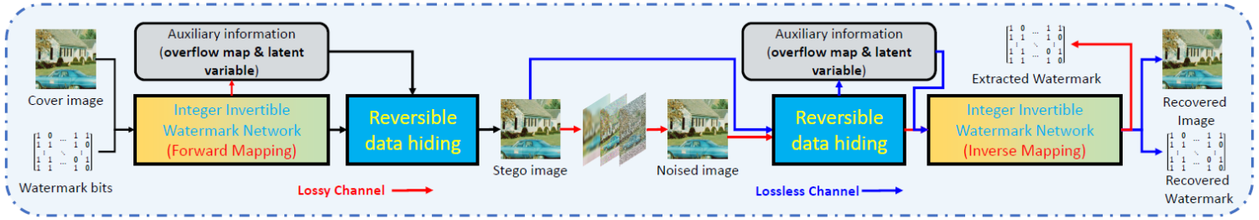
# Learning Robust Image Watermarking with Lossless Cover Recovery

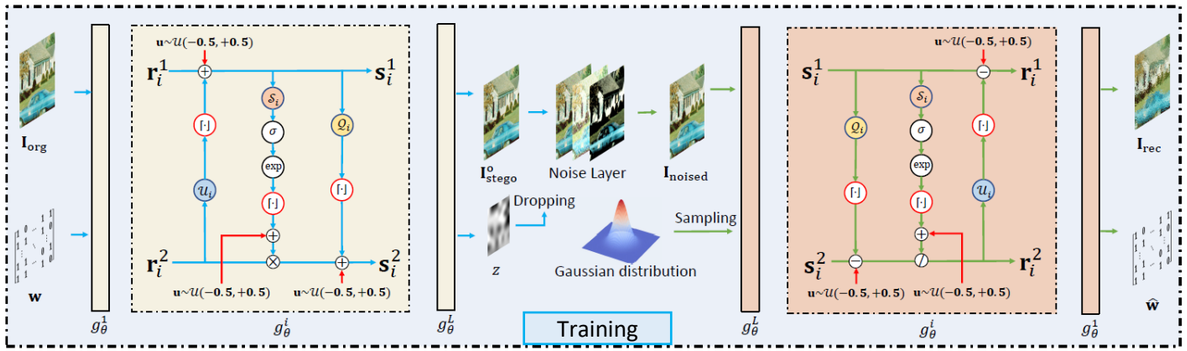
Jiale Chen, Wei Wang*, Chongyang Shi, Li Dong*, Xiping Hu*

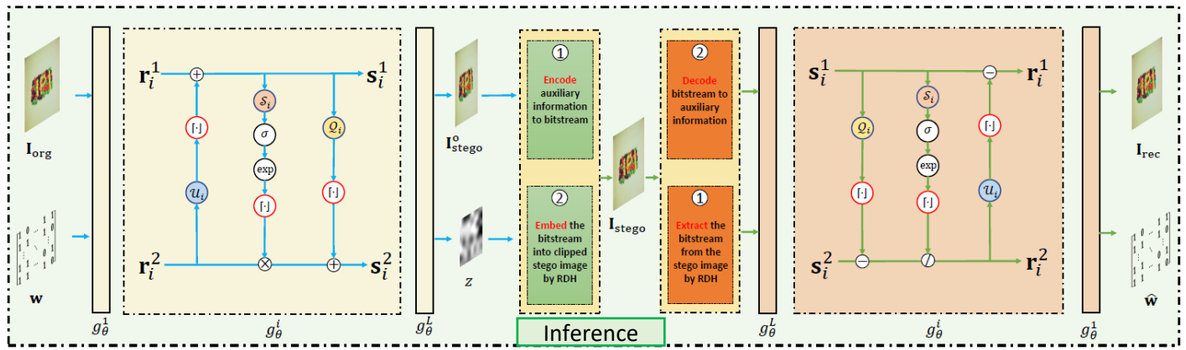
## Does a watermark necessarily cause irreversible damage to an image?

Most existing watermarking methods embed watermarks by adding irremovable perturbations to the cover image, causing permanent distortion. In this work, we propose a Cover-Recoverable WaterMark (CRMark). CRMark can losslessly recover the cover image and watermark in lossless channels and enables robust watermark extraction in lossy channels.

a) CRMark employs an integer Invertible Watermarking Network (iIWN) to achieve a lossless and invertible mapping between the cover image-watermark pair and the stego image.
b) in ***training***, CRMark uses an encoder-noise-layer-decoder architecture to enhance robustness against various distortions.
c) In ***inference***, the cover image and watermark are first mapped to an overflowed stego image and a latent variable, which are then losslessly compressed by arithmetic coding and embedded into the clipped stego image via reversible data hiding.
d) For extraction, in lossy channels, the noisy stego image can be directly inverted through iIWN to recover the watermark; in lossless channels, the latent variable and full stego image are first restored using reversible data hiding, followed by watermark extraction via iIWN.




## Loss functions

**Image loss** $\mathcal{L}_s = \mathrm{MSE}(\mathbf{I}^{\mathrm{o}}_{\mathrm{stego}}, \mathbf{I}_{\mathrm{org}}),\ \mathcal{L}_l = \mathrm{LPIPS}(\mathbf{I}^{\mathrm{o}}_{\mathrm{stego}}, \mathbf{I}_{\mathrm{org}}),$

**regularization loss** $\mathcal{L}_z = \mathrm{MSE}(\mathbf{z}, \mathbf{0}).$

**watermark loss** $\mathcal{L}_{\mathrm{w}} = \mathrm{MSE}(\hat{\mathbf{w}}, \mathbf{w}),$

$\mathbf{I}^{+}_{\mathrm{stego}} = \mathrm{relu}\left(\mathbf{I}^{\mathrm{o}}_{\mathrm{stego}} - 255\right) \quad \mathcal{L}_p = \mathrm{MSE}(\mathbf{I}^{+}_{\mathrm{stego}}, \mathbf{0}) + \mathrm{MSE}(\mathbf{I}^{-}_{\mathrm{stego}}, \mathbf{0}),$
$\mathbf{I}^{-}_{\mathrm{stego}} = \mathrm{relu}\left(\mathbf{0} - \mathbf{I}^{\mathrm{o}}_{\mathrm{stego}}\right)$ **under/overflow penalty loss**

$$\mathcal{L} = \lambda_s \cdot \mathcal{L}_{\mathrm{s}} + \lambda_l \cdot \mathcal{L}_{\mathrm{l}} + \lambda_z \cdot \mathcal{L}_{\mathrm{z}} + \lambda_p \cdot \mathcal{L}_{\mathrm{p}} + \lambda_w \cdot \mathcal{L}_{\mathrm{w}},$$

**Total loss**

a) **Image loss** is used to enhance the imperceptibility of the watermark.
b) **Regularization loss** prevents model parameters from becoming too large, which could cause overflow during inference and prevent perfect recovery of the cover image.
c) **Watermark loss** ensures robust extraction of the watermark under various attacks.
d) **Overflow penalty** penalizes pixel values in the stego image that fall outside the [0,255] range, reducing the length of the auxiliary bitstream for encoding.

## Robustness comparison

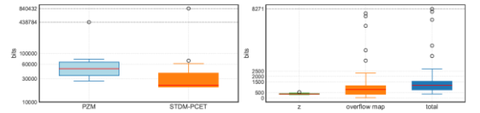
| Methods | JPEG($Q_f$) | | | Gaussian Blur($\sigma$) | | | Gaussian Noise($\sigma$) | | |
|---|---|---|---|---|---|---|---|---|---|
| | 50 | 70 | 90 | 5.0 | 6.0 | 7.0 | 0.05 | 0.15 | 0.25 |
| STDM† [31] | **100.0** | **100.0** | **100.0** | 53.76 | 50.46 | 49.71 | 99.66 | 92.79 | 74.33 |
| PZM† [30] | **100.0** | **100.0** | **100.0** | 61.92 | 54.31 | 50.12 | 99.86 | 88.71 | 62.97 |
| FIN† [11] | 98.64 | 99.14 | 99.40 | 47.98 | 47.58 | 47.73 | 98.88 | 92.28 | 83.42 |
| MBRS† [19] | 97.68 | 99.95 | 100.0 | 87.05 | 74.65 | 62.95 | 99.99 | 97.58 | 90.18 |
| R-CRMark† | 93.33 | 96.06 | 96.84 | 81.18 | 73.05 | 66.64 | 95.30 | 92.30 | 85.00 |
| **CRMark†** | 99.27 | 100.0 | 100.0 | 98.70 | 97.11 | 93.87 | 100.0 | 97.59 | 88.87 |
| **CRMark-R†** | 99.31 | **100.0** | **100.0** | **98.77** | **97.00** | **93.83** | **100.0** | **97.56** | **89.16** |

| Methods | S&P($p$) | | | Dropout($p$) | | | Median Filter($w$) | | |
|---|---|---|---|---|---|---|---|---|---|
| | 0.3 | 0.5 | 0.7 | 0.3 | 0.5 | 0.7 | 11 | 13 | 15 |
| STDM† [31] | 56.46 | 49.81 | 49.48 | 97.98 | 87.24 | 64.01 | 77.18 | 69.82 | 63.76 |
| PZM† [30] | 49.71 | 49.61 | 50.09 | 99.98 | 97.72 | 71.45 | 79.81 | 73.71 | 68.18 |
| FIN† [11] | 72.01 | 62.73 | 57.41 | 98.04 | 94.99 | 88.38 | 53.93 | 50.35 | 49.50 |
| MBRS† [19] | 99.77 | 96.62 | 82.58 | 99.99 | 99.42 | 96.61 | 97.08 | 85.74 | 69.00 |
| R-CRMark† | 95.02 | 87.63 | 74.42 | 93.98 | 84.03 | 69.81 | 94.36 | 90.73 | 85.08 |
| **CRMark†** | **99.86** | 99.06 | 95.53 | 99.98 | 97.75 | 92.35 | 99.41 | 98.83 | 97.52 |
| **CRMark-R†** | 99.85 | **99.28** | **95.88** | **99.99** | **97.68** | **92.40** | **99.39** | **98.77** | **97.61** |

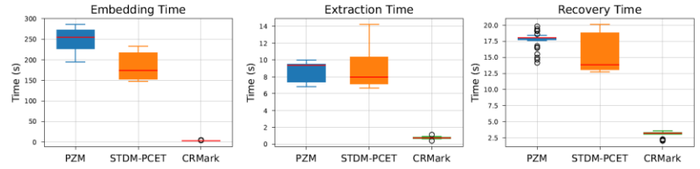
## Comparison of PSNR and SSIM

| Method | SSIM | PSNR |
|---|---|---|
| STDM† [31] | 0.9603 ± 0.0195 | 36.54 ± 1.77 |
| PZM† [30] | 0.9859 ± 0.0078 | 38.17 ± 1.78 |
| FIN† [11] | 0.9988 ± 0.0015 | 39.74 ± 2.49 |
| MBRS† [19] | **0.9996 ± 0.0002** | 40.73 ± 1.67 |
| R-CRMark† | 0.9971 ± 0.0046 | 33.92 ± 5.01 |
| CRMark† | 0.9995 ± 0.0005 | **40.95 ± 1.24** |
| CRMark-R† | 0.9995 ± 0.0005 | 40.90 ± 1.23 |

## Efficiency Comparison

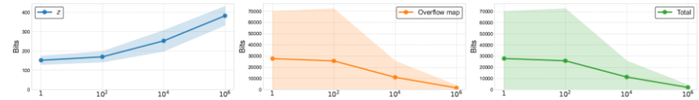



Comparison of auxiliary bitstream length




Comparison of time cost

## Ablation Experiments




Effect of $\lambda_p$ on bitstream length for $\lambda_p \in \{0, 10^2, 10^4, 10^6\}$. Length decreases by $14.18\times$ at $\lambda_p = 10^6$ compared to $\lambda_p = 0$.

**Acknowledgement:** This work was supported by the National Natural Science Foundation of China (62171244), Guangdong Provincial Key Laboratory (Grant 2023B1212060076), and Innovation Team Project of Guangdong Province of China (No. 2024KCXTD017).