2025 IEEE 13th International Conference on Computer Science and Network Technology

# Visual Quality Enhancement of Adversarial Examples based on Adaptive Visible Watermarking

**Qiqi Shen[1] , Jingtian Wang[1], Xiaolong Li[1], Jian Li[2], Bin Ma[2], Yao Zhao[1]**

[1] Institute of Information Science,Beijing Jiaotong University, Beijing, China

[2] School of Cyber Security, Qilu University of Technology,Jinan, China

## Abstract

Recently, adversarial examples have received extensive attention due to their excellent performance in revealing the model vulnerability. However, for most existing attacks, poor visual quality is usually introduced to the generated adversarial examples. Then, a novel adversarial example generation method based on adaptive visible watermarking is proposed in this work. A model-based attention map is utilized for visible watermark embedding, in which the watermark is just embedded into the visually insignificant image regions. Moreover, a gradient-based method is utilized to further improve the attack performance, in which the adversarial perturbation is concealed by the previously embedded visible watermark. Extensive experiments on ImageNet-compatible dataset show that, compared with some state-of-the-art works, the proposed method can achieve better attack while ensuring the image quality.

## Motivation

- In recent years, the rapid development of adversarial samples has made deep neural networks very vulnerable. As a result, risks exist in all application fields of deep neural networks. Research on adversarial sample technology is conducive to improving the robustness of deep networks.
- The existing methods for generating adversarial samples ignore the visual quality of adversarial samples in order to increase the success rate of attacks. As shown in the figure 1, the visual quality of adversarial samples is poor.

(a) Original (b) MI-FGSM [9] (c) MISPSO [12]

Figure 1 Adversarial images generated by MI-FGSM , MISPSO .

## Methodology

To ensure the success rate of adversarial sample attacks while improving the visual quality of adversarial samples, when generating adversarial samples, we generate. As shown in Figure 2, the visible watermark is combined with the gradient attack to generate adversarial samples.

Our method integrates two key strategies for effective adversarial watermarking. First, the watermark is embedded into visually insignificant image regions identified by the model's attention map. Moreover, during watermark embedding, the color of the watermark image is adaptively adjusted based on the background contrast. After that, to further improve the attack performance, a gradient-based adaptive perturbation is applied to introduce imperceptible adversarial noise. In this way, the perturbation is mainly introduced into the watermarking regions and it is well concealed by the embedded visible watermark.

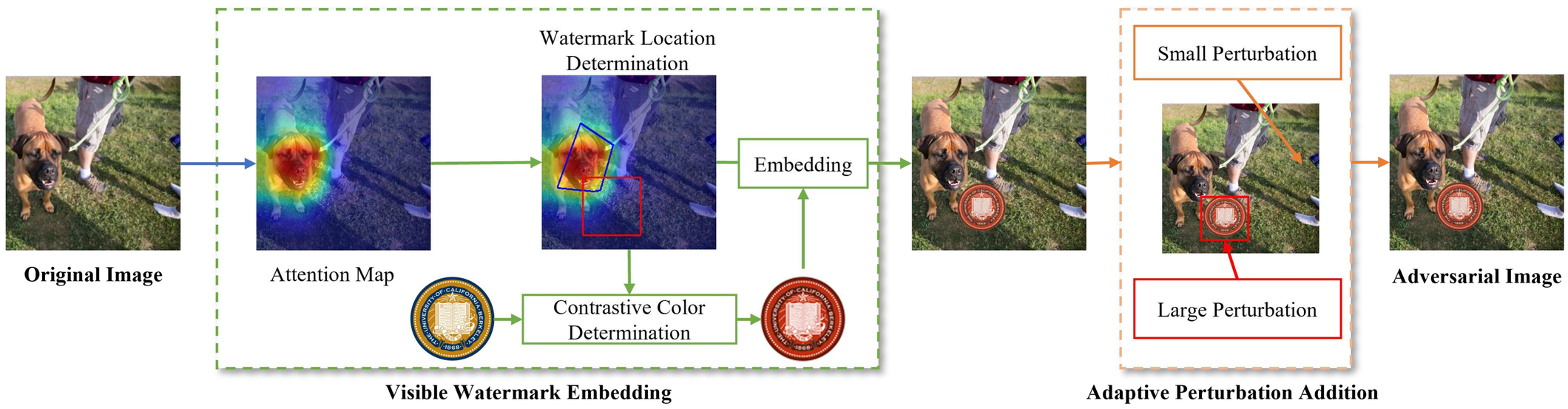


Figure 2 The proposed framework.

## Experiments

(a) Original (b) MI-FGSM [9] (c) NI-FGSM [10] (d) Adv-watermark [11] (e) MISPSO [12] (f) CamoPatch [20] (g) Proposed

Figure 3 Visualization of adversarial images generated by various attacks.

| Surrogate | Attack | Inc-v3 | Inc-v4 | Res-50 | Res-101 | Avg |
|---|---|---|---|---|---|---|
| Inc-v3 | MI-FGSM [9] | **100.0** | 59.2 | 47.4 | 46.3 | 63.2 |
| | NI-FGSM [10] | 96.2 | **62.4** | 50.9 | 52.6 | 65.5 |
| | Adv-watermark [11] | 64.4 | 13.4 | 57.6 | 53.6 | 47.3 |
| | MISPSO [12] | 71.6 | 14.4 | 61.3 | 54.9 | 50.6 |
| | CamoPatch [20] | 92.9 | 36.4 | 21.0 | 20.0 | 42.6 |
| | Proposed | **100** | 45.7 | **76.3** | **69.2** | **72.8** |
| Inc-v4 | MI-FGSM [9] | 57.1 | **100.0** | 46.9 | 44.7 | 62.2 |
| | NI-FGSM [10] | **64.5** | 100.0 | 48.1 | 47.5 | 65.0 |
| | Adv-watermark [11] | 28.7 | 47.4 | 59.6 | 53.7 | 47.4 |
| | MISPSO [12] | 30.0 | 57.4 | 40.5 | 37.6 | 41.4 |
| | CamoPatch [20] | 32.1 | 99.6 | 27.8 | 28.3 | 47.0 |
| | Proposed | 50.3 | **100.0** | **74.9** | **66.6** | **73.0** |
| Res-50 | MI-FGSM [9] | 50.8 | 43.0 | **100.0** | 61.6 | 63.9 |
| | NI-FGSM [10] | **52.3** | **50.4** | **100.0** | 61.8 | **66.1** |
| | Adv-watermark [11] | 16.4 | 8.2 | 72.5 | 34.5 | 32.9 |
| | MISPSO [12] | 13.5 | 10.0 | 78.4 | 29.8 | 32.9 |
| | CamoPatch [20] | 51.5 | 38.1 | 91.7 | 78.0 | 64.8 |
| | Proposed | 35.1 | 24.5 | **100.0** | **95.3** | 63.7 |
| Res-101 | MI-FGSM [9] | **53.6** | **44.8** | 56.9 | 100.0 | 63.8 |
| | NI-FGSM [10] | 53.5 | 44.7 | 56.9 | 100.0 | 63.8 |
| | Adv-watermark [11] | 16.8 | 11.4 | 44.9 | 69.2 | 35.6 |
| | MISPSO [12] | 17.8 | 11.4 | 11.3 | 73.5 | 28.5 |
| | CamoPatch [20] | 46.6 | 23.7 | 26.6 | 92.3 | 47.3 |
| | Proposed | 36.3 | 28.3 | **98.0** | **100.0** | **65.7** |

Table 1 Attack success rates (%) of different adversarial methods against target models. The best results are highlighted in bold..