on Inception-v3 (**Middle**) untargeted attack on Inception-v3 with 8/255 perturbation (**Right**) untargeted attack on robust model with 16/255 perturbation.

Recommendation: when evaluating attacks, should include harder settings (e.g., targeted attacks, against robust models). Untargeted attack on standard models are mostly solved.

Conclusion

- Many interesting and practical settings are not explored.
- Should carefully evaluate baselines within the same threat model.
- Evaluate attacks under well-motivated constraints (e.g., total local runtime of attacks)

UNIVERSITY OF MARYLAND
Cornell University
UNIVERSITY of VIRGINIA
COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK