

DP-RAE: A Dual-Phase Merging Reversible Adversarial Example for Image Privacy Protection (Supplementary Material)

Anonymous Author(s)

A DETAILS AND ANALYSE

In our evaluation of the Threshold-Informed Superpixel Attack (TISA), we investigated the optimal frequency for recording historical data to select points for perturbation enhancement. We tested settings of m equals to 5, 10, and 20, with results detailed in Table 1. We found that recording every ten instances to select enhancement points yielded the best performance. Recording too few instances reduces the availability of effective points, diminishing the enhancement points impact. Conversely, recording too many instances, while allowing for the selection of more suitable points, decreases the frequency of enhancements. Therefore, after careful consideration, we determine that setting m to 10 is a reasonable compromise that efficiently balances point selection with enhancement frequency.

Table 1: The ASR (%) and queries (times) of TISA with different history record size m , "↑" means the bigger the better.

	$m = 5$	$m = 10$	$m = 20$
ASR ↑	95.1	95.3	95.3
Queries ↓	1507	1496	1498

Figure 1 to 3 provide more details about the images, which include the original image, DP-AE, DP-RAE, and the recovered image. Our findings reveal that despite some degradation in visual quality, both DP-AE and DP-RAE remain recognizable to the human eye. However, our experimental results demonstrate that these perturbations significantly disrupt accurate model classification. Importantly, the recovered image effectively restores the visual quality, mitigating any degradation caused by the perturbations. Thus, DP-RAE not only prevents malicious DNNs from analyzing images but also ensures their usability is not compromised. Consequently, we have successfully leveraged adversarial perturbations as a tool for preserving privacy, demonstrating that the restored images eliminate any adverse effects on authorized users.

Figures 4 to 7 present additional results of our attacks on commercial black-box models. The experimental data show that DP-RAE can deceive Baidu's cloud vision API¹ with a high success rate, even under limited query conditions. These results not only demonstrate that our method poses a significant threat in real-world applications but also indirectly confirm DP-RAE's effectiveness in enhancing user security and privacy.

¹<https://ai.baidu.com/tech/imagerecognition/general>

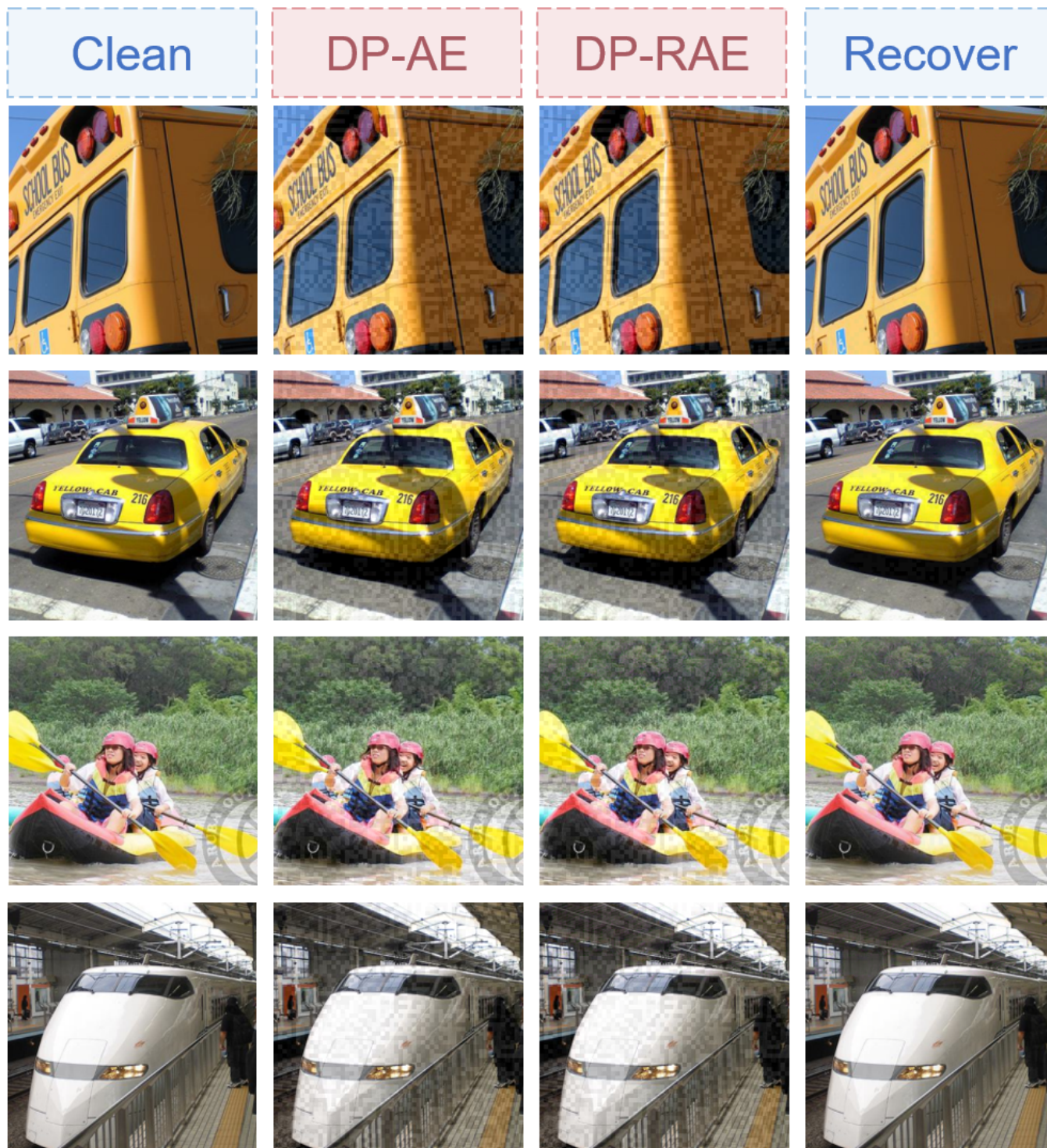


Figure 1: Visual effects before and after DP-RAE recovery.



Figure 2: Visual effects before and after DP-RAE recovery.

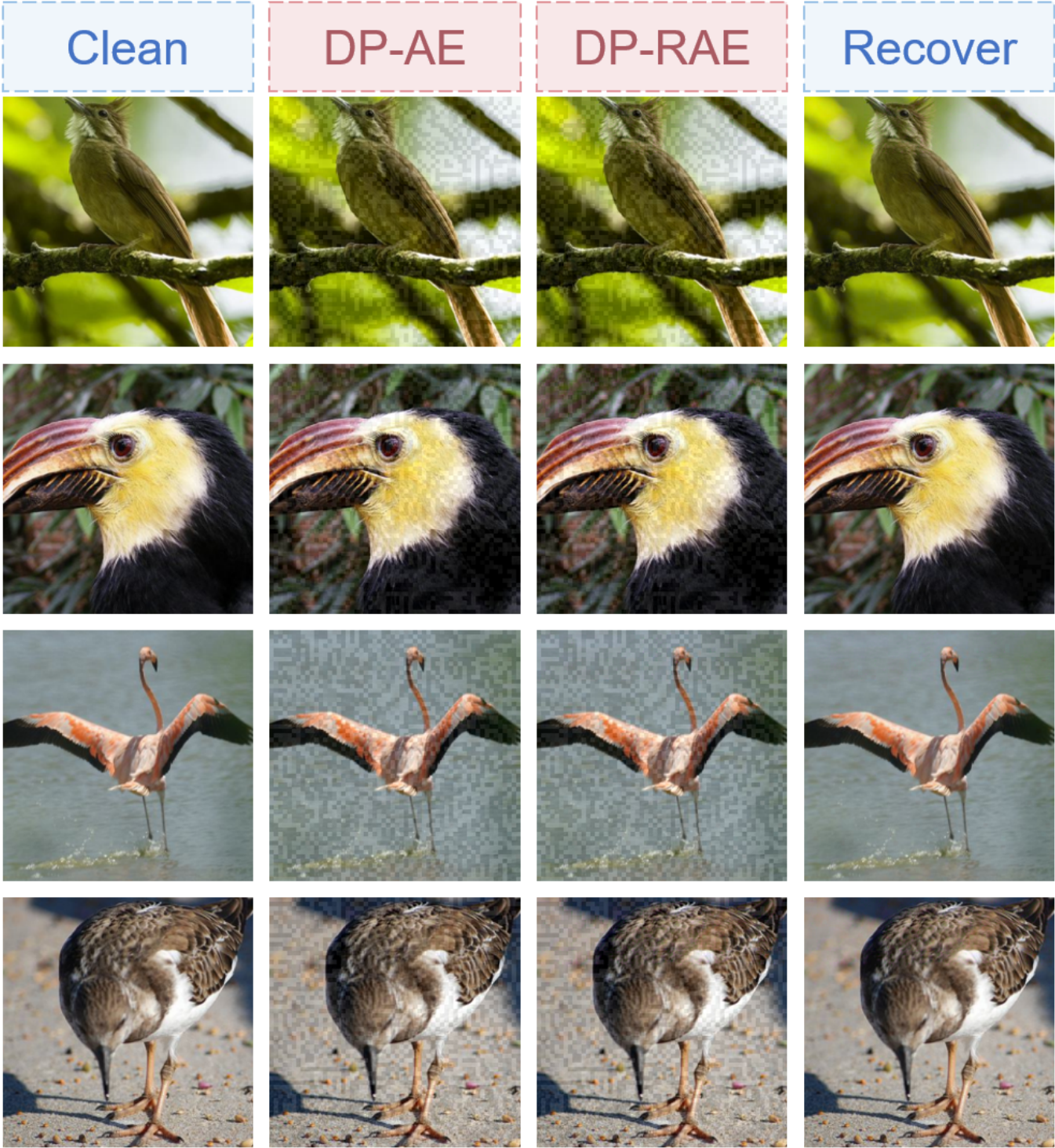


Figure 3: Visual effects before and after DP-RAE recovery.



Figure 4: In the commercial model, clean image identified as "Fig", DP-RAE misclassified as "Aplysia".

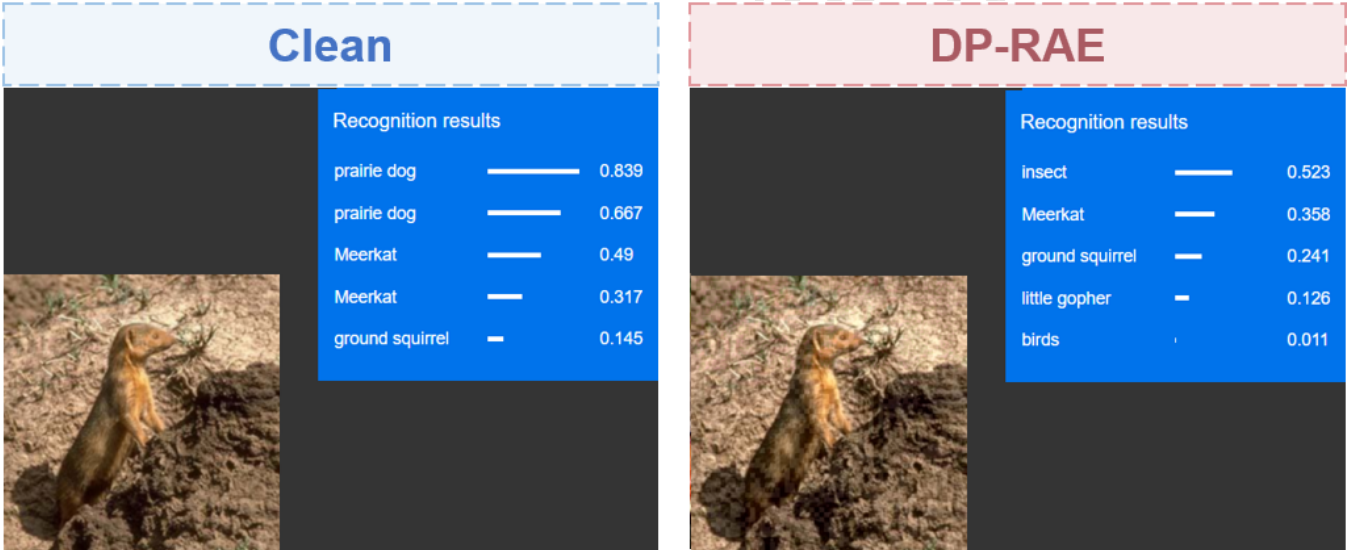


Figure 5: In the commercial model, clean image identified as a "Prairie dog", DP-RAE misclassified as "Insect".

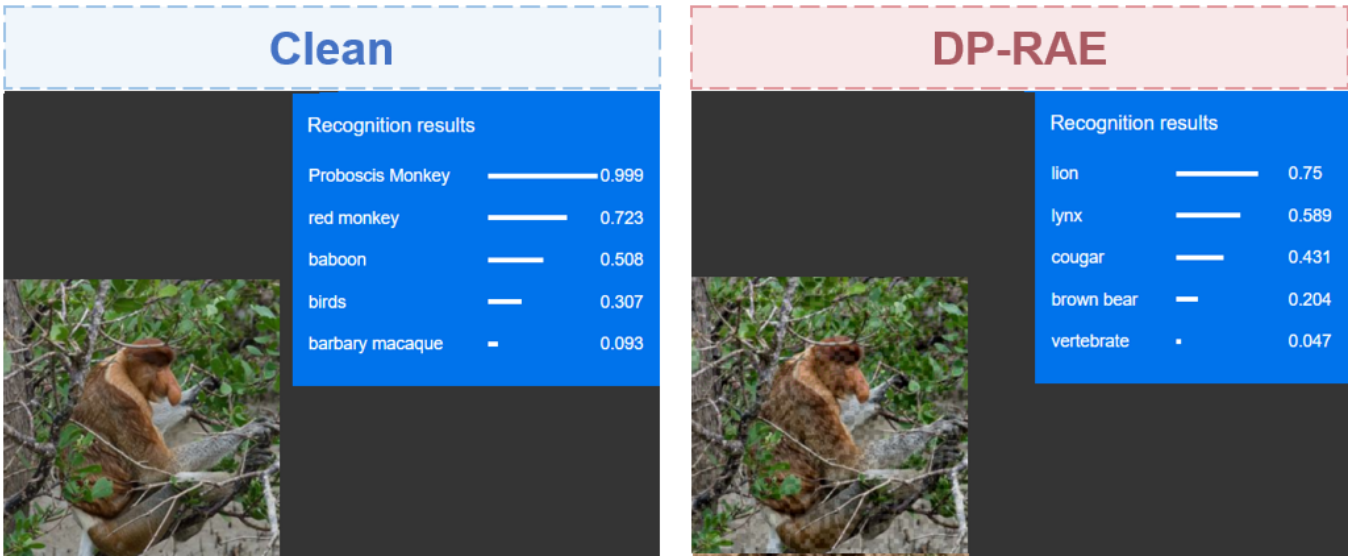


Figure 6: In the commercial model, clean image identified as a "Proboscis monkey", DP-RAE misclassified as a "Lion".

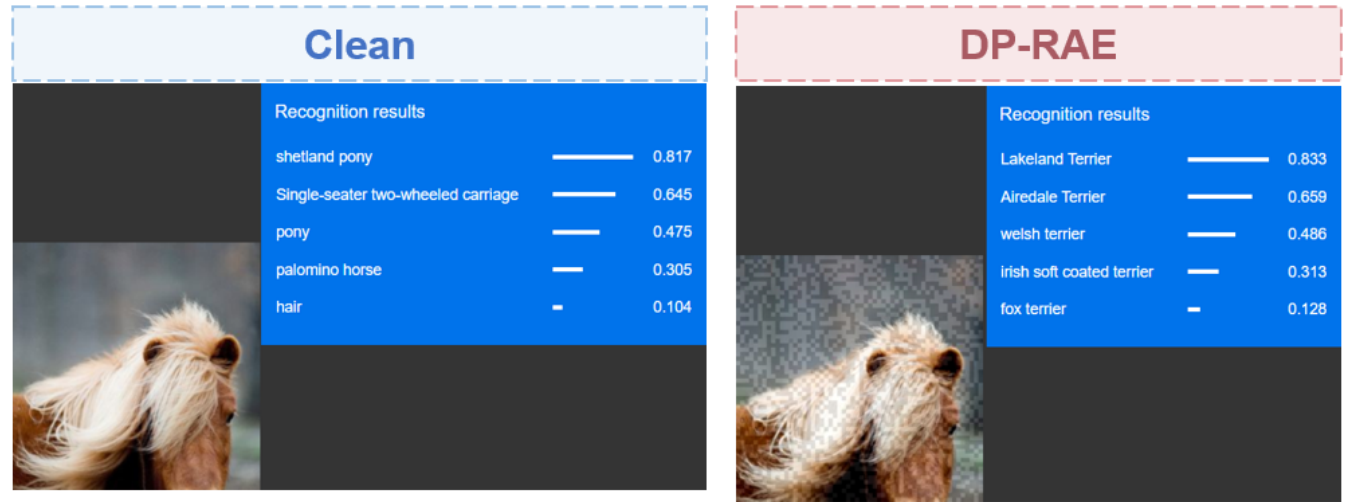


Figure 7: In the commercial model, clean image identified as a "Shetland pony", DP-RAE misclassified as a "Lakekand terrier".