SALIENT CONDITIONAL DIFFUSION FOR BACKDOORS

Brandon May†, N. Joseph Tatro†, Nathan Shnidman, & Piyush Kumar
STR
Woburn, MA 01801, USA
joseph.tatro@str.us

ABSTRACT

We propose a novel algorithm, **Salient Conditional Diffusion** (**Sancdifi**), a state-of-the-art defense against backdoor attacks. **Sancdifi** uses a diffusion model (DDPM) to degrade an image with noise and then recover it. Critically, we compute saliency map-based masks to condition our diffusion, allowing for stronger diffusion on the most salient pixels by the DDPM. As a result, **Sancdifi** is highly effective at diffusing out triggers in data poisoned by backdoor attacks. At the same time, it reliably recovers salient features when applied to clean data. **Sancdifi** is a black-box defense, requiring no access to the Trojan network parameters.

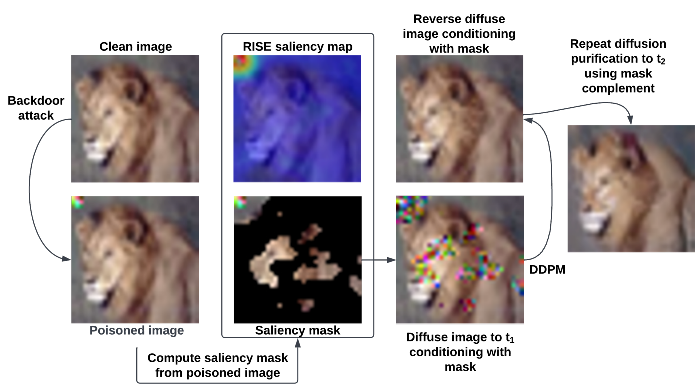

Figure 1: **Sancdifi**: Given a possibly backdoor attacked image, we compute saliency maps via RISE, and use the top-5 class maps to construct a mask, $\boldsymbol{A}$. Notice the trigger is left unmasked. We then apply diffusion purification (DP) conditioned with the saliency mask. Following this, we reapply DP using the reverse mask, $\boldsymbol{I} - \boldsymbol{A}$. **Sancdifi** diffuses out the trigger without large degradation.

1 INTRODUCTION AND RELATED WORK

As machine learning develops, attention is being given to sophisticated attacks aligning closely to practical use cases. We consider backdoor attacks, like *BadNet* (Gu et al., 2017), that are challenging to defend against (Li et al., 2020). The attack poisons data with a visual trigger so that a malicious classifier will purposefully misclassify it, allowing an adversary precision. In this work, we:

1. Propose a novel defense against backdoor attacks, **Sancdifi**, that *purifies* input by diffusing and denoising it with a diffusion model (DDPM) conditioned on a mask derived from saliency maps.
2. Establish state-of-the-art performance among backdoor defense algorithms. While **Sancdifi** is a black-box defense algorithm, it achieves performance competitive with state-of-the-art white-box defenses such as adversarial retraining (Madry et al., 2017) and fine-pruning (Liu et al., 2018a).

Backdoor Attacks Surveyed in TrojanZoo (Pang et al., 2022), there are several types of defenses against backdoor attacks. There are black-box input reformation defenses like our algorithm and

†Equal Contribution

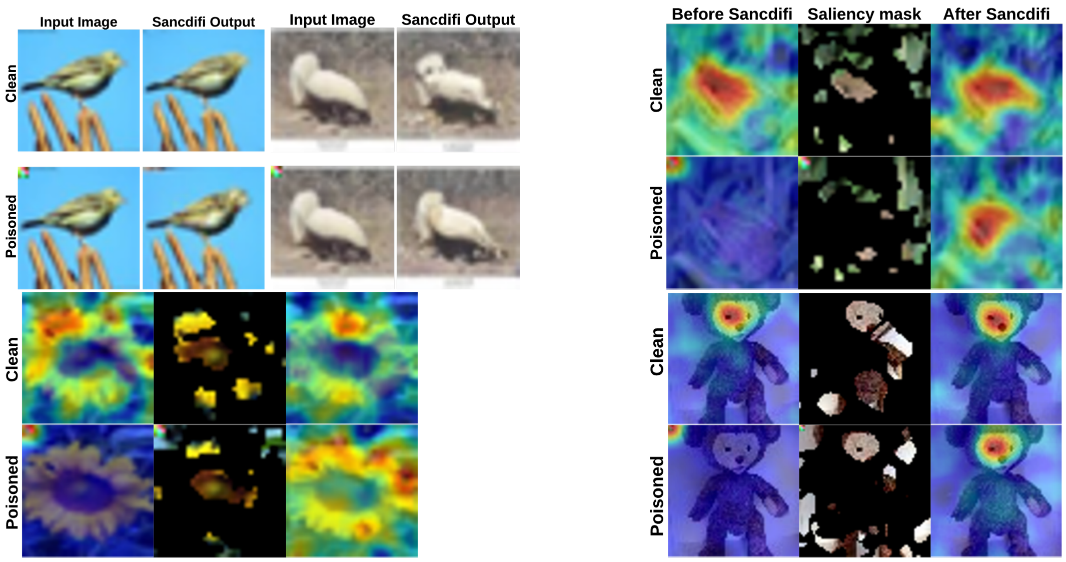

Figure 2: **Upper Left: Sancdifi** defense against BadNet attacks on CIFAR-10 and CIFAR-100 for ResNet-50. Top/bottom rows display our method operating on clean/BadNet attacked images. **Right:** Computed saliency masks. Each column displays; the top class saliency map, the computed saliency mask, and the post-**Sancdifi** saliency map. We have removed the trigger and its saliency.

Algorithm 1 Salient Conditional Diffusion algorithm with image $\boldsymbol{x}$, Trojan network f, N RISE masks, timesteps $\{T_1, T_2\}$, saliency percentile cutoff d, and r of top-r performance.

$\mathbb{C} \leftarrow$ top-k$(f(\boldsymbol{x}), r)$ indices
$\mathbb{S} \leftarrow \{\text{RISE}(\boldsymbol{x}, f, N, c), \quad c \in \mathbb{C}\}$ ▷ see (Petsiuk et al., 2018) for RISE algorithm
$\boldsymbol{M} \leftarrow \{\boldsymbol{S}_i \leq percentile(\boldsymbol{S}_i, d), \quad \boldsymbol{S}_i \in \mathbb{S}\}$
$\boldsymbol{A} \leftarrow \prod_i \boldsymbol{M}_i, \quad \boldsymbol{M}_i \in \boldsymbol{M}$
for i in $\{1, 2\}$ **do**
  $\boldsymbol{z} \leftarrow$ sample $q(\boldsymbol{x}_{T_i} | \boldsymbol{x}_0)$
  $\hat{\boldsymbol{x}}_{T_i} \leftarrow \boldsymbol{A}\boldsymbol{x}_0 + (\boldsymbol{I} - \boldsymbol{A})\boldsymbol{z}$
  for t in $\{T_i, T_{i-1}, ..., 0\}$ **do**
    $\boldsymbol{z} \leftarrow$ sample $p(\hat{\boldsymbol{x}}_{T_{i-1}} | \hat{\boldsymbol{x}}_{T_i})$ ▷ trained DDPM parameterizing $p(\hat{\boldsymbol{x}}_{T-1} | \hat{\boldsymbol{x}}_T)$
    $\hat{\boldsymbol{x}}_{T_{i-1}} \leftarrow \boldsymbol{A}\hat{\boldsymbol{x}}_{T_i} + (\boldsymbol{I} - \boldsymbol{A})\boldsymbol{z}$
  $\boldsymbol{A} \leftarrow \boldsymbol{I} - \boldsymbol{A}$

manifold projection (MP) (Meng & Chen, 2017). State-of-the-art defenses include two model sanitization defenses, adversarial retraining (AR) (Madry et al., 2017) and fine-pruning (FP) (Liu et al., 2018a). These alter the Trojan model and are white-box. We also compare to the Februus algorithm (FB), which uses a GAN to inpaint an image after saliency-based masking (Doan et al., 2020).

Diffusion Models DDPMs (Ho et al., 2020) are a recent popular generative model. Of interest, Nie et al. (2022) and Wu et al. (2022) both proposed *diffusion purification*, using DDPMs to defend against PGD attacks. Consider that a defense for PGD attacks is not necessarily valid for backdoor attacks (Weng et al., 2020). Others have guided DDPMs with deep features (Dhariwal & Nichol, 2021; Voynov et al., 2022) Unlike these methods, our work does not require user input, such as class.

2 Salient Conditional Diffusion for Backdoor Defense

To the best of our knowledge, this work is the first to propose the use of diffusion models (DDPMs) as a defense against backdoor attacks. The main novel contribution of **Sancdifi** is the use of saliency masks for conditioning diffusion purification. We begin by stating our attack model.

Figure 1, illustrating the salient conditional diffusion process, displays a typical BadNet trigger, a small 3x3 patch. Concretely, a backdoor trigger is a pattern, $p(\boldsymbol{x})$, that may depend on the data,

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

$\boldsymbol{x} \in \mathbb{R}^d$, as well as hyperparameters such as transparency. Trojan networks are trained to handle both poisoned data and clean data. The clean data is associated with a label, y, while the target label for poisoned data is t. In this work, we use BadNet (Gu et al., 2017) and TrojanNN (Liu et al., 2018b) as our attack models. In BadNet, the trigger $\boldsymbol{r}$ is fixed, while TrojanNN optimizes the pixel colors of the trigger to maximize certain neuron activations.

Methodology A core component of our algorithm is the use of saliency to condition diffusion. To be clear, a saliency map $\boldsymbol{S}_k$ for a given image $\boldsymbol{x}$, class k, and classifier network f measures the importance of each pixel of $\boldsymbol{x}$. This importance is relative to $f's$ determination of the k-class probability of $\boldsymbol{x}$. We compute the maps using the black-box RISE algorithm (Petsiuk et al., 2018).

Given input image $\boldsymbol{x}$, **Sancdifi** starts by computing the RISE saliency maps of $\boldsymbol{x}$ for the top r classes determined by the Trojan network f_θ. Examples of RISE saliency maps can be seen in the right of Figure 2. The most probable saliency map for clean images highlights meaningful areas. In contrast, the top saliency map for BadNet-poisoned images has the strongest response on the trigger. From the saliency map $\boldsymbol{S}_k$, we threshold the top d percentile of values to create a k-class saliency mask, $\boldsymbol{M}_k$. We desire robust performance over different metrics such as top-5 accuracy. With that in mind, given the set of masks for the top-r most probable classes $\mathbb{S}_{\boldsymbol{M}}$, we can define a composite saliency mask $\boldsymbol{A}$ as their elementwise product. Formal definitions are visible in Algorithm 1.

We use $\boldsymbol{A}$ to condition our diffusion processes. Intuitively, the composite mask ignores all but the most salient pixels of the most likely classes. Our method of diffusion is taken from OpenAI's improved-diffusion DDPM (Nichol & Dhariwal, 2021). Given a trained DDPM, there is an associated conditional distribution for forward diffusion, $q(\boldsymbol{x}_t|\boldsymbol{x}_0)$, and the learned prior distribution for backward diffusion, $p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$. For brevity, we omit their definitions referring the reader to Nichol & Dhariwal (2021). The diffusion conditioned with mask is formalized in Algorithm 1. Following the first diffusion purification step, we reapply diffusion purification using the complement of our salient mask, $\boldsymbol{I} - \boldsymbol{A}$, with time $\hat{t}$ where $\hat{t} < t$. Our reason for doing this is to safeguard against other attacks with support across the entire image.

3 Numerical Experiments

With our algorithm defined, we outline the setting of our experiments. We concern ourselves with image classification in the presence of backdoor attacks. We use three **datasets**: CIFAR-10, CIFAR-100, (Krizhevsky et al., 2009), and Tiny-ImageNet (Le & Yang, 2015). Any metrics reported are for the dataset validation subsets. Additionally, we use three **architectures**; ResNet-50 (He et al., 2016), EfficientNet-B7 (Tan & Le, 2019), and a transformer ViT-Base-16 (Dosovitskiy et al., 2020).

Concerning our algorithm, we use pretrained DDPMs from OpenAI's improved-diffusion repository (Nichol & Dhariwal, 2021). For the CIFAR datasets, we diffuse out to 300 time steps for the first

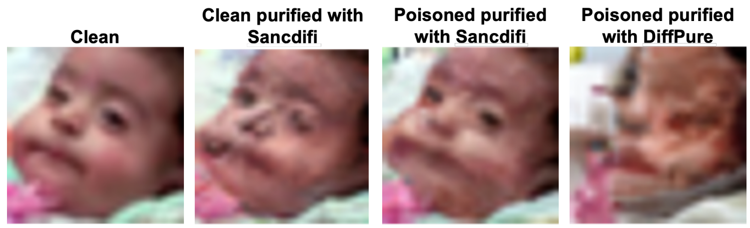

Figure 3: We display a clean image, it purified with **Sancdifi**, and the attacked image purified with **Sancdifi** and DiffPure respectively. Without salient conditioning, the face in the image is destroyed.

diffusion purification. For Tiny-ImageNet, we use 450 time steps as we find that it is needed to sufficiently defend against BadNet attacks. For the second diffusion purification step using the complement mask, $\boldsymbol{I} - \boldsymbol{A}$, we diffuse out to 100 time steps. Our backdoor attacks are generated using the TrojanZoo suite (Pang et al., 2022) with their default parameters. Regarding saliency, the number of binary masks used to compute RISE maps was set at 2000. For the saliency thresholding cutoff, we set a value of 95%. The composite saliency map is aggregated across the top-5 classes to align us with top-5 validation metrics. We compare with the other defenses stated in Section 1.

Performance on BadNet Attack Table 1 contains the results of defending against BadNet attacks on ResNet-50 using **Sancdifi**. Performance is given in terms of clean accuracy reduction (CAR) and attack success rate (ASR). While we focus mainly on top-1 classification accuracy, we include top-5 classification results. To be clear, CAR denotes the reduction in accuracy on clean images after applying the defense. Intuitively, we desire CAR and ASR to be low. Clearly **Sancdifi** outperforms the other black-box defenses. Additionally, its competitive with adversarial retraining and fine-pruning. The winner among our method and the last two is largely a question of the tradeoffs between CAR and ASR as well as top-1 and top-5 performance. Importantly, its black-box style gives our defense wider applicability. The **Sancdifi** defense is visible in Figure 2. The BadNet trigger has clearly been diffused after purification. In contrast, the other salient regions of the images not covered by salient mask $\boldsymbol{A}$ have been reliably recovered by the DDPM.

To further validate our performance, we repeat the previous experiment for our other architectures. The results can be found in Table 2 in the Appendix. The behavior is similar to Table 1, showing that our performance generalizes to various classes of neural networks. We also show that our performance generalizes to other backdoor and adversarial attacks in Table 3. Regarding the Invisible BadNet attack on CIFAR-10, we find that **Sancdifi** has a CAR/ASR of (3.0% \11.0%) compared to Februus with (2.0% \88.0%). Thus our algorithm works where inpainting methods fail.

Impact of Saliency Masks Naively, one might assume that vanilla diffusion purification a la DiffPure (Nie et al., 2022) is sufficient against backdoor attacks. Table 4 in the Appendix provides results on ResNet-50 where we have performed no salient thresholding and omit the second diffusion purification step. Strikingly, CAR is much higher without salient masking. Notably, DiffPure at 30% has the highest CAR across all defenses for the CIFAR-100 dataset. A comparison of the output with and without salient conditioning is visible in Figure 3. We can see that while it is not the most salient, the masked part of the image offers a strong prior for the DDPM. This prior allows us to reliably recover the unmasked part of the image excluding the backdoor trigger. The mask prevents image-wide degradation that we can see in DiffPure. The second diffusion purification step with the mask complement is necessary to safeguard against backdoor attacks with larger triggers such as Invisible BadNet. This attack has image-wide support and is L_∞-bounded to prevent perceptability.

4 Conclusion and Acknowledgements

Salient conditional diffusion, **Sancdifi**, is a state-of-the-art black-box defense against backdoor attacks with wide generalization. Salient conditioning plays a major role in diffusing out backdoor triggers while preventing massive degradation to other parts of an image. We believe conditional diffusion will play a strong role in the future in defending against backdoor attacks. This work was supported by the DARPA AIE program, Geometries of Learning (HR00112290078).

REFERENCES

Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis, 2021. URL https://arxiv.org/abs/2105.05233.

Bao Gia Doan, Ehsan Abbasnejad, and Damith C. Ranasinghe. Februus: Input purification defense against trojan attacks on deep neural network systems. In *Annual Computer Security Applications Conference*. ACM, dec 2020. doi: 10.1145/3427228.3427264. URL https://doi.org/10.1145%2F3427228.3427264.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2020. URL https://arxiv.org/abs/2010.11929.

Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain, 2017. URL https://arxiv.org/abs/1708.06733.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2016. doi: 10.1109/cvpr.2016.90. URL http://dx.doi.org/10.1109/cvpr.2016.90.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Alex Krizhevsky et al. Learning multiple layers of features from tiny images. 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. 2015. URL http://vision.stanford.edu/teaching/cs231n/reports/2015/pdfs/yle_project.pdf.

Yiming Li, Baoyuan Wu, Yong Jiang, Zhifeng Li, and Shutao Xia. Backdoor learning: A survey. *IEEE transactions on neural networks and learning systems*, PP, 2020.

Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Fine-pruning: Defending against backdooring attacks on deep neural networks. *Lecture Notes in Computer Science*, pp. 273–294, 2018a. ISSN 1611-3349. doi: 10.1007/978-3-030-00470-5_13. URL http://dx.doi.org/10.1007/978-3-030-00470-5_13.

Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Weihang Wang, and X. Zhang. Trojaning attack on neural networks. In *Network and Distributed System Security Symposium*, 2018b.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Dongyu Meng and Hao Chen. Magnet. *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, Oct 2017. doi: 10.1145/3133956.3134057. URL http://dx.doi.org/10.1145/3133956.3134057.

Alex Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models, 2021. URL https://arxiv.org/abs/2102.09672.

Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Anima Anandkumar. Diffusion models for adversarial purification, 2022. URL https://arxiv.org/abs/2205.07460.

Ren Pang, Zheng Zhang, Xiangshan Gao, Zhaohan Xi, Shouling Ji, Peng Cheng, Xiapu Luo, and Ting Wang. Trojanzoo: Towards unified, holistic, and practical evaluation of neural backdoors. *2022 IEEE 7th European Symposium on Security and Privacy*, Jun 2022. doi: 10.1109/eurosp53844.2022.00048. URL http://dx.doi.org/10.1109/EuroSP53844.2022.00048.

Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*, 2018.

Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pp. 6105–6114. PMLR, 2019.

Andrey Voynov, Kfir Aberman, and Daniel Cohen-Or. Sketch-guided text-to-image diffusion models, 2022. URL `https://arxiv.org/abs/2211.13752`.

Cheng-Hsin Weng, Yan-Ting Lee, and Shan-Hung Wu. On the trade-off between adversarial and backdoor robustness. In *Neural Information Processing Systems*, 2020.

Quanlin Wu, Hang Ye, and Yuntian Gu. Guided diffusion model for adversarial purification from random noise, 2022. URL `https://arxiv.org/abs/2206.10875`.

A Appendix

Table 2: **Sancdifi** (SD) defense results on BadNet for CIFAR100 for other networks. Other methods and metrics defined are in Table 1. Our performance extends to architectures beyond ResNet-50, with top-1 ASR comparable to white-box defenses, adversarial retraining and fine-pruning.

		Backdoor Defenses							
		top-1				top-5			
Network	**Metric**	**SD**	AR	FP	MP	**SD**	AR	FP	MP
Efficient-Net	CAR	16.0	26.0	18.0	-2.0	5.0	11.0	6.0	1.0
	ASR	1.0	0.0	1.0	6.0	13.0	8.0	2.0	17.0
ViT	CAR	26.0	15.0	1.0	2.0	10.0	5.0	0.0	0.0
	ASR	1.0	1.0	1.0	76.0	18.0	3.0	10.0	99.0

Table 3: **Sancdifi** (SD) defense results on other backdoor attacks for CIFAR-100 and ResNet-50. We also include PGD attacks. Clearly, our algorithm can handle other backdoor attacks such as TrojanNN as well as the traditional PGD attack. So **Sancdifi** can be used for both backdoor and adversarial robustness. Our worst scenario is the image-wide Invisible BadNet attack, though we can resolve this by running the second diffusion for longer than 100 steps.

		Backdoor Defenses							
		top-1				top-5			
Attack	**Metric**	**SD**	AR	FP	MP	**SD**	AR	FP	MP
Invisible BadNet	CAR	5.0	14.0	-3.0	4.0	10.0	7.0	-1.0	1.0
	ASR	20.0	0.0	1.0	0.0	72.0	3.0	4.0	13.0
TrojanNN	CAR	9.0	22.0	4.0	2.0	10.0	18.0	5.0	3.0
	ASR	1.0	0.0	2.0	93.0	4.0	8.0	11.0	98.0
PGD	CAR	18.0	16.0	—	1.0	10.0	11.0	—	1.0
	ASR	0.0	8.0	—	18.0	3.0	24.0	—	53.0

Table 4: Diffusion results **without** salient conditioning for ResNet-50. This reduces to the DiffPure algorithm (Nie et al., 2022). Diffusion times are denoted relative to the maximum 1000 time steps. As diffusion time increases, ASR decreases at the cost of increased CAR. At less than 30% diffusion, ASR can become too high as in the case of Tiny-ImageNet. Yet the high diffusion leads to worse CAR. Notice that in the case of CIFAR-100, CAR is much higher at 30% than our algorithm (SD) in Table 1. Thus, saliency masking is needed.

		Diffusion Times					
		top-1			top-5		
Dataset	**Metric**	10%	20%	30%	10%	20%	30%
CIFAR-10	CAR	5.0	5.0	9.0	1.0	1.0	2.0
	ASR	90.0	14.0	11.0	100.0	63.0	61.0
CIFAR-100	CAR	13.0	31.0	47.0	2.0	15.0	31.0
	ASR	42.0	1.0	0.0	82.0	3.0	3.0
Tiny ImageNet	CAR	2.0	2.0	13.0	0.0	3.0	6.0
	ASR	99.0	47.0	9.0	99.0	53.0	15.0