Supplementary Material: Decoupling Defense Strategies for Robust Image Watermarking

1. Attacked Image Visualization

To better illustrate different attacks' performance, we present various attacked images x_a against AdvMark's final output x_{w_2} in Fig. 1. The bit accuracy can be found in Table 2 in our main paper. Different from watermarked images which are visually indistinguishable from original images x_o, the attacked images already reveal visible artifacts due to strong noises. For example, Gaussian noise, brightness and combined distortion attacks achieve only 25.8, 17.8 and 17.4 PSNR respectively. The regeneration attacks [12] also present noticeable modifications such as the clothes tags "LIONS". This is reasonable since diffusion model [7] tends to perturb image features during the noise injection in the forward process, whose original idea is to exactly render different image styles with the given prompt.

Regarding the adversarial attacks, WEvade [3] maintains high image quality with PSNR 36.8 because its motivation is to evade the target decoder with the minimum perturbation, thanks to the white-box access to the decoder model. While Black-Surrogate attack achieves worse results because the pseudo model itself does not generalize well (refer to the validation accuracy in Fig. 7 (b) in our main experiments). The Black-Query attack always guarantees successful evasion via random image initialization, resulting in significant quality loss. In general, all existing attacks fail to evade AdvMark within acceptable budget, which in turn validates AdvMark's outstanding defense performance.

2. Discussion

We present some clarification for our method design.

1. The advanced regeneration and adversarial attacks are practical in the real world.

On the one hand, regeneration attack maintains high-level semantic features even with low PSNR via diffusion model, as demonstrated by Saberi et al. [8], and diffusion model technique has also been applied to root watermark in Tree-Rings in Wen et al. [10]. On the other hand, adversarial attack evades the decoder with small perturbation that appears natural to human perception, while recent LLaMA model leakage from Meta in 2023 [5] and the open-source decoder of Stable Diffusion both verify the possibility of decoder exposure. Moreover, it is only until recent research in WEvade [3] and Saberi et al. [8] that these two attacks are applied in the watermarking domain, which in turn proves the necessity of a new defense like AdvMark.

2. The proposed two insights are supported by theoretical guarantees and empirical experiments.

Regarding insight 1, we have cited previous work to demonstrate the inevitable tradeoff between robustness and clean accuracy, which has been well studied with theoretical guarantee in the literature such as [6, 11]. In addition, we have conducted experiments in Fig. 2 in the main paper to further prove that MBRS-EAT exhibits lower clean accuracy, despite the higher robustness compared with original MBRS.

Regarding insight 2, we have analyzed the drawback of image optimization on adversarial attack. On the one hand, higher-order derivatives lead to significantly high memory and computation overhead. On the other hand, most watermarking methods include ReLU networks, which are almost locally linear. Therefore the higher-order derivatives only incur slow and difficult convergence, which has also been demonstrated in past works with theoretical guarantee [1, 2].

3. Comparison of AdvMark with previous work: Distortion Agnostic Deep Watermarking.

DADW [4] replaces the typical noise layer with a CNN-based network to generate adversarial examples, which are leveraged to train the encoder and decoder. However, such training paradigm fails to suffice when subjected to more advanced regeneration and adversarial attacks, because the capacity of CNN to simulate various attacks is limited. Specifically, diffusion model reconstructs the image with an intricate U-Net which is way deeper than CNN, and adversarial attack targets the specific decoder instead of the CNN network. Our experiments in Table 2 have further demonstrated the inferior robustness performance of DADW.

On the contrary, AdvMark tackles each attack in accordance with its design mechanism, e.g. encoder-based adversarial training for adversarial attack and direct optimization for regeneration attack. Henceforth AdvMark outperforms DADW in terms of methodology and extensive experiments.

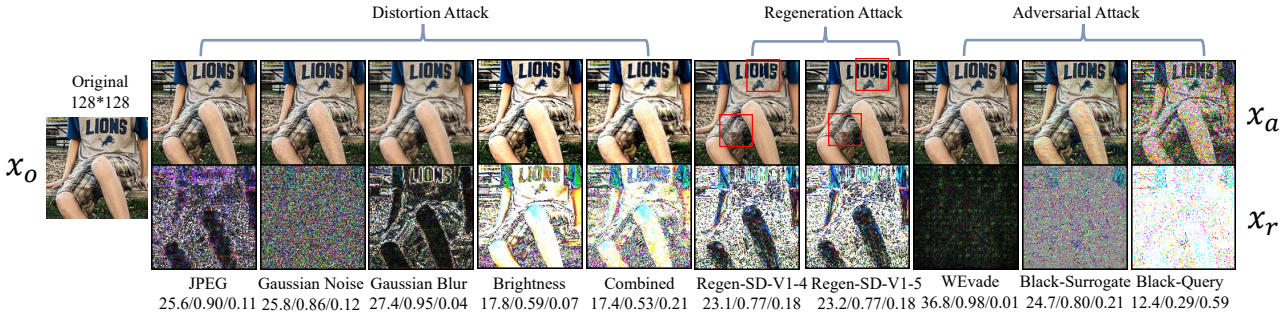

Figure 1. The original x_o, attacked x_a and residual images x_r of different attacks, $x_r = |x_a - x_o| \times 10$. We present below each attack as PSNR↑/SSIM↑/LPIPS↓ (from the attacker's view).

4. According to Table 3, the result of AdvMark ($\lambda_{i_2} = 7$) is better than AdvMark ($\lambda_{i_2} = 5$) in most cases, but we use the latter as default setting.

Higher $\lambda_{i_2} = 7$ gains more robustness and quality at the cost of more optimization iteration overhead, while lower $\lambda_{i_2} = 3$ may lead to incomplete optimization. $\lambda_{i_2} = 5$ is the balanced tradeoff.

5. The performance of AdvMark does not depend on the setting of the PGD perturbation budget ϵ.

As explained in Section 4.3, we have disregarded the typical ϵ-ball projection and leveraged a quality-aware early-stop to guarantee visual quality. In this way, we directly limit the lower bound of PSNR to ensure high image quality.

6. The results in Table 4 show that other baselines like HiDDeN also achieve robustness with stage 2, which demonstrates our effectiveness instead.

As explained in Section 5.4, we tend to demonstrate the effectiveness of stage 2 on enhancing robustness against distortion and regeneration attacks. By incorporating such module with other baselines, they also achieve excellent results (though not better than AdvMark itself), which indeed validates the positive improvement of our proposed stage 2.

Moreover, we also present the results of AdvMark without stage 2 in Table 3, the degradation again demonstrates the effectiveness of such stage.

7. The results in Table 3 present ablation study on AdvMark with different image size and message length n, which ensures fair comparison with baselines with different settings.

As explained in Section 5.4 and Table 1, the image size and n for some baselines are fixed, e.g. $256 \cdot 256$ for StegaStamp [9]. The combination of image size and n in Table 3 has covered all specific baseline in Table 1, whose setting is different from AdvMark's default value, i.e. size is $128 \cdot 128$ and $n = 32$.

3. More Image Examples

To provide comprehensive demonstration, we present 100 watermarked images from AdvMark in Fig. 2, Fig. 3 and Fig. 4. The results have validated the superior image quality without noticeable artifacts.

References

[1] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017. 1

[2] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. pmlr, 2015. 1

[3] Zhengyuan Jiang, Jinghuai Zhang, and Neil Zhenqiang Gong. Evading watermark based detection of ai-generated content. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, pages 1168–1181, 2023. 1

[4] Xiyang Luo, Ruohan Zhan, Huiwen Chang, Feng Yang, and Peyman Milanfar. Distortion agnostic deep watermarking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13548–13557, 2020. 1

[5] Meta. Meta's powerful ai language model has leaked online. https://www.theverge.com/2023/3/8/23629362/meta-ai-language-model-llama-leak-online-misuse., 2023. 1

[6] Tianyu Pang, Xiao Yang, Yinpeng Dong, Hang Su, and Jun Zhu. Bag of tricks for adversarial training. *ICLR*, 2021. 1

[7] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1

[8] Mehrdad Saberi, Vinu Sankar Sadasivan, Keivan Rezaei, Aounon Kumar, Atoosa Chegini, Wenxiao Wang, and Soheil Feizi. Robustness of ai-image detectors: Fundamental limits and practical attacks. *ICLR*, 2024. 1

[9] Matthew Tancik, Ben Mildenhall, and Ren Ng. Stegastamp: Invisible hyperlinks in physical photographs. In *Proceedings*

of the IEEE/CVF conference on computer vision and pattern recognition, pages 2117–2126, 2020. 2

[10] Yuxin Wen, John Kirchenbauer, Jonas Geiping, and Tom Goldstein. Tree-rings watermarks: Invisible fingerprints for diffusion images. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[11] Huan Zhang, Hongge Chen, Chaowei Xiao, Sven Gowal, Robert Stanforth, Bo Li, Duane Boning, and Cho-Jui Hsieh. Towards stable and efficient training of verifiably robust neural networks. *arXiv preprint arXiv:1906.06316*, 2019. 1

[12] Xuandong Zhao, Kexun Zhang, Zihao Su, Saastha Vasan, Ilya Grishchenko, C Kruegel, G Vigna, YX Wang, and L Li. Invisible image watermarks are provably removable using generative ai. *arXiv*, 2023. 1

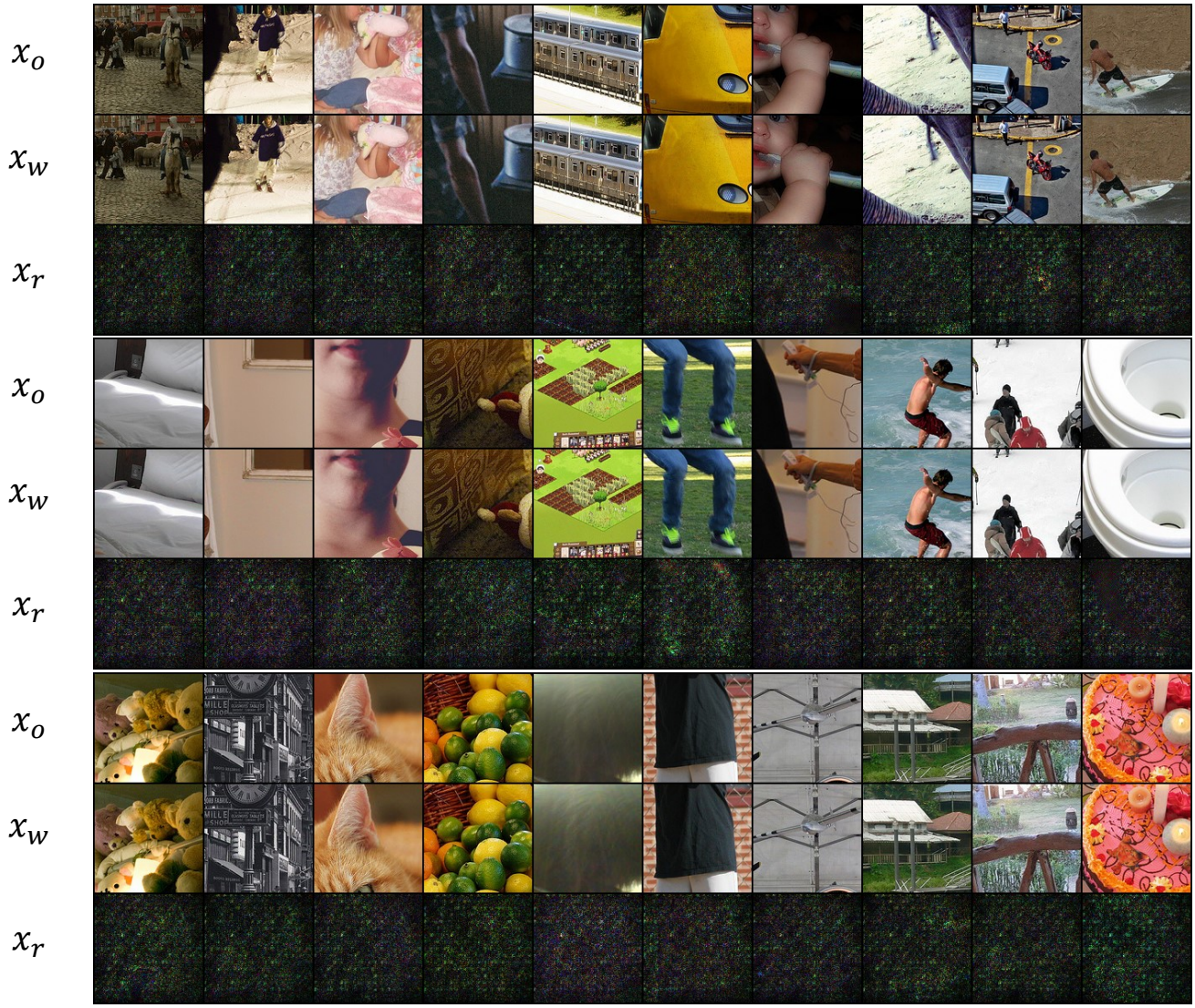

Figure 2. The 1-30 original x_o, watermarked x_w and residual images x_r for AdvMark, $x_r = |x_w - x_o| \times 10$.

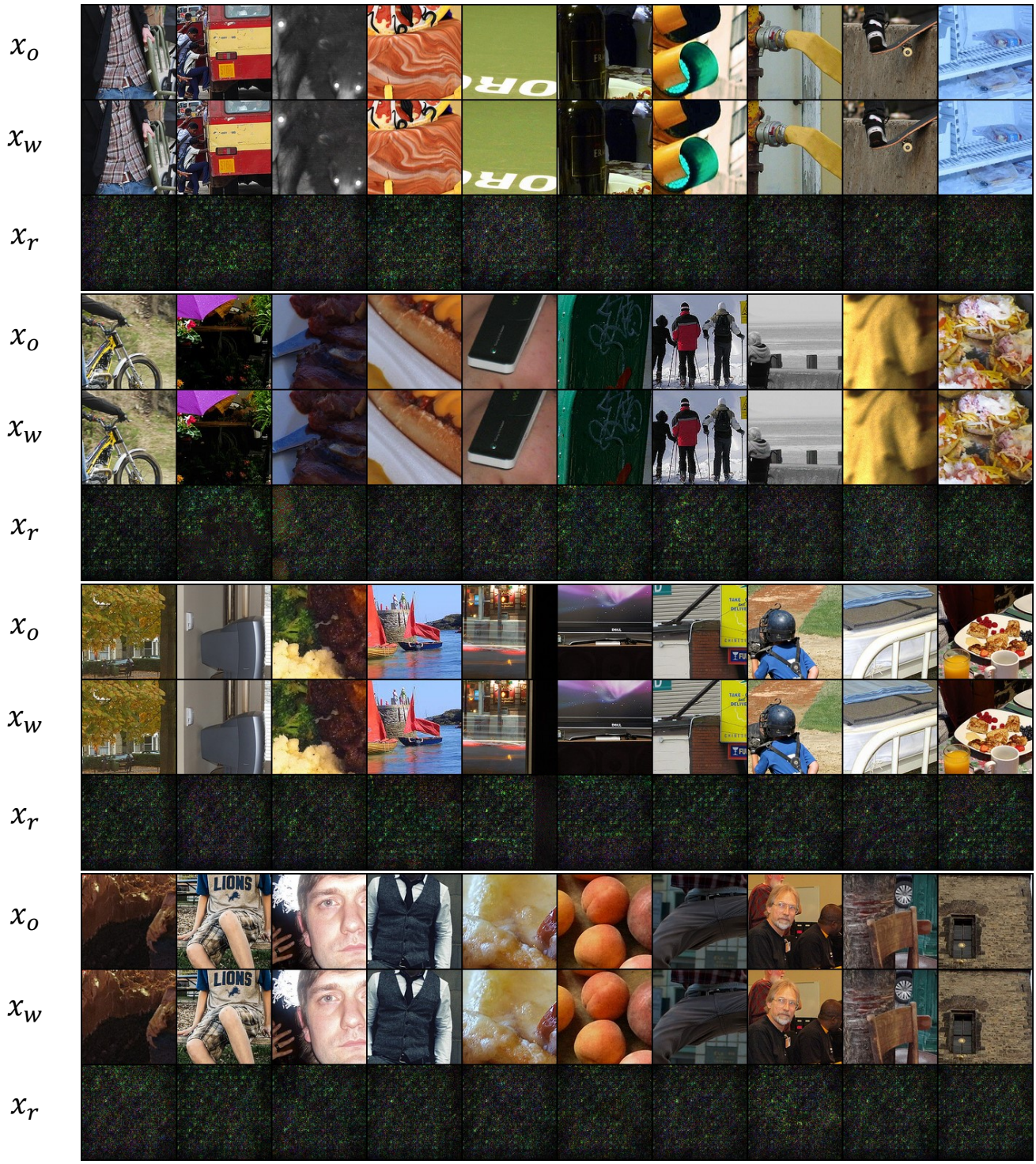

Figure 3. The 31-70 x_o, watermarked x_w and residual images x_r for AdvMark, $x_r = |x_w - x_o| \times 10$.

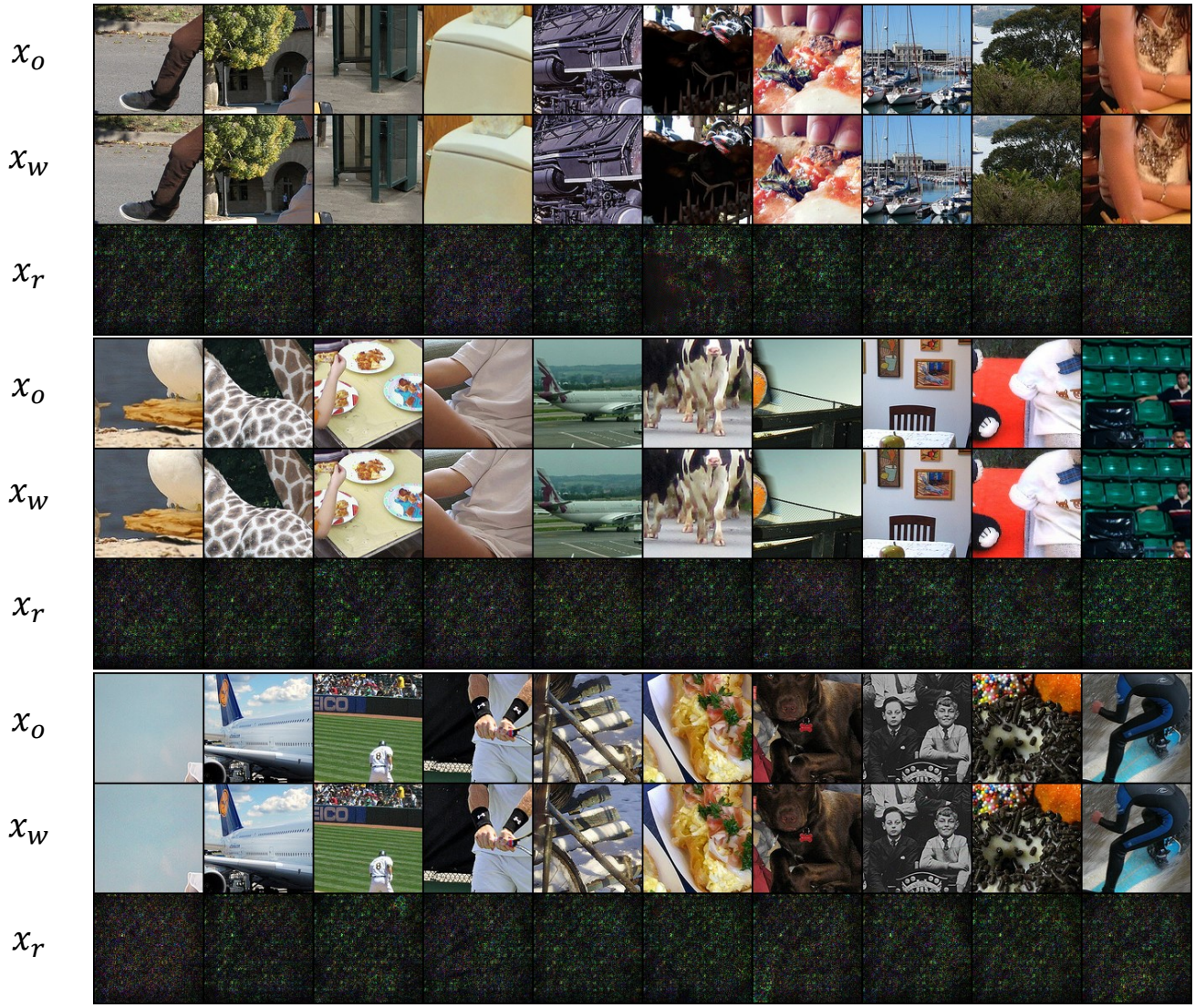

Figure 4. The 71-100 x_o, watermarked x_w and residual images x_r for AdvMark, $x_r = |x_w - x_o| \times 10$.