RAEncoder: A Label-Free Reversible Adversarial Examples Encoder for Dataset Intellectual Property Protection

Supplementary Material

A. Extended Robustness Study

A.1. Diffusion Model Purification

Following [27]'s released code for CIFAR10 purification, Table 6 shows our work is slightly behind SRAE. However, as shown in Table 4, SRAE's robustness sacrifices visual quality (SSIM=0.42), while ours maintains higher quality (SSIM=0.85).

Table 6. Diffusion model purification results.

Attack Method	Clean	SRAE	Ours
Purified Accuracy	85.28	46.72	61.72

A.2. The Robustness of Interference Pattern

This section discusses the design of the Interference Pattern for ensuring robustness in distinguishing authorized and unauthorized users. Specifically, it should achieve this distinction without compromising RAE performance. Let $\mathcal{S}^i_{blank}$ denote the random signature noise used in this study, where i represents the similarity between the specified signature noise $\mathcal{S}$ and the random signature noise(e.g., $i = 100$ indicates complete similarity). As shown in Figure 10, the results indicate that ASR remains high (83.51) even at $i = 90$, with a significant drop only at $i = 94$ (to 66.66). These findings confirm the robustness of our design.

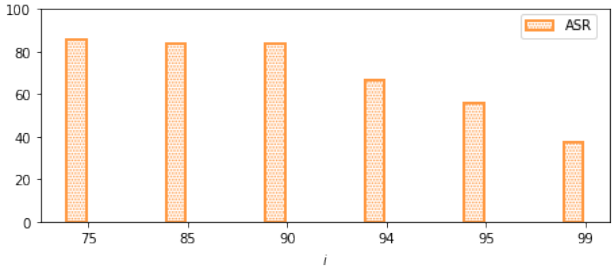

Figure 10. The ASR of RAE when using random signature noise with varying similarity to the specified signature noise as input.

B. Algorithmic Efficiency

As shown in Table 7, RAEncoder's parameter count, latency, and FLOPs on ImageNet (batch size=1) indicate room for improvement compared to SRAE. However, our work primarily addresses the critical challenge of protecting unlabeled datasets in SSL by proposing an RAE solution that balances adversarial strength and recovery capability. Algorithmic efficiency remains a secondary consideration for current methods in this domain. However, future work may explore ways to further optimize RAEncoder's computational cost while maintaining or even enhancing its robustness and recovery performance.

Table 7. Algorithm efficiency results.

Method	Params (M)	FLOPs (B)	Latency (s)
Ours	12.66	1.66	0.08
SRAE	0.18	0.84	0.04