# Sim-to-Real: An Unsupervised Noise Layer for Screen-Camera Watermarking Robustness

**Yufeng Wu[1], Xin Liao[1]*, Baowei Wang[2,3], Han Fang[4], Xiaoshuai Wu[1], Mingyue Chen[1], Guiling Wang[5]**

[1]College of Cyber Science and Technology, Hunan University, Changsha 410082, China
[2]the Engineering Research Center of Digital Forensics, Ministry of Education, the School of Computer Science,
Jiangsu Collaborative Innovation Center of Atmospheric Environment and Equipment Technology,
Nanjing University of Information Science and Technology, Nanjing 210044, China
[3]Jiangsu Yuchi Blockchain Technology Research Institute, Nanjing 210000, China
[4]School of Computing, National University of Singapore, Singapore
[5]Department of Computer Science, New Jersey Institute of Technology, Newark, NJ 07102, USA
wuyufeng0523@163.com; xinliao@hnu.edu.cn; wang@nuist.edu.cn; fanghan@nus.edu.sg;
shinewu@hnu.edu.cn; chenmingyue@hnu.edu.cn; gwang@njit.edu

## Abstract

Unauthorized screen capturing and dissemination pose severe security threats such as data leakage and information theft. Several studies propose robust watermarking methods to track the copyright of Screen-Camera (SC) images, facilitating post-hoc certification against infringement. These techniques typically employ heuristic mathematical modeling or supervised neural network fitting as the noise layer, to enhance watermarking robustness against SC. However, both strategies cannot fundamentally achieve an effective approximation of SC noise. Mathematical simulation suffers from biased approximations due to the incomplete decomposition of the noise and the absence of interdependence among the noise components. Supervised networks require paired data to train the noise-fitting model, and it is difficult for the model to learn all the features of the noise. To address the above issues, we propose Simulation-to-Real (S2R). Specifically, an unsupervised noise layer employs unpaired data to learn the discrepancy between the modeled simulated noise distribution and the real-world SC noise distribution, rather than directly learning the mapping from sharp images to real-world images. Learning this transformation from simulation to reality is inherently simpler, as it primarily involves bridging the gap in noise distributions, instead of the complex task of reconstructing fine-grained image details. Extensive experimental results validate the efficacy of the proposed method, demonstrating superior watermark robustness and generalization compared to state-of-the-art methods.

**Code** — https://github.com/ttz0523/S2R-main

## Introduction

With the widespread use of digital images in areas such as digital photography, presentations, social media, and online publications, image piracy has become an increasing concern. Although digital watermarking techniques have proven effective for copyright protection in purely digital environments (Zhu et al. 2018; Jia, Fang, and Zhang 2021; Wang, Wu, and Wang 2023; Fu et al. 2024), they frequently fail when content is presented on a screen and subsequently captured by a camera, which is a common method of unauthorized reproduction (Fang et al. 2018; Tancik, Mildenhall, and Ng 2020). In these Screen-Camera (SC) scenarios, watermarks are subject to complex physical degradations introduced during the display and capture process, which can lead to message loss. Ensuring robustness against SC degradation remains a critical challenge for watermarking systems in real-world copyright protection.

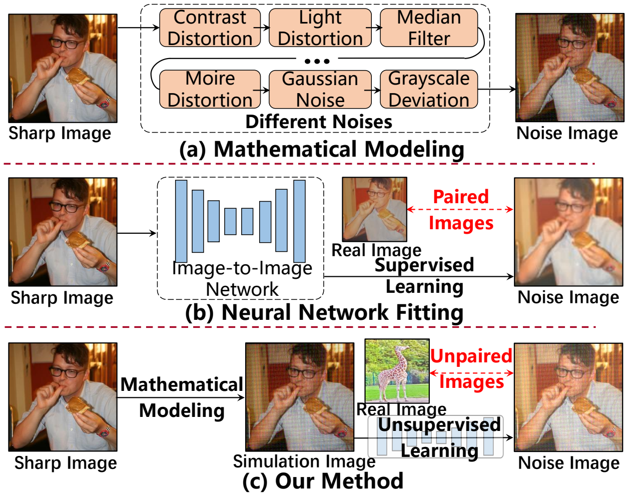


Figure 1: Overview of traditional noise approximation strategies and proposed Simulation-to-Real. (a) Mathematical modeling-based noise approximation. (b) Supervised neural fitting via paired images. (c) Our method: transforming sharp images to a certain noise domain, then mapping them unsupervised to an unknown domain, achieving more realistic noise approximation.

[*]Corresponding author.

During the past few years, a large number of SC resistant watermarking methods have emerged. Adversarial training is an effective tool for resisting SC by approximating the noise introduced during SC. This approach helps to adapt the watermarking network to resist changes in the image caused by SC during training. To bridge the gap between digital and real environments, existing SC resistant watermarking works are categorized into two strategies: mathematical modeling (Tancik, Mildenhall, and Ng 2020; Fang et al. 2022; Li, Liao, and Wu 2024) and neural network fitting (Wengrowski and Dana 2019).

Mathematical modeling allows for a flexible and generalized SC noise approximation through a combination of different mathematical formulas, which explicitly expresses the impact of each noise component, as shown in Figure 1 (a). However, most mathematical modeling methods separate the noise components by linear superposition of independent noises (Tancik, Mildenhall, and Ng 2020; Fang et al. 2022; Li, Liao, and Wu 2024), ignoring the coupling of the noise sources in the real scene, leading to the deviation of the modeling from the real SC noise. Furthermore, mathematical modeling typically focuses on large-scale noise characteristics and struggles to model fine-grained, localized distortions. This is because the formulas and parameterizations employed in mathematical modeling are more suited to describing regular, widespread noise patterns, such as global blur, perspective distortions, or Gaussian noise.

Neural network fitting-based methods are capable of precisely fitting nonlinear SC noise features, as shown in Figure 1 (b). However, deep learning methods for noise approximation are hindered by the high demand for training data and limited noise modeling capacity. Current methods primarily adopt supervised learning, but obtaining diverse and high-quality paired SC samples is frequently a challenge. Constructing paired real samples typically involves manual rectification and alignment with sharp images, a process prone to spatial misalignments. Such inconsistencies introduce labeling bias, rendering data collection both error-prone and labor-intensive. Moreover, unlike mathematical models, directly mapping sharp images to the complex and variable noise patterns associated with SC can overwhelm the model due to the vast diversity of noise types, leading to suboptimal performance. Although the model is capable of generating precision noise, it may not capture detailed noise features, diminishing the precision of the approximation, making the training process more complex and hindering the convergence.

To address the aforementioned issues, in this work, we propose Simulation-to-Real (S2R), a novel framework that leverages mathematical modeling and neural network fitting to formulate the transition from sharp images to real-world SC noise. This work pioneers a neural network-based framework for fitting approximate noise, guided by mathematical modeling. As shown in Figure 1 (c), first, a rough representation of a certain domain is derived from a mathematical model. Then, the rough domain is further refined through unsupervised learning on unpaired data. This allows the model to capture unknown domain noise features that are not considered by the mathematical model. We conduct extensive experiments to compare the effectiveness of our model with other state-of-the-art SC resistant watermarking methods. The results demonstrate that S2R outperforms existing methods, highlighting its superior performance in addressing the challenges of watermark robustness against SC in real-world environments.

The contributions of the proposed method are shown as follows:

- We propose Simulation-to-Real (S2R), the first noise approximation framework guided by mathematical modeling. It maps simulated noise to real screen-camera noise and narrows the gap between synthetic and real data. The approach is supported by a feasibility proof.
- We innovatively introduce an unsupervised method without the requirement of paired data to bridge the distributional discrepancies between noise in simulation and real, resulting in improved accuracy in noise approximation.
- We build a scalable structure based on mathematical modeling and unsupervised learning. The mathematical modeling module supports replacement and flexible modification of algorithms according to requirements.
- Extensive experiments show that our method surpasses state-of-the-art methods in watermark robustness and image quality under real screen-camera conditions.

# Related Work

## Screen-Camera Resistant Watermarking

Recent SC resistent watermarking methods increasingly adopt deep learning frameworks (Zhu et al. 2018; Wang, Wu, and Wang 2023) due to their superior robustness and imperceptibility compared to traditional techniques (Fang et al. 2020; He et al. 2024; Wang et al. 2024; Zhu et al. 2024). These methods typically use an encoder–decoder structure, with a noise layer inserted between them to simulate distortions during training. The noise layer designed specifically to simulate SC noise can be constructed either through mathematical modeling or neural network fitting.

Mathematical modeling-based strategy is widely adopted in watermarking frameworks targeting SC distortions, due to its differentiability and controllable approximation capability. StegaStamp (Tancik, Mildenhall, and Ng 2020) proposed a differentiable pipeline to simulate physical distortions such as perspective distortion, blur, color shifts, noise, and JPEG compression. Unlike earlier single-noise strategies, it applies all distortions sequentially to improve robustness. Luo et al. (Luo et al. 2020) added lighting variations to enhance realism. PIMoG (Fang et al. 2022) modeled SC noise as a mixture of perspective transform, lighting, moiré, and Gaussian noise. SSDS (Li, Liao, and Wu 2024) further introduced grayscale deviation. Although effective, these models still struggle to capture complex artifacts in real SC scenarios.

Neural network fitting-based strategy learn distortion mappings directly from data. (Wengrowski and Dana 2019) trained a distortion network, CDTF, with a 1.9 TB real dataset, allowing the network to learn the features of SC distortion. Similarly, CDTF has also been successfully applied in challenging Print-Camera scenarios (Qin et al. 2023).

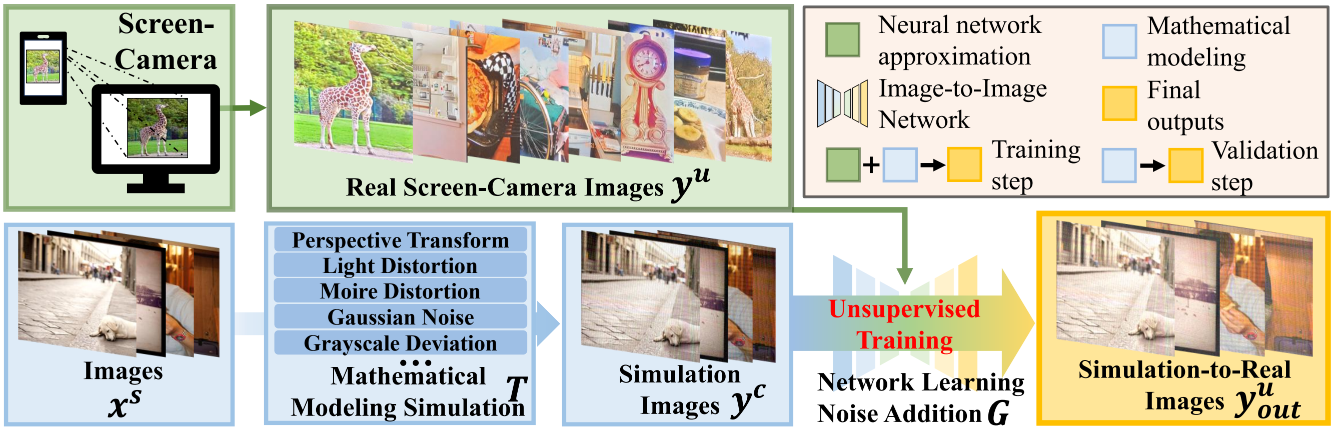


Figure 2: Overview of the proposed S2R. In the training phase, given a set of sharp images $x^s$ and real SC images $y^u$, the sharp images are first transformed into images with a certain simulation noise distribution $y^c$ using a pre-defined mathematical modeling transformation $T$. Through unsupervised training, the Image-to-Image Network $G$ gradually adjusts $y^c$ to match the distribution of $y^u$, ultimately outputting the approximate images $y^u_{\text{out}}$. In the validation phase, given sharp images $x^s$, after passing through the transformation $T$ and the fixed-weight network $G$, the outputs are $y^u_{\text{out}}$.

However, these supervised methods rely heavily on paired data and frequently generalize poorly to unseen devices, making them less suitable in real-world settings. Thus, the neural network fitting-based strategy requires further exploration.

### Unsupervised Learning with Unpaired Data

To overcome the dependence on paired data, unsupervised learning methods based on Generative Adversarial Networks (GANs) (Goodfellow et al. 2014) have been extensively studied. In particular, CycleGAN (Zhu et al. 2017) and DualGAN (Yi et al. 2017) introduce cycle consistency loss to enable unpaired image translation, facilitating learning from unpaired data. While Pix2Pix (Isola et al. 2017) represents a supervised image-to-image translation framework relying on paired data and adversarial loss with a PatchGAN discriminator, its core principles have been extended to unsupervised settings (Pham et al. 2024). This has led to progress in deblurring (Pham et al. 2024), deraining (Chang et al. 2023), and denoising (Pang et al. 2021).

However, these frameworks struggle with bridging the gap between sharp and noisy domains due to the complexity of real-world noise. To address this limitation, our S2R builds upon employing unsupervised techniques to learn and approximate the differences between simulated noise distributions and real SC noise. This approach overcomes the challenge of direct mapping between simulated and real noise complexities.

## Proposed Method

### Motivation

Our goal is to learn a noise approximation function $F_{\mathcal{U}}(\cdot)$ that transforms sharp images $x^s \in \mathcal{S}$ into SC images $y^u \in \mathcal{U}$, i.e., $F_{\mathcal{U}}(x^s) = y^u$.

One strategy for approximating SC noise is to decompose it into independent components via mathematical modeling, thereby enabling targeted simulation. However, this strategy has two significant limitations. First, it tends to introduce superfluous or overly pronounced distortions, which compromises imperceptibility. Second, by assuming component independence, it disregards the inherent interdependencies among different noise types, leading to unrealistic approximations of real-world conditions.

An alternative approach is to employ supervised image translation networks to directly learn the mapping $F_{\mathcal{U}}(\cdot)$ from paired data. However, these methods face two main limitations. First, obtaining reliable supervised data is inherently challenging. SC images require manual rectification and cropping to align with their original counterparts, a process that often introduces spatial misalignments. These misalignments lead to significant labeling bias, making the collection of truly accurate data pairs both costly and labor-intensive. Second, even with ideal data, neural networks with limited capacity struggle to model the highly complex and diverse nature of SC noise. The underlying noise distribution often comprises a mixture of intricate, overlapping patterns that exceed the representational ability of a single network. As a result, the network fails to capture the full diversity and complexity of real-world noise variations.

### Method Overview

Based on the above discussion, when neither mathematical modeling nor neural network fitting is ideal due to their inherent limitations, a hybrid strategy that leverages the strengths of both becomes a promising alternative for approximating SC noise. We introduce an innovative method of learning $F_{\mathcal{U}}(\cdot)$, rather than directly learning this function, which is very challenging. We treat $F_{\mathcal{U}}(\cdot)$ as a composition of the mathematical modeling transformation $T$ from the certain domain $\mathcal{C}$ and the network mapping function $G$ from the certain domain to the unknown domain $\mathcal{U}$:

$$F_{\mathcal{U}}(\cdot) = T * G. \tag{1}$$

We leverage the prior noise constraints provided by mathematical modeling to offer a foundation of noise robustness and generalization for deep learning, and then utilize deep learning to bridge the gap between the certain modeling noise distribution and the unknown noise distribution. The challenge here is that we cannot obtain paired noise datasets for supervised training, as acquiring paired datasets of simulated images and real SC images is extremely difficult. Therefore, the only feasible choice is to adopt unpaired data. Fortunately, we can capture a set of real SC images $U$ arbitrarily, without being restricted by the set of sharp images $S$. These two sets of images are unpaired, which means that there is no need for a one-to-one correspondence between images from $\mathcal{S}$ and $\mathcal{U}$. Consequently, collecting these datasets is relatively easy and straightforward.

Our goal is then shifted to learning $G$ to bridge the gap between domain $\mathcal{C}$ and domain $\mathcal{U}$. In particular, our task is to learn a mapping function $G$ that maps each input image $y^c$ simulated by mathematical modeling, defined in Eq. (2), to an image $y^u$ that has the same sharp visual representation $x^s$ but belongs to the unknown noise distribution $\mathcal{U}$.

$$T : x^s \to y^c, G : y^c \to y^u, \quad \text{where} \quad y^u = G(T(x^s)). \tag{2}$$

The general method is illustrated in Figure 2. Our method decomposes a complex task into two more manageable tasks. One task is to simulate noise through mathematical modeling, which, although challenging, benefits from existing research. We can select a well-performing noise model $T$ that has already achieved good simulation results in its domain $\mathcal{C}$. The other task is to learn the transformation from a certain noise domain $\mathcal{C}$ to the unknown domain $\mathcal{U}$. The difficulty of this task depends on the difference between $\mathcal{C}$ and $\mathcal{U}$, but it is much easier than directly learning the mapping from $\mathcal{S}$ to $\mathcal{U}$. Moreover, we can flexibly select the most appropriate $T$ and $G$ for our specific SC noise domain, ensuring that the simulation accuracy of SC noise is maximized.

**Design of S2R.** To implement S2R framework, the key is to train a noise-to-noise transformation network $G$ that can convert any noisy image from a certain noise domain $\mathcal{C}$ to an unknown noise domain $\mathcal{U}$, while preserving the content of the image. To train $G$, we need two datasets: sharp images from $\mathcal{S}$, and real SC images from an unknown source $\mathcal{U}$. We design $G$ to operate on multiple scales and carefully design the training loss to achieve the desired results. We adopt MIMO-UNet (Fan et al. 2021) for its efficiency and simplicity. For a more detailed description of the framework, please refer to Section A.3 of the supplementary material.

## Loss Function

**Adversarial Loss** We employ adversarial loss to constrain the generative network $G$ to produce images with the characteristics of target noise. To achieve this, we introduce a discriminator network $D$ to distinguish between real SC noisy images and generated images. Networks $G$ and $D$ are alternately trained within a minimax game framework. The adversarial loss is defined as follows:

$$\begin{aligned} L_{\text{cGAN}}(G, D) = &\ \mathbb{E}_{y\sim\mathcal{U}} [\log D(y)] \\ &+ \mathbb{E}_{y\sim\mathcal{C}} [\log (1 - D(G(y)))] . \end{aligned} \tag{3}$$

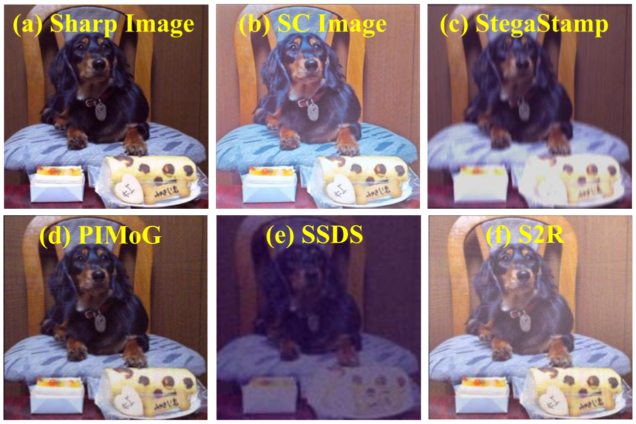


Figure 3: Approximation results of different methods converting sharp images into noise images: (a) sharp images; (b) real SC images; (c) StegaStamp (Tancik, Mildenhall, and Ng 2020); (d) PIMoG (Fang et al. 2022); (e) SSDS (Li, Liao, and Wu 2024); (f) our S2R.

We train $G$ to minimize the above loss term, while training the discriminator $D$ to maximize it. Additionally, we apply the regularization of gradient penalty on the discriminator to enforce the Lipschitz continuity constraint (Gulrajani et al. 2017). The gradient penalty loss is defined as:

$$L^D_{\text{grad}}(D) = \mathbb{E}_{\tilde{y}\sim\tilde{C}} \left[ \left( \|\nabla_{\tilde{y}} D(\tilde{y})\|_2 - 1 \right)^2 \right], \tag{4}$$

where $\tilde{C}$ represents the set of samples $\tilde{y}$ obtained through random interpolation between an image $y \in C$ and the generated image $G(y)$.

**Reconstruction Loss** Previous methods (Zhu et al. 2017) have found it beneficial to combine the GAN objective with more traditional losses (such as the $L_1$, $L_2$ distance, etc.). The discriminator's task remains unchanged, but the generator's objective is to both deceive the discriminator and generate outputs consistent with the real noise. We adopt a weighted multi-scale perceptual loss which reconstructs image content from coarse to fine without being overly constrained by pixel-level accuracy:

$$L_P(G) = \frac{1}{2^{k-1}} \sum_{i=1}^{k} \frac{1}{t_i} \mathbb{E}_{y\sim\mathcal{U}} \left\| \varphi(y_i^c) - \varphi(G(y_i^c)) \right\|, \tag{5}$$

where $\varphi$ is a feature extractor from a pretrained network.

**Total Loss** By combining $L_{\text{cGAN}}$ with $L_P$, the generator loss $L_G$ and the discriminator loss $L_D$ are given as follows:

$$L_G = L_{\text{cGAN}}(G, D) + \lambda_G L_P(G), \tag{6}$$

$$L_D = -L_{\text{cGAN}}(G, D) + \lambda_{\text{grad}} L^D_{\text{grad}}(D), \tag{7}$$

where $\lambda_G$ and $\lambda_{\text{grad}}$ are the weight factors.

## Feasibility Proof of Inter-domain Migration for Noise

In this work, to address the difficulty of approximating SC noise, our method transforms from $\mathcal{S}$ to $\mathcal{C}$, and then estimates the unknown real SC noise domain $\mathcal{U}$, i.e., $\mathcal{S} \to \mathcal{C} \to$

$\mathcal{U}$. We focus on the noise in the SC input, treating the added noise as multiplicative and additive noise (Lim 1990). The noise image $y^u$ is denoted as a function of the corresponding sharp image $x^s$ through the noise operators $k^u$ and $n^u$, with these operators being associated with the noise domain $\mathcal{U}$ that corresponds to the SC noise:

$$y^u = k^u \cdot x^s + n^u, \tag{8}$$

where $k^u$ is the multiplicative noise term, and $n^u$ is the additive noise term. Then, we decompose the new unknown noise formula into the following operators:

(a) Multiplicative noise term:

$$k^u = k^{(c \to u)} \cdot k^{(s \to c)}. \tag{9}$$

(b) Additive noise term:

$$\begin{aligned} n^u &= k^{(c \to u)} \cdot n^c + n^{(c \to u)} \\ &= k^{(c \to u)} \cdot \Big( k^{(s \to c)} \cdot n^s + n^{(s \to c)} \Big) + n^{(c \to u)}, \end{aligned} \tag{10}$$

where $k^{(c \to u)}$ denotes the degree to which $k^u$ deviates from any noise $k^c$ sampled from domain $\mathcal{C}$ (multiplicative noise operator), and $n^{(c \to u)}$ denotes the degree to which $n^u$ deviates from any noise $n^c$ sampled from domain $\mathcal{C}$ (additive noise operator). Similarly, other operators can be derived. Given the above formulas, we observe that the multiplicative noise $k^u$ can be decomposed as $k^{(c \to u)} \cdot k^{(s \to c)}$, and the additive noise can be decomposed as:

$$n^u = k^{(c \to u)} \cdot \Big( k^{(s \to c)} \cdot n^s + n^{(s \to c)} \Big) + n^{(c \to u)}. \tag{11}$$

Therefore, we can express the noise image $y^u$ as follows:

$$\begin{aligned} y^u &= k^u \cdot x^s + n^u \\ &= k^{(c \to u)} \cdot k^{(s \to c)} \cdot x^s \\ &\quad + k^{(c \to u)} \cdot \Big( k^{(s \to c)} \cdot n^s + n^{(s \to c)} \Big) \\ &\quad + n^{(c \to u)}. \end{aligned} \tag{12}$$

Further expansion:

$$\begin{aligned} y^u = {} & k^{(c \to u)} \cdot k^{(s \to c)} \cdot x^s + k^{(c \to u)} \cdot k^{(s \to c)} \cdot n^s \\ & + k^{(c \to u)} \cdot n^{(s \to c)} + n^{(c \to u)}. \end{aligned} \tag{13}$$

Factor out the common term $k^{(c \to u)}$:

$$\begin{aligned} y^u = {} & k^{(c \to u)} \cdot \Big( k^{(s \to c)} \cdot x^s + n^{(s \to c)} \Big) \\ & + n^{(c \to u)} + k^{(c \to u)} \cdot k^{(s \to c)} \cdot n^s. \end{aligned} \tag{14}$$

Since $y^c = k^{(s \to c)} \cdot x^s + n^{(s \to c)}$, substituting this gives:

$$y^u = k^{(c \to u)} \cdot y^c + n^{(c \to u)} + k^{(c \to u)} \cdot k^{(s \to c)} \cdot n^s. \tag{15}$$

In our method, since all the sharp images are assumed to be noise-free, we can set $n^s = 0$. The equation simplifies to:

$$y^u = k^{(c \to u)} \cdot y^c + n^{(c \to u)}. \tag{16}$$

We convert $k^{(c \to u)}$ and $n^{(c \to u)}$ into the corresponding neural network mappings $k_\delta$ and $n_\delta$, respectively. Ultimately, we can conclude:

$$y^u = k_\delta \cdot y^c + n_\delta. \tag{17}$$

Through the above derivation, the distribution alignment task is transformed from directly learning the sharp image $x^s$ to $y^u$ into learning the bias between the certain noise image $y^c$ and the unknown noise image $y^u$. This significantly reduces the difficulty of the unsupervised distribution alignment task and facilitates the network's ability to correctly focus on learning the differences between domain $\mathcal{C}$ and domain $\mathcal{U}$. Therefore, we are able to utilize neural networks to accomplish the conversion from the certain noise domain to the unknown noise domain to obtain an accurate simulation of the real SC noise.

# Experimentation

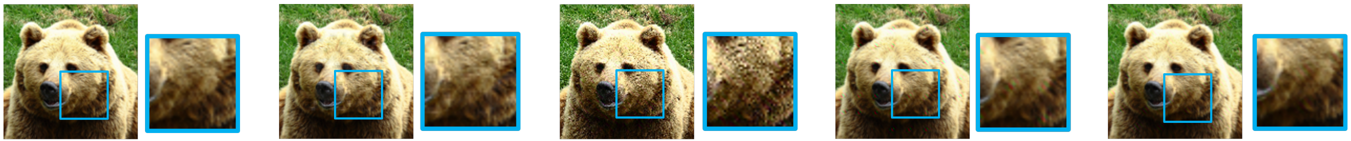
## Implementation Details

**Watermarking and S2R Framework** We adopt MCFN (Wu et al. 2024) as the default watermarking framework and employ the COCO dataset (Lin et al. 2014) for training. Following previous works (Jia, Fang, and Zhang 2021; Fang et al. 2022; Wang, Wu, and Wang 2023), 10,000 images are selected for training, each resized to $128 \times 128$ and embedded with a random 64-bit binary watermark. To simulate SC degradation, we propose the S2R framework, which consists of a simulated noise layer $T$ and a transformation network $G$. We use PIMoG (Fang et al. 2022) as $T$, and an improved MIMO-UNet (Fan et al. 2021) as $G$ with the default configuration in unsupervised Pix2Pix implementation (Isola et al. 2017; Pham et al. 2024). To build the training dataset for S2R, we randomly select COCO images and capture their SC versions by three device pairs: Samsung Galaxy S20 FE with Lenovo Legion Y9000P (S+L), iPhone 13 with Envision G249G (I+E), and MEIZU 20 Pro with ASUS ROG Strix SCAR Edition 8 (M+A). Each pair contributes 900 SC images, forming a combined dataset referred to as SIM+LEA. All training is conducted on an NVIDIA RTX 4090 GPU.

For default testing, we train S2R on the SIM+LEA dataset. We then capture 100 encoded images from the COCO dataset, which was not used during training, with the S+L pair for evaluation. Viewpoint angles follow a left-to-right axis, where negative and positive values indicate bottom/left and top/right tilts, respectively. Comparative methods are tested under the same conditions to ensure fairness. More details are provided in Section A of the supplementary material.

**Metrics** We evaluate watermarking performance in terms of invisibility and robustness. Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) measure visual quality between the watermarked and original images, where higher values indicate less distortion. Robustness is assessed via the Bit Error Rate (BER), with lower values indicating more accurate extraction.

**Baseline** To ensure fair comparison, we benchmark S2R against StegaStamp (Tancik, Mildenhall, and Ng 2020), PIMoG (Fang et al. 2022), and SSDS (Li, Liao, and Wu 2024). Since SSDS and CDTF (Wengrowski and Dana 2019) lack open-source implementations, we reproduce SSDS and replace CDTF with a supervised variant of S2R.



| (a) Sharp Images | (b) StegaStamp | (c) PIMoG | (d) SSDS | (e) S2R |
|---|---|---|---|---|
| **PSNR(dB)** | 37.29 | 40.84 | 40.32 | **42.27** |
| **SSIM** | 0.940 | 0.959 | 0.954 | **0.962** |
| **BER(%)** | 4.6 | 3.3 | 2.7 | **2.1** |

Figure 4: Visual quality and robustness comparison of watermarking methods: (a) Original, (b) StegaStamp (Tancik, Mildenhall, and Ng 2020), (c) PIMoG (Fang et al. 2022), (d) SSDS (Li, Liao, and Wu 2024), (e) Proposed S2R (trained on SIM+LEA).

| Distance = 30 cm | | Training dataset | | | | | |
|---|---|---|---|---|---|---|---|
| | | S+L | | I+E | | M+A | |
| PSNR (dB) | | 41.94 | | 41.00 | | 42.57 | |
| SSIM | | 0.969 | | 0.957 | | 0.964 | |
| Angle | | 0° | 40° | 0° | 40° | 0° | 40° |
| Device pair | S+L | 2.6 | 6.1 | 2.2 | 4.9 | 1.6 | 5.1 |
| | I+E | 2.1 | 5.0 | 2.5 | 4.5 | 2.2 | 5.6 |
| | M+A | 2.2 | 5.3 | 2.3 | 5.3 | 2.0 | 5.9 |

Table 1: Generalization Robustness Test Results of S2R Trained on Datasets with Different Device Pairs.

| Distance = 30 cm | Image quality | | BER (%) | | |
|---|---|---|---|---|---|
| Method | PSNR (dB) | SSIM | 0° | 20° | 40° |
| StegaStamp | 39.89 | 0.948 | 5.5 | 7.1 | 7.3 |
| PIMoG | 41.41 | 0.950 | 6.2 | 8.8 | 9.5 |
| SSDS | 41.05 | 0.956 | 5.1 | 6.0 | 7.6 |
| S2R | **42.27** | **0.962** | **2.1** | **3.3** | **6.0** |

Table 2: Performance Comparison of Models with Different Noise Layers under Same Watermarking Framework.

## Noise Approximation Experiments

In noise approximation experiments, we visually compare the output of different methods and analyze the noise differences. As shown in Figure 3, other methods exhibit limitations in detail and noise realism. StegaStamp produces darker images because of lower saturation, brightness, and contrast. PIMoG improves visually but struggles with lightness variations. SSDS shows improvements but still suffers from darkening. In contrast, S2R generates more natural images, closely resembling real SC noise and bridging the gap between physical and approximate environments. Histogram comparisons further confirm S2R's superior approximation of subtle noise variations, detailed in Section C of the supplementary material. These results confirm that S2R outperforms existing modeling methods in both perceptual fidelity and statistical consistency, offering a more accurate and realistic noise layer.

## Ablation Experiments

**Cross-Device Generalization** To validate the generalization of S2R, we conduct a cross-device generalization test. Specifically, we train the model on datasets from three device pairs (S+L, I+E, M+A) and then test the trained S2R on other device pairs. In the experiment, the shooting conditions are consistent with two capture angles (0° and 40° from the left) for comparison, and a shooting distance of 30 cm is maintained. The results of the generalization experiments, as shown in Table 1, indicate that despite hardware differences between devices, the S2R is still able to effectively extract and recover the watermark, demonstrating high accuracy between datasets from different devices.

**Cross-Dataset Generalization** We further evaluate the generalization of the framework generalization by testing on datasets unseen during watermarking training. Despite variations in image semantics and textures, the method maintains similar robustness and perceptual quality as on the original dataset, demonstrating strong adaptability under domain shifts. This robustness mainly arises from the model learning SC noise characteristics rather than content-specific features. Detailed results are reported in Section D.1 of the supplementary material.

**Cross-Source Training for S2R and Watermarking** To further test the generalization, we train the noise generation module and watermarking network on different datasets. Even with this cross-source training, the system maintains low BER and preserves image quality, with only minor perceptual quality drops due to source differences. These findings confirm that our framework supports flexible, decoupled training of its components without sacrificing robustness. Additional analyses are available in Section D.2 of the supplementary material.

**Comparison of Different Noise Layers Under the Same Watermarking Framework** To validate the effectiveness of the S2R noise layer, we train multiple models under an identical watermarking framework, MCFN (Wu et al. 2024), with different noise layers. We select different noise layers from StegaStamp, PIMoG, and the state-of-the-art SSDS, which are widely adopted in contemporary SC resistant watermarking schemes, as baselines, and compare their performance with S2R. As shown in Table 2, S2R outperforms other methods in all aspects. This is because the core advantage of S2R lies in the ability to adaptively adjust the noise

| BER (%) | Angle = 0° | | | | | Distance = 30 cm (Left to Right) | | | | | | Distance = 30 cm (Up to Down) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Methods | 20cm | 25cm | 30cm | 35cm | 40cm | -60° | -40° | -20° | 20° | 40° | 60° | -60° | -40° | -20° | 20° | 40° | 60° |
| StegaStamp | 2.9 | 3.9 | 4.6 | 4.7 | 4.4 | 5.9 | 7.2 | 4.1 | 5.8 | 7.7 | 7.6 | 15.1 | 4.8 | 4.7 | 3.5 | 5.7 | 9.4 |
| PIMoG | 1.5 | 1.4 | 3.3 | 3.2 | 2.6 | 9.0 | 8.7 | 5.2 | 5.3 | 9.3 | 9.7 | 15.3 | 4.7 | 4.5 | 2.1 | 5.0 | 6.5 |
| SSDS | 2.4 | 2.7 | 2.1 | 2.7 | 4.1 | 7.5 | 5.1 | 3.9 | 4.2 | 6.1 | 6.2 | 9.3 | 4.0 | 3.3 | 2.9 | 4.8 | 6.2 |
| S2R (SIM+LEA) | **1.2** | **1.1** | **2.1** | **2.5** | **2.2** | **5.8** | **3.9** | **3.2** | **3.3** | **6.0** | **5.9** | 10.2 | **3.3** | **2.3** | **1.8** | **4.2** | 6.3 |

Table 3: Comparison of Bit Accuracy for Extracted Watermark Message under Different Capture Distances and Angles.

distribution. The model can rely on predefined mathematical modeling of noise and bridge the noise distribution based on the statistical characteristics of the data. Therefore, S2R can more accurately approximate the noise distribution in real SC, surpassing traditional noise layer designs.

## Comparison Experiments

**Comparison of Visual Quality** In this experiment, we compare the invisibility of watermark images generated by different methods, with the results presented in Figure 4. Compared to other methods, S2R reduces visible artifacts due to the superior design of its noise layer, which bridges the gap between mathematical modeling and real-world distortions. This design enables S2R to better approximate the statistical properties and intensity distribution of real SC noise, allowing the adversarial training process to be more closely aligned with real scenes. Consequently, the model requires less effort to counteract extraneous noise, enhancing its visual quality in real SC scenarios.

**Comparison of Robustness Under Different Shooting Distances** To assess robustness against variations in shooting distance, we conduct experiments across five distances ranging from 20 cm to 40 cm under fixed perpendicular capture angles. As shown in Table 3, S2R consistently achieves the lowest BER on all distances, outperforming existing methods. This robustness is primarily due to S2R's ability to handle resolution degradation and potential defocus effects introduced by distance changes. By combining mathematical modeling simulation and unsupervised refinement, S2R learns to adapt to these changes and maintain reliable watermark recovery in real SC scenarios.

**Comparison of Robustness Under Different Shooting Angles** We conduct experiments by capturing watermark images at angles ranging from 0° to 60°, both from left to right and from bottom to top, and compare the BER. The experimental results in Table 3 show that within smaller angle ranges, the model is able to recover the watermark effectively, with the BER remaining at a low level. At larger shooting angles, the degradation in watermarked image quality results in increased BER. However, S2R consistently achieves lower error rates than competing approaches, indicating stronger robustness to angle-induced distortions.

## Scalability Experiments

To assess the scalability of S2R and compare different noise modeling strategies, we integrate noise layers from StegaStamp, SSDS, and PIMoG into the unsupervised S2R framework. These variants, denoted as StegaStamp-based, SSDS-based, and PIMoG-based (S2R), demonstrate that the framework supports plug-and-play replacement of degradation models without altering its core structure. To further assess S2R, we train a supervised variant of S2R to learn a direct mapping from sharp to noisy images based on 900 paired samples from I+E. As shown in Table 4, although the S2R-supervised benefits from paired supervision, its performance remains limited due to insufficient generalization and reliance on labeled data. In addition, we implement DualGAN and CycleGAN as alternative unsupervised methods. Although they underperform compared to S2R, both methods outperform counterparts that do not employ unsupervised networks for noise modeling. This highlights the advantage of incorporating learning-based strategies over purely mathematical modeling. In summary, S2R achieves greater robustness, imperceptibility, and generalization.

| Distance = 30 cm | PSNR | BER (%) | | |
|---|---|---|---|---|
| Model | | 0° | 20° | 40° |
| StegaStamp-based (SIM+LEA) | 40.47 | 2.4 | 3.7 | 7.1 |
| SSDS-based (SIM+LEA) | 41.25 | 5.0 | 8.1 | 10.6 |
| S2R-supervised (I+E) | 41.29 | 3.8 | 5.5 | 7.9 |
| S2R-CycleGAN (SIM+LEA) | 41.85 | 2.9 | 4.5 | 6.9 |
| S2R-DualGAN (SIM+LEA) | 41.55 | 3.5 | 5.2 | 7.6 |
| S2R (I+E) | **42.57** | **1.6** | **3.1** | **5.1** |
| S2R (SIM+LEA) | 42.27 | 2.1 | 3.3 | 6.0 |

Table 4: Comparison of Models with Different Mathematical Modeling Methods and Supervised Methods.

# Conclusion

In this paper, we propose a novel SC-resistant watermarking framework, Simulation-to-Real (S2R), which improves robustness against SC noise in real-world scenarios. S2R adopts a two-step strategy: it first simulates rough noise via mathematical modeling, then refines them with an unsupervised image-to-image network to approximate real SC noise. Experiments show that S2R achieves superior robustness and generalization compared to existing methods. In future work, we will enhance flexibility by enabling adaptive noise refining during end-to-end training.

## Acknowledgments

This work is supported by National Key R&D Program of China (Grant Nos. 2024YFF0618800, 2022YFB3103500), National Natural Science Foundation of China (Grant Nos. U22A2030, U22B2062), Hunan Provincial Funds for Distinguished Young Scholars (Grant No. 2024JJ2025), Hunan Provincial Key Research and Development Program (Grant Nos. 2024AQ2027, 2025AQ2022), Nanjing Major Science and Technology Special Project (Grant No. 202405002).

## References

Chang, Y.; Guo, Y.; Ye, Y.; Yu, C.; Zhu, L.; Zhao, X.; Yan, L.; and Tian, Y. 2023. Unsupervised deraining: Where asymmetric contrastive learning meets self-similarity. *IEEE Trans. Pattern Anal. Mach. Intell.*

Fan, M.; Lai, S.; Huang, J.; Wei, X.; Chai, Z.; Luo, J.; and Wei, X. 2021. Rethinking bisenet for real-time semantic segmentation. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 9716–9725.

Fang, H.; Chen, D.; Huang, Q.; Zhang, J.; Ma, Z.; Zhang, W.; and Yu, N. 2020. Deep template-based watermarking. *IEEE Trans. Circuits Syst. Video Technol*, 31(4): 1436–1451.

Fang, H.; Jia, Z.; Ma, Z.; Chang, E.-C.; and Zhang, W. 2022. Pimog: An effective screen-shooting noise-layer simulation for deep-learning-based watermarking network. In *Proc. ACM Int. Conf. Multimedia*, 2267–2275.

Fang, H.; Zhang, W.; Zhou, H.; Cui, H.; and Yu, N. 2018. Screen-shooting resilient watermarking. *IEEE Trans. Inf. Forensics Secur.*, 14(6): 1403–1418.

Fu, L.; Liao, X.; Guo, J.; Dong, L.; and Qin, Z. 2024. WaveRecovery: Screen-shooting Watermarking based on Wavelet and Recovery. *IEEE Trans. Circuits Syst. Video Technol.*

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *Neural Inf. Process. Syst. (NIPS)*, 27.

Gulrajani, I.; Ahmed, F.; Arjovsky, M.; Dumoulin, V.; and Courville, A. C. 2017. Improved training of wasserstein gans. *Neural Inf. Process. Syst. (NIPS)*, 30.

He, M.; Wang, H.; Zhang, F.; and Xiang, Y. 2024. Exploring Accurate Invariants on Polar Harmonic Fourier Moments in Polar Coordinates for Robust Image Watermarking. *IEEE Trans. Multimedia*, 26: 5435–5449.

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-image translation with conditional adversarial networks. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 1125–1134.

Jia, Z.; Fang, H.; and Zhang, W. 2021. Mbrs: Enhancing robustness of dnn-based watermarking by mini-batch of real and simulated jpeg compression. In *Proc. ACM Int. Conf. Multimedia*, 41–49.

Li, Y.; Liao, X.; and Wu, X. 2024. Screen-Shooting Resistant Watermarking with Grayscale Deviation Simulation. *IEEE Trans. Multimedia.*

Lim, J. S. 1990. *Two-dimensional signal and image processing*. Prentice-Hall, Inc.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 740–755.

Luo, X.; Zhan, R.; Chang, H.; Yang, F.; and Milanfar, P. 2020. Distortion agnostic deep watermarking. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 13548–13557.

Pang, T.; Zheng, H.; Quan, Y.; and Ji, H. 2021. Recorrupted-to-recorrupted: Unsupervised deep learning for image denoising. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2043–2052.

Pham, B.-D.; Tran, P.; Tran, A.; Pham, C.; Nguyen, R.; and Hoai, M. 2024. Blur2Blur: Blur Conversion for Unsupervised Image Deblurring on Unknown Domains. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2804–2813.

Qin, C.; Li, X.; Zhang, Z.; Li, F.; Zhang, X.; and Feng, G. 2023. Print-camera resistant image watermarking with deep noise simulation and constrained learning. *IEEE Trans. Multimedia.*

Tancik, M.; Mildenhall, B.; and Ng, R. 2020. Stegastamp: Invisible hyperlinks in physical photographs. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2117–2126.

Wang, B.; Wu, Y.; and Wang, G. 2023. Adaptor: Improving the robustness and imperceptibility of watermarking by the adaptive strength factor. *IEEE Trans. Circuits Syst. Video Technol.*

Wang, K.; Wu, S.; Yin, X.; Lu, W.; Luo, X.; and Yang, R. 2024. Robust image watermarking with synchronization using template enhanced-extracted network. *IEEE Trans. Circuits Syst. Video Technol.*

Wengrowski, E.; and Dana, K. 2019. Light field messaging with deep photographic steganography. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 1515–1524.

Wu, Y.; Wang, B.; Wang, G.; and Liao, X. 2024. MCFN: Multi-scale Crossover Feed-forward Network for high performance watermarking. *Neurocomputing*, 129282.

Yi, Z.; Zhang, H.; Tan, P.; and Gong, M. 2017. Dualgan: Unsupervised dual learning for image-to-image translation. In *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2849–2857.

Zhu, J.; Kaplan, R.; Johnson, J.; and Fei-Fei, L. 2018. Hidden: Hiding data with deep networks. In *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 657–672.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2223–2232.

Zhu, L.; Fang, Y.; Zhao, Y.; Peng, Y.; Wang, J.; and Ni, J. 2024. Lite Localization Network and DUE-Based Watermarking for Color Image Copyright Protection. *IEEE Trans. Circuits Syst. Video Technol.*