Full version available at:
https://arxiv.org/abs/2301.13862

Salient Conditional Diffusion for Defending Against Backdoor Attacks

Brandon May* Joseph Tatro* Piyush Kumar Nathan Shnidman

Vision and Image Understanding Group

STR

Woburn, MA

DOCUMENT RESTRICTIONS

*Equal contribution

Corresponding author: joseph.tatro@str.us

Salient Conditional Diffusion (Sancdifi)

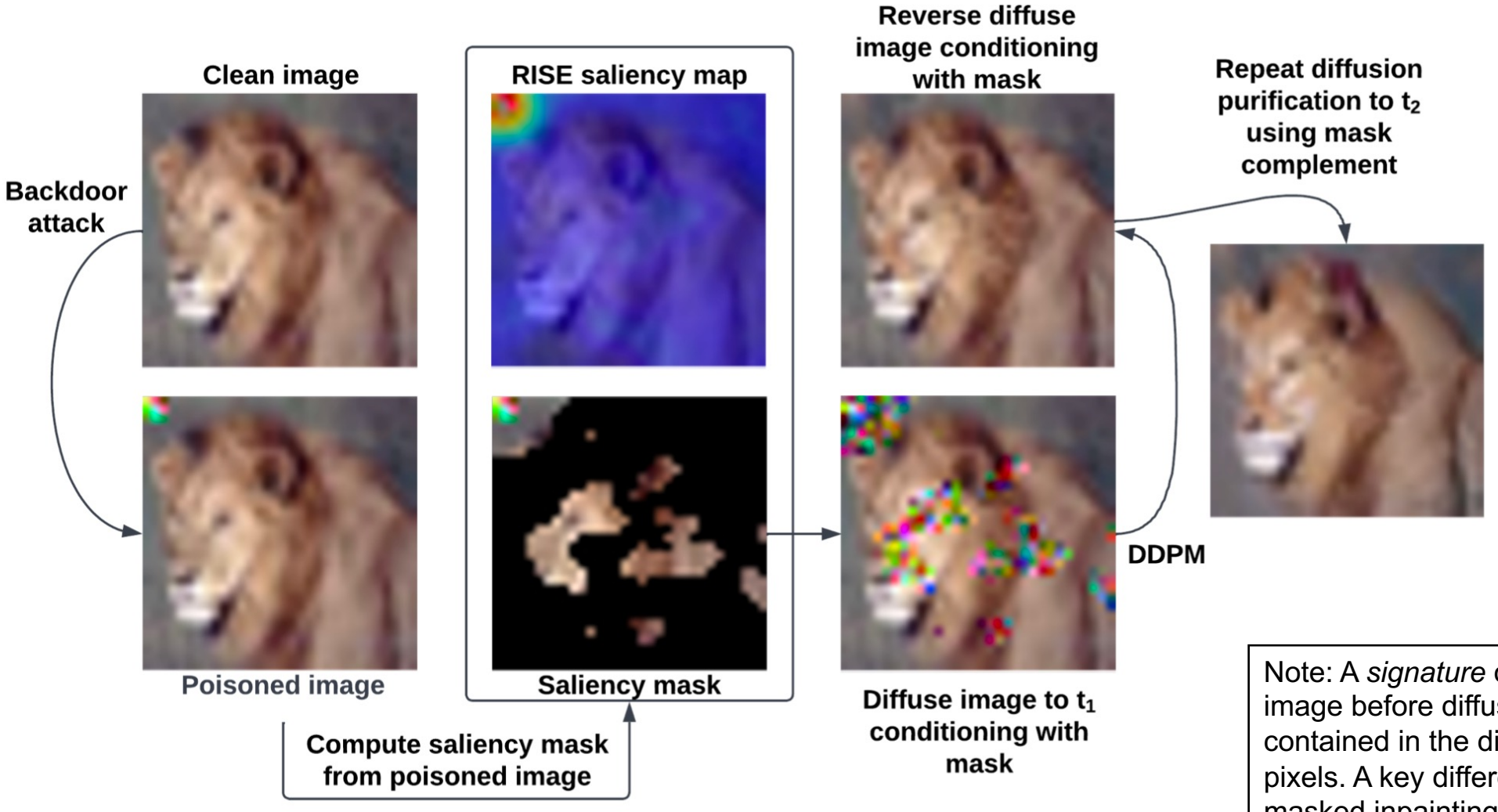

Illustration of our defense algorithm.

Note: A *signature* of the image before diffusion is contained in the diffused pixels. A key difference from masked inpainting.

Key Takeaways

- **Sancdifi is a <u>black-box</u> defense**
  - The RISE algorithms allows for a black-box computation of saliency maps
    - Zeroth-order approximation to gradient
- **Sancdifi also defends against traditional PGD attacks**
  - The second diffusion purification on the mask complement fuels this robustness
  - Key distinction over related methods such as *Februus*
    - Attack success rate of PGD attacks for CIFAR-10/ResNet-50 models are:
    - 11% (Ours) vs. 88% (Februus)
- **Conditioning via saliency is critical for the success of diffusion purification**
- **This work was supported by the DARPA AIE program, Geometries of Learning (HR00112290078)**

Without salient conditioning (right), the structure of this infant from TinyImageNet is severely degraded by diffusion which hurts accuracy. Conversely, diffusing for less time increases attack success as the trojan trigger is not completely diffused out.

Table 1: **Sancdifi** (SD) results on BadNet for ResNet-50. Our metrics include clean accuracy reduction (CAR) and attack success rate (ASR) for top-1 and top-5 class performance. **Sancdifi** outperforms the other reformation algorithms, manifold projection (MP) and Februus (FB). Our algorithm CAR outperforms adversarial retraining (AR), while our top-1 ASR is competitive with both adversarial retraining and fine-pruning (FP).

		top-1					top-5			
Dataset	**Metric**	**SD**	AR	FP	MP	FB	**SD**	AR	FP	MP
CIFAR-10	CAR	2.0	6.0	-1.0	-1.0	13.0	0.0	0.0	0.0	0.0
	ASR	12.0	9.0	36.0	100.0	11.0	55.0	41.0	95.0	100.0
CIFAR-100	CAR	18.0	20.0	15.0	6.0	—	11.0	11.0	5.0	3.0
	ASR	0.0	1.0	1.0	33.0	—	7.0	4.0	3.0	91.0
Tiny ImageNet	CAR	7.0	27.0	0.0	2.0	—	5.0	28.0	0.0	1.0
	ASR	3.0	0.0	1.0	99.0	—	7.0	2.0	6.0	99.0