LRS: Enhancing Adversarial Transferability through Lipschitz Regularized Surrogate

Tao Wu[1]; Tie Luo[1*]; Donal C. Wunsch II[2]
Missouri University of Science and Technology
[1]Department of Computer Science, [2]Department of Electrical and Computer Engineering

BACKGROUND & KEY TAKEAWAY

- **Adversarial examples (AE)** are created by adding human imperceptible perturbations to benign inputs to induce misclassifications.
- **Adversarial transferability**: AE created on surrogate models (source; white-box) can also fool target models (black-box).
- **Objective:** Improve adversarial transferability (more transferable AE).
- **Key Takeaway:** 1) Instead of designing AE creation algorithms on a given surrogate model (the vast majority of existing work), **transform surrogate** models toward flatter and smoother loss landscape (characterized by smaller local Lipschitz constant) and stronger adversarial robustness.
  2) LRS acts as a **"cushion"**: existing AE creation algorithms can run on LRS-transformed surrogates **w/o any modification**, yet attaining **much improved** performance (i.e., generating AE that are more transferable).

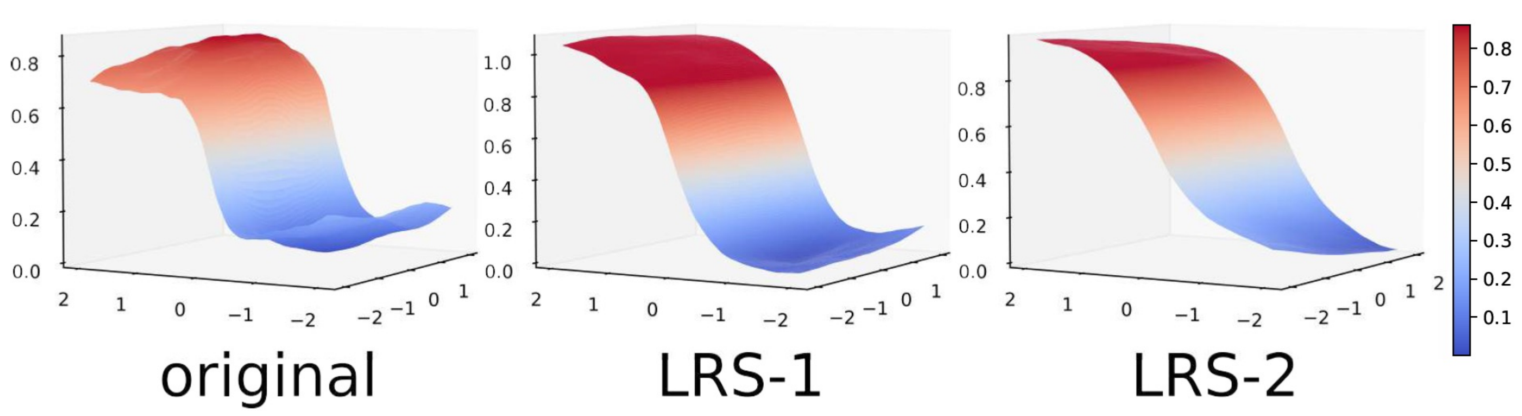
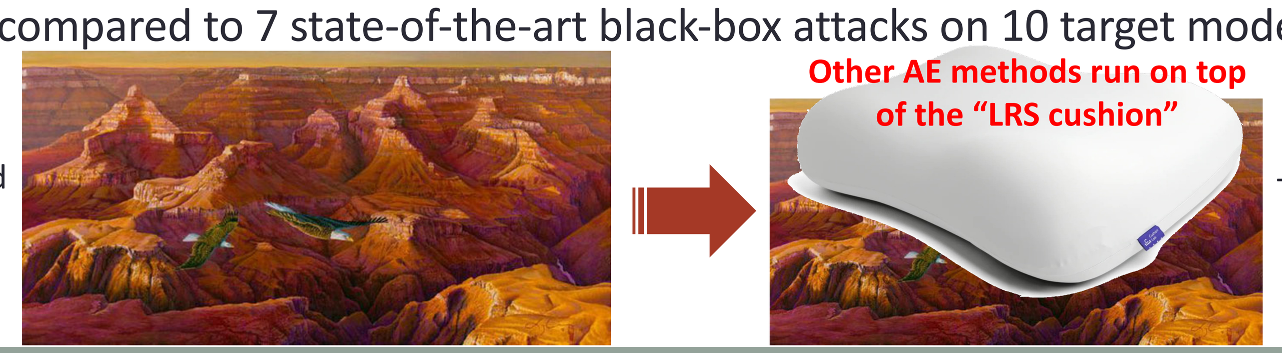


Figure 1. Loss landscape of original (corrugated) and transformed (smooth) surrogate model. Transformed surrogate models offer more stable input gradients and more generalizable AE, enabling more potent attacks.

CONTRIBUTIONS

- **LRS is a ``cushion'' method:** It **transforms** surrogate models (rather than taking them as is) such that *any existing transfer-based black-box AE generation methods* can simply run on LRS-transformed surrogate models w/o any change yet achieving much better performance.
- We identify three properties of surrogate models---smaller local Lipschitz constant, smoother loss landscape, and stronger adversarial robustness---and provide theoretical and empirical explanations of their relationship and how they favor adversarial transferability.
- We conduct extensive evaluation on ImageNet and demonstrate that, by applying LRS to even a basic AE generation method (PGD), it yields superior adversarial transferability (**>7%** abs. improvement on average) compared to 7 state-of-the-art black-box attacks on 10 target models.



RESOURCES AND CONTACT

- **Paper:** https://arxiv.org/abs/2312.13118
- **Code:** https://github.com/TrustAIoT/LRS
- **Contact:** wuta@mst.edu, tluo@mst.edu, dwunsch@mst.edu

METHODS

- **LRS-1: Lipschitz Regularization on the First Order of Loss Landscape**

$$L(x,y)=\ell(x,y)+\lambda_1\left\|\nabla_x \ell(x,y)\right\|_2^2$$

- **LRS-2: Lipschitz Regularization on the Second Order of Loss Landscape**

$$L(x,y)=\ell(x,y)+\lambda_2\left\|\nabla_x^2 \ell(x,y)\right\|_2^2$$

- **LRS-F: sum of the two regularization terms applied to the loss function**

- In view of high-dimensional data, approximate using *finite difference method* (FDM):

$$\left\|\nabla_x \ell(x,y)\right\|_2^2 \approx \left(\frac{\ell(x+h_1 d,y)-\ell(x,y)}{h_1}\right)^2$$

$$\left\|\nabla_x^2 \ell(x,y)\right\|_2^2 \approx \left(\frac{\nabla_x \ell(x+h_2 d,y)-\nabla_x \ell(x,y)}{h_2}\right)^2$$

Algorithm 1: LRS-1 (using PGD as an example base)

Input: A clean sample x with ground-truth label y; a pretrained surrogate model $f(\cdot)$;
Hyper-parameters: Finetune epochs n; batch size m; learning rate η; training dataset D; step size h; perturbation size ϵ; maximum iterations T; regularization coefficient λ
Output: A transferable AE x^{adv}

1: Pretrained surrogate model f_0 with weight w_0
2: **for** epoch $=0$ to $n-1$ **do**
3: **for** t $=0$ to $len(D)/m$ **do**
4: sample minibatch $\{(x_i, y_i)\}_{i=1,\ldots,m}$
5: $g_i=\nabla_x \ell\left(x_i, y_i; w_t\right)$
6: $d_i=\operatorname{sign}(g_i)$
7: $z_i=x_i+h d_i$
8: $\mathcal{L}(w_t)=\sum_{i=1}^m \ell\left(x_i, y_i; w_t\right)$
9: $\mathcal{R}(w_t)=\sum_{i=1}^m\left(\ell\left(z_i, y_i; w_t\right)-\ell\left(x_i, y_i; w_t\right)\right)^2$
10: $w_{t+1}=w_t-\frac{1}{m}\eta \nabla_w\left(\mathcal{L}\left(w_t\right)+\frac{1}{h^2}\lambda \mathcal{R}\left(w_t\right)\right)$
11: **save** finetuned surrogate model f_n with weight w_n
12: $\alpha=\epsilon/T$; $x_0^{adv}=x$
13: **for** $t=0$ to $T-1$ **do**
14: $g_t=\nabla_x \ell(x, w_n)$
15: $x_{t+1}^{adv}=x_t^{adv}+\alpha\cdot\operatorname{sign}(g_t)$
16: $x_{t+1}^{adv}=\operatorname{clip}\left(x_{t+1}^{adv}, 0, 1\right)$
17: **return** $x^{adv}=x_T^{adv}$

RESULTS

Method	ResNet-50*	VGG-19	ResNet-152	Inception v3	DenseNet	MobileNet
PGD (2018)	100.00%	39.22%	29.18%	15.60%	35.58%	37.90%
TIM (2019)	100.00%	44.98%	35.14%	22.21%	46.19%	42.67%
SIM (2020)	100.00%	53.30%	46.80%	27.04%	54.16%	52.54%
LinBP (2020)	100.00%	72.00%	58.62%	29.98%	63.70%	64.08%
Admix (2021)	100.00%	57.95%	45.82%	23.59%	52.00%	55.36%
TAIG (2022)	100.00%	54.32%	45.32%	28.52%	53.34%	55.18%
ILA++ (2022)	99.96%	74.94%	69.64%	41.56%	71.28%	71.84%
LRS-1 (ours)	100.00%	76.02%	72.36%	42.01%	71.23%	69.36%
LRS-2 (ours)	100.00%	78.24%	75.96%	46.14%	73.01%	73.45%
LRS-F (ours)	100.00%	**80.64%**	**78.21%**	**50.10%**	**75.19%**	**76.24%**

Method	SENet	ResNeXt	WRN	PNASNet	MNASNet	**Average**
PGD (2018)	17.66%	26.18%	27.18%	12.80%	35.58%	27.69%
TIM (2019)	22.47%	32.11%	33.26%	21.09%	39.85%	34.00%
SIM (2020)	27.04%	41.28%	42.66%	21.74%	50.36%	41.69%
LinBP (2020)	41.02%	51.02%	54.16%	29.72%	62.18%	52.65%
Admix (2021)	30.28%	41.94%	42.78%	21.91%	52.32%	42.40%
TAIG (2022)	24.82%	38.36%	42.16%	17.20%	54.90%	41.41%
ILA++ (2022)	53.12%	65.92%	65.64%	44.56%	70.40%	62.89%
LRS-1 (ours)	54.27%	66.85%	67.21%	45.29%	72.03%	64.53%
LRS-2 (ours)	57.19%	69.48%	71.13%	48.39%	75.68%	67.57%
LRS-F (ours)	**59.68%**	**71.96%**	**74.61%**	**52.43%**	**76.87%**	**69.91%**

Table 1. Attack success rates of transfer-based untargeted attacks on ImageNet using ResNet-50 as the surrogate model and PGD as the base attack method.

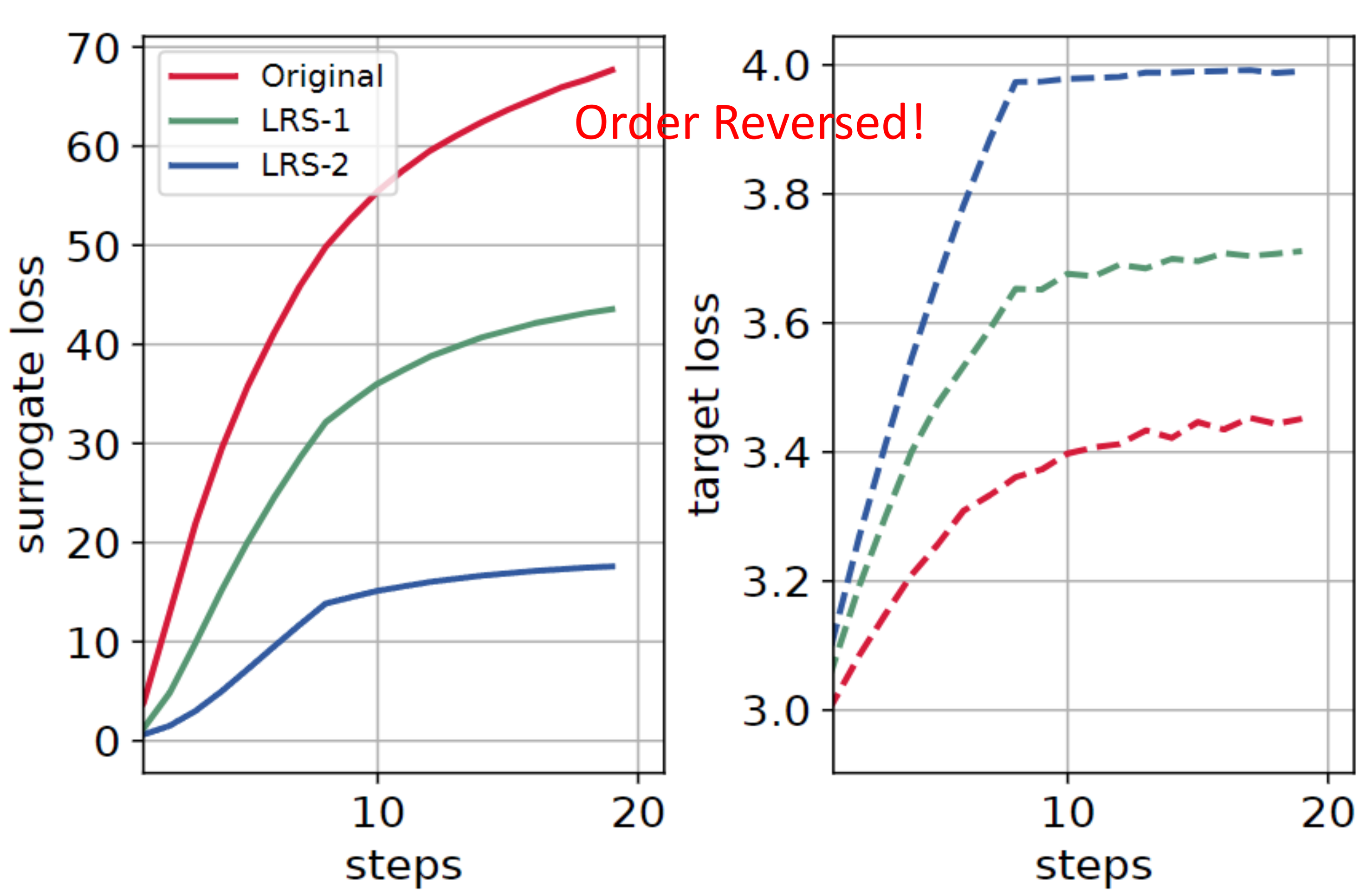


Figure 2. Loss of surrogate model (DenseNet100) and target model (ResNet18) under PGD-based attacks. It reveals that LRS-transformed surrogate exhibits stronger robustness and, in turn, enables more transferable (potent) attacks.

Surrogate model	DenseNet100	ResNet50
Original pretrained	5.53	976.59
Transformed by LRS-1	0.79	57.62
Transformed LRS-2	0.67	53.21
Transformed LRS-F	**0.59**	**49.64**

Table 2. Smoothness quantified by *empirical local Lipschitz constant*. DenseNet100 is evaluated on CIFAR10 and ResNet50 is evaluated on ImageNet.

$$L_{emp}=\frac{1}{n}\sum_{i=1}^{n}\max_{\boldsymbol{x}_i'\in\mathbb{B}_\infty(\boldsymbol{x}_i,\epsilon)}\frac{\left\|f\left(\boldsymbol{x}_i\right)-f\left(\boldsymbol{x}_i'\right)\right\|_2}{\left\|\boldsymbol{x}_i-\boldsymbol{x}_i'\right\|_2}$$