RevINN: An End-to-End Invertible Neural Network for Reversible Adversarial Examples Generation

Supplementary Material

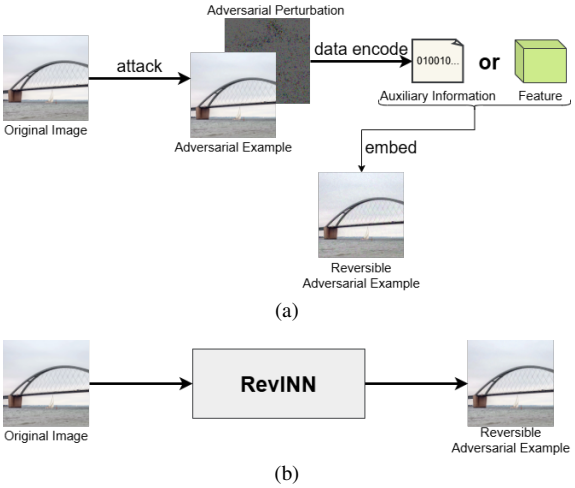

Figure 6. (a) The two-stage generation paradigm of existing methods (b) Our proposed RevINN for one stage generation

A. Paradigm Comparison with Existing Methods

Fig. 6 illustrates the paradigm of our proposed method in comparison with existing approaches. Most existing methods first generate an adversarial example and then apply RDH techniques or DNN-based encoders to embed the perturbation information reversibly into the adversarial example, thereby producing the RAE. However, this two-stage paradigm inevitably introduces both visual distortion and adversarial loss between the two separate processes, which leads to a degradation in the quality and robustness of the resulting RAE.

In contrast, we formulate RAE generation as an image-to-image task. By designing an invertible network, our approach directly produces the RAE from the original image in an end-to-end manner, without introducing any additional information. Extensive experiments demonstrate the superiority and soundness of our proposed one-stage generation paradigm.

B. Reverse Process of RevINN

Fig. 7 illustrates the recovery process of RevINN. Since both proposed modules are bijective mappings, the recovered image can be obtained by reversing the information flow of the forward process.

Specifically, the RAE is first decomposed by the wavelet transform into three high-frequency components, ${x_{LH}}^2$, ${x_{HL}}^2$, and ${x_{HH}}^2$. Then, the HFPE module performs information restoration according to Eq. 6 to obtain the preliminary recovered high-frequency component ${x_{HC}}^1$. Subsequently, the CFMA module further applies the similar restoration process described in Eq. 4 to combine this component with ${x_{LL}}^1$, producing frequency components that closely match the original ones. Finally, the recovered image $x_{recover}$ is reconstructed through the inverse wavelet transform.

C. Loss Function of Targeted Attack

For the targeted attack, given a targeted label y_{tgt} of an image, the objective for generating RAE is reformulated as:

$$\begin{cases} C(x_{RAE}) = y_{tgt}, \\ C(x_{recover}) = y. \end{cases} \tag{12}$$

Accordingly, in our RevINN we modify the adversarial loss to implement the targeted attack:

$$\mathcal{L}_{adv} = \ell_{CE}(C(x_{\text{recover}}), y) + \ell_{CE}(C(x_{RAE}), y_{tgt}) \tag{13}$$

In addition to steering the generated RAE toward the semantics of the target class, we must also ensure that the recovered image preserves the semantics of the original input.

For the RAE-RDH, RIT, and RAE-YUV methods used for comparison, all of them first generate adversarial examples using I-FGSM. Under the targeted attack setting, their attack loss can be formulated as:

$$\mathcal{L}_{I-FGSM}(x, y_{tgt}) = \ell_{CE}\big(C(x_{AE}), y_{tgt}\big), \tag{14}$$

where x_{AE} denotes the adversarial example. Minimizing $\mathcal{L}_{I-FGSM}$ drives the adversarial example toward the target class. Besides, the SRAE and W-RAE methods also use an adversarial loss similar to Eq. 12 to optimize their attacks.

D. Attack Performance on Transformer-based Models

To verify the universality of our method across different transformer architectures, we compare RevINN with the RAE-YUV scheme on several representative transformer-based models, including ViT-B, Swin-T, CaiT-S, and Visformer-S. The results are summarized in Tab. 5. It can be observed that RevINN can effectively maintain strong attack capability while producing high-quality RAEs, demonstrating that the architectural generality of the proposed approach.

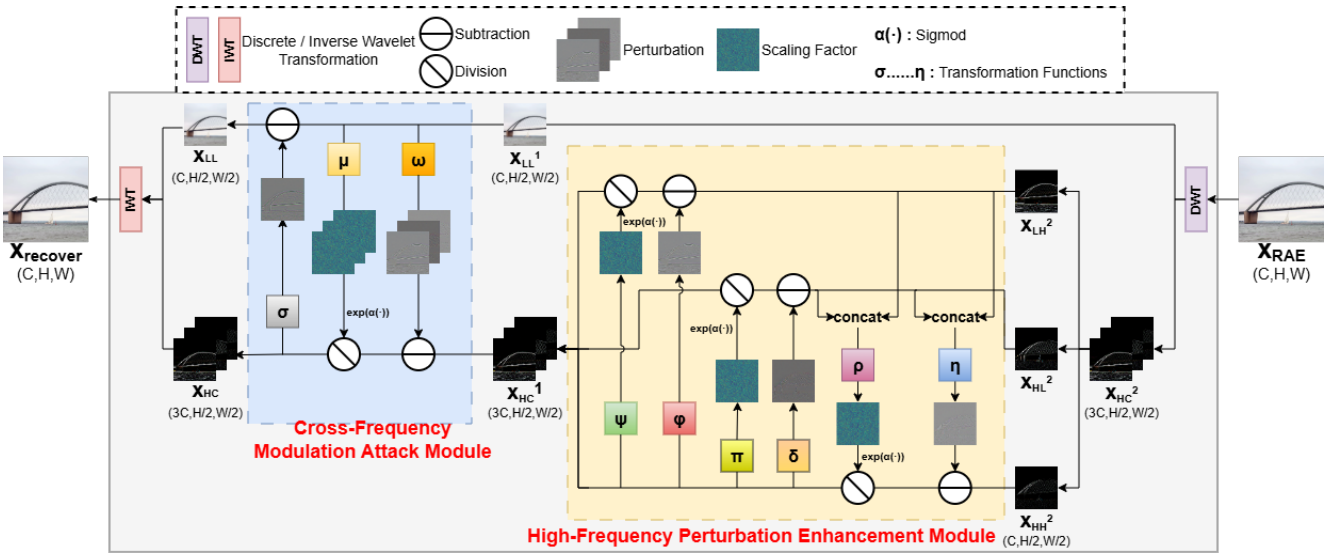

Figure 7. Reverse process of our RevINN for restoring images from reversible adversarial examples.

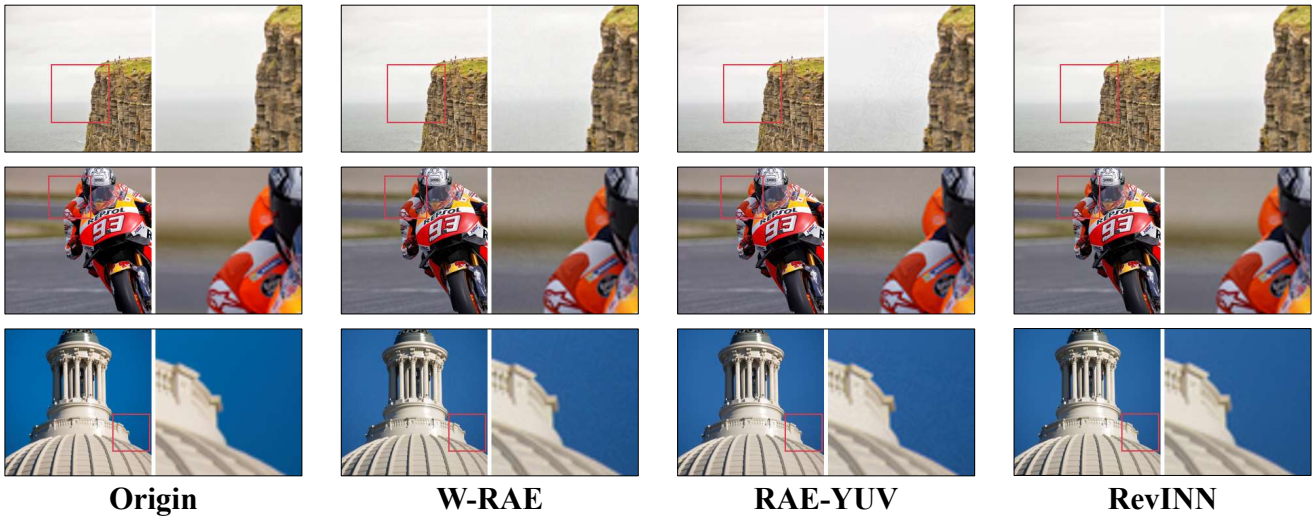

Figure 8. Visual comparison of reversible adversarial examples obtained by SRAE, W-RAE and our RevINN.

Model	Method	ASR	PNSR	SSIM	LPIPS
ViT-B	RAE-YUV [42]	86.4	41.04	0.976	0.070
	Ours	**86.7**	**51.38**	**0.999**	**0.005**
Swin-T	RAE-YUV [42]	86.5	42.09	0.980	0.057
	Ours	**99.0**	**50.62**	**0.998**	**0.007**
CaiT-S	RAE-YUV [42]	**84.2**	41.50	0.978	0.071
	Ours	82.9	**51.43**	**0.998**	**0.007**
Visformer-S	RAE-YUV [42]	84.8	41.84	0.979	0.083
	Ours	**96.3**	**50.08**	**0.997**	**0.013**

Table 5. Attack performance of RevINN on vision transformers.

E. More Visual Qualitative Comparisons

To better illustrate the excellent visual quality of RAEs generated by RevINN, we qualitatively compare them with RAEs produced by the suboptimal W-RAE and RAE-YUV methods, and present enlarged local details, as shown in Fig. 8.

It can be observed that RAEs generated by the proposed RevINN exhibit higher visual imperceptibility compared to other methods. This advantage likely stems from RevINN's use of invertible neural networks to perform information exchange in the frequency domain.In contrast, other methods that generate RAEs based on the image domain usually restrict the perturbation strength by the norm ℓ_p, leading to

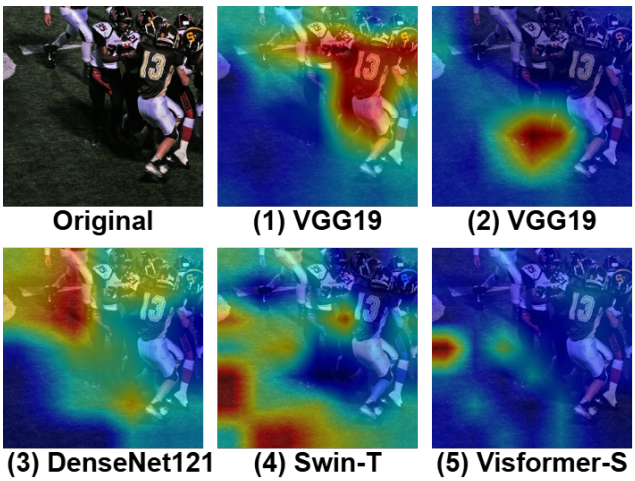

Figure 9. Visualization of Grad-CAMs for different models. (1): the result for the raw image on VGG19. (2)∼(5): the results for RAEs generated by RevINN on different models.

uniformly distributed adversarial perturbations throughout the image.

F. Model Attention Analysis

To shed light on how our method works, we visualize the Grad-CAMs by different models in Fig.9. For the raw image, the model attends to the main object (e.g., the player's torso). In contrast, RevINN consistently redirects different models to focus on erroneous features that are unrelated to the actual content of the image, leading to misclassification.