Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.077803

ARTICLE

ArtFlow: Flow-Based Watermarking for High-Quality Artwork Images Protection

Yuanjing Luo[1,2,#], **Xichen Tan**[1,#], **Yinuo Jiang**[1] **and Zhiping Cai**[1,*]

[1]College of Computer Science and Technology, National University of Defense Technology, Changsha, China
[2]College of Computer and Mathematics, Central South University of Forestry and Technology, Changsha, China

*Corresponding Author: Zhiping Cai. Email: zpcai@nudt.edu.cn

[#]These authors contributed equally to this work

Received: 17 December 2025; Accepted: 18 March 2026; Published: 08 May 2026

ABSTRACT: With increasing artwork plagiarism incidents, the necessity of using digital watermarking technology for high-quality artwork copyright protection is evident. Current digital watermarking methods are limited in imperceptibility and robustness. To address this, based on comprehensive copyright protection research, we develop a novel watermark framework named ArtFlow, using Invertible Neural Networks (INN). Our framework treats watermark embedding and recovery as inverse image transformations, implemented through forward and reverse processes of INN. To ensure high-quality watermark embedding, we utilize frequency domain transformations and attention mechanisms to guide the watermark into high-frequency areas of the image that have greater protective weighting. These areas are attractive to plagiarizers yet have minimal impact on the artistic integrity of the artwork itself. For strong plagiarism-resistant, we design a noise layer that includes various infringement methods—transmission, plagiarism action, and camera-shooting—to train robust watermark recovery process. Additionally, an image quality enhancement module is introduced to minimize the distortions that may arise from infringement before the watermark recovery. Experimental results across four datasets confirm that our ArtFlow surpasses existing advanced watermarking methods.

KEYWORDS: Deep watermarking; invertible neural networks; artwork copyright protection; plagiarism resistance

1 Introduction

"*Over 80% of the items created with this tool were plagiarized works, fake collections, and spam*[1] ", reported by Opensea, the largest marketplace for non-fungible tokens (NFTs). This widespread issue is concerning, especially given the broad and easy access to networks that exposes original artworks to large-scale plagiarism [1,2]. Unfortunately, high-quality artwork images—such as photographs, digital paintings, and NFTs [3]—have increasingly become unintended "victims" of this trend, much to the dismay of numerous designers. These artworks, often used to express ideas and sentiments, are vulnerable to unauthorized use and duplication. In response, both academic and industrial sectors have stepped up efforts in copyright protection [4], concentrating on advancements in technical measures and enhancements in legal frameworks to combat plagiarism[2].

It is encouraging to observe that digital watermarking, a leading technique for copyright protection, has been extensively adopted across various sectors, including social media and artistic creation, among

[1]https://stealthoptional.com/crypto/opensea-80-percent-nfts-scams
[2]www.creativebloq.com/features/how-can-designers-deal-with-plagiarism

others [5,6]. Traditionally, unique watermarks are crafted by extracting distinctive information through the process of image transformation [7], yet these techniques have been criticized for causing significant visual distortion [8], which can detract from the viewer's experience. Emerging with the rise of deep learning, the auto-encoder approach has risen to prominence within the realm of digital watermarking, prized for its capacity for imperceptible information concealment and as an innovative solution for plagiarism detection [9]. Throughout the end-to-end training process, auto-encoder models are tailored to integrate novel components and accommodate minor distortions for optimization [10–17], granting them a level of robustness in extracting watermarks from partially altered artwork. Despite these advantages, it is important to recognize that the auto-encoder architecture has inherent limitations due to its relatively simplistic embedding approach: The encoder embeds the watermark into the cover image, while the noise layer applies various differentiable distortions. The decoder then tries to extract the watermark from these images. Although joint training typically ensures robustness, the automatic end-to-end training of the framework might be undermined by the weak coupling between the encoder and decoder. This issue arises because they are constructed as two parameter-unshared forward networks, connected merely by simple concatenation. Such tenuous link in the process can lead to the inadvertent omission of critical data during the forward propagation, resulting in issues such as color aberrations and the replication of textural artifacts. This can be particularly detrimental to the integrity of artwork images.

In response to this challenge, previous studies have suggested harnessing normalized flow through the application of Invertible Neural Networks (INN) for image concealment tasks, which treat the processes of hiding and revealing images as reversible. This approach aims to retain the fine details of the input, showing potential benefits over traditional auto-encoder models. However, while these ready-made methods, such as HiNet [18] and ISN [19], offer promising results and significant utility in image concealment, they do not perfectly align with our specific needs: 1) an excessive reliance on reversibility at the expense of robustness, making the embedded information susceptible to broken; 2) reversible mechanisms may be exploited by those intent on copyright infringement to recovery original images devoid of watermarks, a.k.a., watermark removal, which constitutes a severe breach of intellectual property rights.

Motivated by the initial achievements of INN in the realm of image hiding, we previously utilized INN in our watermark framework IRWArt [20], embedding watermarks into the high-frequency areas of artwork images to protect copyrights against common forms of plagiarism. Although this work demonstrated the efficacy of using INN for watermarking artwork images, it addressed only a limited range of plagiarism scenarios and did not optimally leverage the unique features of the artworks, leaving room for improvement. Building upon comprehensive copyright protection research, we further develop a new watermark framework, ArtFlow, using a flow-based paradigm. This framework treats watermark embedding and recovery as inverse image transformations, achieved through the forward and reverse processing of INN. To ensure high-quality watermark embedding, we employ frequency domain transformations and attention mechanisms to direct the watermark into areas of the image with higher protective weighting. These areas are attractive to plagiarizers yet minimally impact the integrity of the artwork itself. To enhance anti-plagiarism capabilities, we refined the construction of the noise layer to include various infringement methods (such as transmission, plagiarism action, and camera-shooting), coupled with an image quality enhancement module to train robust watermark recovery processes. Moreover, ArtFlow continues to use contrastive learning, considering the embedded image and the recovered watermark as positive samples derived from the original, while the recovered cover image is regarded as a negative sample aimed at thwarting the removal of watermarks. The main contributions are outlined as follows:

- We design an end-to-end deep watermarking network architecture dedicated to protecting precious artwork images. This framework provides high embedding visual quality and is effective in common plagiarism scenarios.

- We explore four key insights into artwork images that inform our system design, featuring specialized anti-plagiarism noise layers and a highlight-guided embedding strategy. We further integrate an enhancement module and contrastive learning-based loss functions to weaken ArtFlow's dependency on reversible blocks.
- Empirical studies, encompassing both qualitative and quantitative analyses across four distinct datasets, reveal that ArtFlow outperforms five state-of-the-art (SoTA) methods, showcasing superior imperceptibility (11.7%↓ image distortion rate) and enhanced robustness evidenced (11.6%↓ watermark distortion rate).

The subsequent sections of this paper are structured as follows: Section 2 provides an overview of pertinent literature concerning watermarking techniques, the application of invertible neural networks, and the role of attention mechanisms. Section 3 elucidates four pivotal discoveries pertinent to the domain of artwork images. Section 4 delineates the detailed network architecture and the adopted training methodologies. The experimental configurations and the ensuing analyses are sequentially detailed in Sections 5 and 6. Section 7 offers concluding remarks on the research presented.

2 Related Work

2.1 Watermarking Approaches

Digital watermarking embeds short messages into images [21] for authorship statements, necessitating high imperceptibility and robustness [22]. Traditional watermarking strategies have primarily leaned on human intuition and manually devised methods for selecting suitable pixels for information embedding [23,24], coupled with the development of sophisticated encoding mechanisms [25,26]. While these methods have demonstrated efficacy to some extent. Frequently, they raise statistical red flags and prove insufficiently resilient to the manipulations associated with plagiarism. This vulnerability largely stems from their design, which is tailored to resist specific types of attacks but often falls short when confronted with novel, unforeseen attacks [27].

With the development of deep learning, a new horizon in watermarking techniques has been unveiled. Deep learning-based models, particularly those employing convolutional neural networks (CNNs) for separate encoder and decoder designs, have introduced a paradigm shift [10–17,28–30]. These models facilitate the incorporation of innovative modules aimed at optimization, thereby achieving performance metrics that significantly outstrip those of their traditional counterparts. The process begins with an input watermark and original image, whereby the encoder is tasked with generating an encoded image. This encoded image is visually indistinguishable from the original, yet it hides the watermark in such a manner that the decoder can accurately recover it. HiDDeN [28] pioneered the auto-encoder architecture by introducing the joint training of the encoder and decoderwith an additional noise layer and deploying a suite of novel training strategies. This foundational work has paved the way for the subsequent evolution of numerous sophisticated auto-encoder-based watermarking approaches [10–16]. Udh [17] refines the architecture by streamlining the encoder input to be exclusively associated with the watermark, thereby unlocking new avenues for exploration and fostering innovation within the watermarking research domain. Despite these advancements, challenges remain. In these methodologies, the encoding and decoding processes, which are sequential and operate through two distinct forward networks without shared parameters, frequently lead to the inadvertent loss of essential information during the encoding phase. Consequently, it poses a challenge for current auto-encoder techniques to strike a balance between producing high-quality encoded images and retrieving watermarks with fidelity, potentially leading to issues such as color distortion and the replication of textural features, as highlighted in [18].

2.2 Invertible Neural Network

Lately, invertible neural networks have gained popularity due to their ability to facilitate reversible image transformations by learning a stable, invertible mapping between data and latent distributions [31,32]. Given a variable y and the forward computation $x = f_\theta(y)$, y can be recovered directly by $y = f_\theta^{-1}(x)$, where the inverse function f_θ^{-1} shares same θ with $x = f_\theta$ [33]. INN achieves effective retention of input details by incorporating both forward and backward propagation mechanisms within a unified network framework and utilizing supplementary implicit output variables to safeguard information that could be otherwise forfeited during the forward traversal [19]. Consequently, INNs have demonstrated exceptional performance across a multitude of image-centric applications, including, but not limited to, image colorization [34], rescaling [31,35], and compression [36]. The processes of embedding and recovering a watermark, similarly, can be conceptualized as invertible operations executed via an INN's forward and reverse functions. Previous applications of INNs to image hiding [18,19,33,37,38] have not fully addressed robustness against plagiarism or the risk of "reversible structures being exploited for watermark removal", posing significant copyright infringement issues. Thus, these approaches cannot be directly applied to protect artwork images' copyrights. While our prior work [20] has improved upon these aspects by developing an INN-based watermarking framework tailored for artwork copyright protection, thereby proving the effectiveness of INNs for watermarking artwork images, it covered limited plagiarism scenarios and did not optimally leverage the unique characteristics of artwork during watermark embedding, leaving room for further improvement.

2.3 Attention Mechanism

Attention mechanisms enhance deep learning models by focusing on crucial input data for tasks. This includes spatial attention targeting specific locations within images, channel attention focusing on the content aspects of images, and the Convolutional Block Attention Module (CBAM) that synergizes both channel and spatial attentiveness [39]. Integrating attention mechanisms in image information hiding allows for selective attention to key content while disregarding low-perceptual information, thus effectively guiding the embedding locations. In existing research, various methods have achieved higher embedded image quality by dynamically adjusting channel features within deep representations of images [38,40,41], assigning different levels of importance to individual pixels [16], and mining the global features of the original image to generate attention masks [12,42]. However, these approaches are either designed for CNNs, not suitable for INNs' reversible nature [12,16,40–42], or they integrate with INNs for multi-image hiding without considering the robustness of the embedded image [38].

3 Understanding Artwork Images

Understanding the expectations for protecting artwork images is crucial before designing a watermarking scheme. We began by consulting 26 design students and professionals, conducting both online and in-person interviews. Except for one participant who believed artwork should remain unaltered, the majority favored watermarks that ensure *complete imperceptibility* (23/25) and *plagiarism resistance* (21/25). A typical comment was from participant #9: "*Watermarks are important for proving ownership, but they shouldn't affect the artwork's appearance.*"

From these insights, we proceeded to explore two critical questions: **[Q1]** Which areas of the artwork are least impacted by the embedding of a watermark? **[Q2]** Beyond simple duplication or network transmission, what techniques or processes does plagiarism typically involve?

Frequency domain analysis for Q1. Analyzing watermark embedding in the frequency domain, which divides image information into high-frequency (e.g., textures and edges) and low-frequency (e.g., smooth areas) components, offers a precise approach to identifying optimal embedding regions compared to the

pixel domain [43]. To validate the effectiveness of watermark embedding across various frequency bands of artwork images, we randomly curated a collection of 100 pieces from the Wiki Art database [44]. Utilizing wavelet transformation, we embedded watermarks into the respective LL, LH, HL, and HH sub-bands of these artworks. Subsequently, we assessed the quality of the original and watermarked images, as well as the original and recovered watermarks, through the computation of Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) values. The findings, depicted in Fig. 1, reveal an intriguing outcome. While our initial goal was to identify the embedding region that minimally affects the original image's integrity, it was discovered that the high-frequency region not only fulfilled this criterion but also facilitated the most effective watermark recovery (highest PSNR & SSIM).

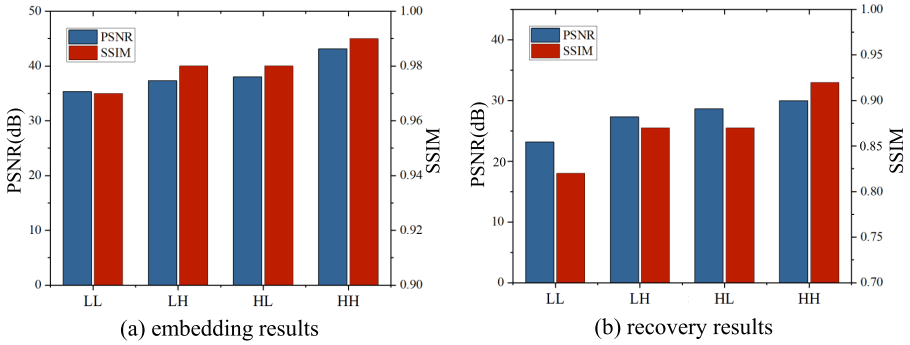

Figure 1: The artwork images' embedding and extraction outcomes are evaluated across the LL, LH, HL, and HH sub-bands. Superior image quality is denoted by elevated PSNR and SSIM values, signifying enhanced fidelity and structural similarity.

Remark 1: The high frequency area of the image is the most suitable for watermark embedding.

Investigations for Q2. In reality, the act of plagiarism is intricate and manifests in numerous forms [45]. To explore the nuances of image plagiarism, our prior research [20] engaged a focus group comprising 10 design majors. Each participant was briefed on the study's objectives. Throughout the focus group discussions, five distinct image plagiarism manipulations were identified: image cropping, where parts of the image are removed; image stretching, which alters the aspect ratio; adding or deleting patterns, such as overlaying text or objects to cover original elements; color adjustment, which modifies hue, saturation, or brightness; and angle adjustment, a.k.a., rotation, which changes the orientation of the image.

Remark 2: Typical plagiarism processing actions include image cropping, image stretching, adding or deleting patterns, color adjustment, and angle adjustment.

Given the above processing actions, we additionally gather 207 anonymous questionnaires from a diverse pool of respondents, with over half being design students or professionals. Our analysis focuses on common techniques associated with plagiarism that participants have encountered or might consider using. Statistically, *subject elements copy* (73.47% positive[3]) stands out as the preferred method, surpassing other techniques such as composition copy (55.1% positive), color copy (28.57% positive). Conversations with design professionals indicate that plagiarizers would try to reuse the most captivating aspects of a design and minimize the efforts of such processes if s/he intended to produce his/her own work via plagiarism.

Remark 3: The subject elements of an artwork represent the focal points of design, capturing a viewer's attention through their intricate textures and vibrant color schemes. These elements are frequently the target of plagiarism, often coupled with the act of cropping to tailor the copied content.

[3] Positive indicates a score > 3 on our 5-level Likert scale.

During our investigation, we also discover a widespread and overlooked infringement: unauthorized camera-shooting, which often occurs at art exhibitions. Due to lax oversight, visitors can easily capture images of both offline exhibits and digital displays using their mobile phones or cameras. These images may then be used for unauthorized reproduction, sale, or online distribution. To further examine this issue, we conduct field visits to three major local art exhibitions, where we randomly interview 40 visitors to analyze their photographic behaviors and intentions. The survey revealed that over 60% of visitors admitted to photographing artworks, with more than half clarifying that they had no intention of using the images for commercial or other illicit purposes. As one visitor (#12) poignantly stated, *"I'm really sorry if I hurt the creator without meaning to. I just wanted to snap these pics to show off their beauty online..."*

Remark 4: Unauthorized camera-shooting of artwork is a common, yet often overlooked infringement that, despite originating from photographers' lack of copyright awareness, poses a potentially significant threat to artists' rights.

4 The Proposed Approach

4.1 Motivation

Building upon the above insights, we are pleasantly surprised to discover that the high-frequency areas of an image, ideal for watermark embedding (*Remark 1*), coincide with regions of heightened interest. These areas, rich in complex textures, are prone to plagiarism and thus essential for protection (*Remark 3*). Motivated by this observation, *we aim to embed watermarks in these high-frequency, high-interest areas to enhance copyright protection*. Additionally, we recognize the necessity of designing a robust watermarking system capable of resisting potential copyright infringement attacks (*Remark 2, Remark 4*). Thus, in the development of the noise layer of our watermarking model, *we are motivated to guide the embedded watermark to resist potential infringement, e.g., plagiarism processing and camera-shooting*. This strategic approach not only safeguards key image areas but also strengthens against future plagiarism attempts.

4.2 Overview

ArtFlow's primary objective is to craft an all-encompassing framework dedicated to the copyright protection of artwork images. Drawing inspiration from the aforementioned characteristics of artworks, our approach is to *strategically embed the watermark within the highlight regions of the artworkand make it plagiarism/capture-resistant* when devising the framework. Table 1 presents the main notations employed throughout this manuscript. Fig. 2 illustrates the overarching architecture of the proposed ArtFlow system, which encompasses a Flow-based Invertible Module featuring re-architected DenseNets integrated with spatial-channel attention mechanisms, an array of Noise Layers, and a dedicated Quality Enhancement Module. Within the ArtFlow framework, the tasks of watermark embedding and extraction are conceptualized as a set of inverse operations:

$$\begin{aligned} \mathbf{I_{em}} &= AF_{Fwd}(\mathbf{I_{co}}, \mathbf{W_{or}}) \\ (\mathbf{I_{co}}, \mathbf{W_{or}}) &= AF_{Bwd}(\mathbf{I_{em}}), \end{aligned} \tag{1}$$

where the embedding function $AF_{Fwd}(\cdot)$ is derived from ArtFlow's forward processing, while the recovery function $AF_{Bwd}(\cdot)$ is derived from its backward processing. During the forward embedding phase, the framework takes in a cover image $\mathbf{I_{co}}$ and watermark $\mathbf{W_{or}}$ as inputs. These inputs are initially processed through a Discrete Wavelet Transform (DWT), which decomposes them into low and high-frequency wavelet sub-bands. These sub-bands are then sequentially introduced into a series of invertible blocks. The culminating output from this series undergoes an Inverse Wavelet Transform (IWT) to produce the

embedded image $\mathbf{I_{em}}$, alongside the incidental loss information $\mathbf{z}$. For the backward recovery phase, the embedded image $\mathbf{I_{em}}$ is first enhanced by a Quality Enhancement Module (QEM) to precondition the input for the reverse operation. Subsequently, akin to the embedding procedure, an auxiliary variable $\widetilde{\mathbf{z}}$ accompanies the enhanced image $\mathbf{I_{em}}$ through a frequency domain transformation and traverses a sequence of invertible operations to facilitate the recovery of the watermark $\mathbf{W_{re}}$. This process treats $\mathbf{I_{em}}$ and $\mathbf{W_{re}}$ as proximate 'positive samples' of the original inputs $\mathbf{I_{co}}$ and $\mathbf{W_{or}}$, respectively, while $(\mathbf{I_{re}})$ is designated as a distant 'negative sample' of $\mathbf{I_{co}}$. The objective is for the positive samples to converge closely, while maintaining a wider separation from the negative sample, an outcome attainable through supervised contrastive learning. Moreover, noise layers are strategically positioned between the forward and reverse phases to bolster the system's resilience against the distortions in $\mathbf{I_{em}}$ due to plagiarism or capture interventions.

Table 1: The notations.

Notation	Description
$\mathbf{I_{co}}$	The cover image to be protected
$\mathbf{W_{or}}$	The original watermark
$\mathbf{I_{em}}$	The image with watermark embedded
$\mathbf{W_{re}}$	The recovered watermark from $\mathbf{I_{em}}$
$\mathbf{I_{re}}$	The recovered cover image from $\mathbf{I_{em}}$
$\mathbf{I_{fake}}$	The fake image unrelated to the cover image
$\mathbf{z}$	The information lost in forward propagation
$\widetilde{\mathbf{z}}$	The auxiliary variable to help recover $\mathbf{W_{re}}$
$AF_{Fwd}(\cdot)$	The forward process of ArtFlow
$AF_{Bwd}(\cdot)$	The backward process of ArtFlow
Θ_{Fwd}	The forward training variables
Θ_{Bwd}	The backward training variables

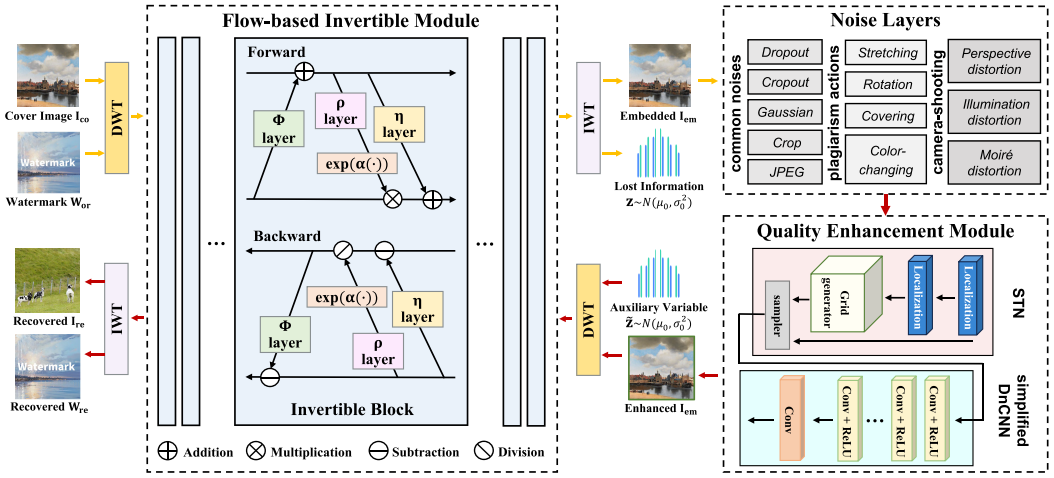

Figure 2: The framework of ArtFlow (the case image is posted by Johannes Vermeer on WikiArt. https://www.wikiart.org/en/johannes-vermeer/view-on-delft). The ArtFlow utilizes a flow-based Invertible module with several neural blocks for forward-embedding (marked by yellow arrows) and backward-recovery (marked by red arrows). A noise layer applied between passes distorts the embedded image for recovery training. Additionally, a quality enhancement module alleviates distortion and perspective changes during watermark recovery.

4.3 Network Architectures

4.3.1 Flow-Based Invertible Module

Flow-based invertible module consists of several invertible blocks. For the i-th invertible block in the forward operation, $i \in \{1, \ldots, 15\}$, the inputs are $\mathbf{I}_{\mathbf{co}}^{\mathbf{i}}$ and $\mathbf{W}_{\mathbf{or}}^{\mathbf{i}}$, and the corresponding outputs $\mathbf{I}_{\mathbf{co}}^{\mathbf{i+1}}$ and $\mathbf{W}_{\mathbf{or}}^{\mathbf{i+1}}$ are formulated as follows:

$$\begin{aligned} \mathbf{I}_{\mathbf{co}}^{\mathbf{i+1}} &= \mathbf{I}_{\mathbf{co}}^{\mathbf{i}} + \phi(\mathbf{W}_{\mathbf{or}}^{\mathbf{i}}) \\ \mathbf{W}_{\mathbf{or}}^{\mathbf{i+1}} &= \exp(\alpha(\rho(\mathbf{I}_{\mathbf{co}}^{\mathbf{i+1}}))) \odot \mathbf{W}_{\mathbf{or}}^{\mathbf{i}} + \eta(\mathbf{I}_{\mathbf{co}}^{\mathbf{i+1}}), \end{aligned} \tag{2}$$

where exp denotes the natural exponential function, $\odot$ signifies the Hadamard product and α is a sigmoid function scaled by a constant factor served as a clamp. $\phi(\cdot)$, $\rho(\cdot)$ and $\eta(\cdot)$ are arbitrary functions, represented by 5-layer dense blocks. To enhance the network's focus on pertinent features while ensuring structural reversibility, we adopt a re-engineered dense architecture with spatial-channel attention for $\phi(\cdot)$, $\rho(\cdot)$ and $\eta(\cdot)$. Following the final block in the forward pass, we implement the Inverse Wavelet Transform (IWT) on the two resulting outputs, $\mathbf{I}_{\mathbf{co}}^{\mathbf{16}}$ and $\mathbf{W}_{\mathbf{or}}^{\mathbf{16}}$, to synthesize the watermarked image $\mathbf{I}_{\mathbf{em}}$ and the residual information $\mathbf{z}$. This $\mathbf{z}$ encapsulates both the lost watermark data and the degraded cover image details. Consequently, in the reverse operation, the auxiliary variable $\widetilde{\mathbf{z}}$ is leveraged to precisely reconstruct the watermark $\mathbf{W}_{\mathbf{re}}$. This is drawn from a distribution that is independent of the specific case and is anticipated to mirror the statistical properties of $\mathbf{z}$. The characteristics of this distribution are established during training through the recovery loss, as detailed in Section 4.5. The specific backward propagation operation we employ is as follows:

$$\begin{aligned} \mathbf{I}_{\mathbf{em}}^{\mathbf{i}} &= \mathbf{I}_{\mathbf{em}}^{\mathbf{i+1}} - \phi(\widetilde{\mathbf{z}}^{\mathbf{i}}) \\ \widetilde{\mathbf{z}}^{\mathbf{i}} &= \exp(-\alpha(\rho(\mathbf{I}_{\mathbf{em}}^{\mathbf{i+1}}))) \odot (\widetilde{\mathbf{z}}^{(\mathbf{i+1})} - \eta(\mathbf{I}_{\mathbf{em}}^{\mathbf{i+1}})), \end{aligned} \tag{3}$$

where the input $\widetilde{\mathbf{z}}^{\mathbf{16}}$ is generated by the auxiliary variable $\widetilde{\mathbf{z}}$ performing DWT, and $\widetilde{\mathbf{z}}$ is randomly sampled from a Gaussian distribution, i.e., $\widetilde{\mathbf{z}} \sim N(\mu_0, \sigma_2^0)$. After the last block in the backward operation, the output $\mathbf{I}_{\mathbf{em}}^{\mathbf{1}}$ is processed through IWT to generate the recovery watermark $\mathbf{W}_{\mathbf{re}}$.

4.3.2 Noise Layers

Our objective is to create a watermarking model that is resistant to plagiarism. Implementing adversarial learning within carefully crafted noise layers enhances the robustness of the embedded watermarks [46,47]. Based on *Remark 2* and *Remark 4*, and accounting for typical image transmission losses, we strategically incorporate three distinct types of noise at the juncture between the forward and reverse processes. Fig. 3 illustrates an instance of each noise variant.

- Common transmission distortions processing. Regarding these previously considered distortions, we meticulously adhere to the established modification parameter settings as recognized in existing scholarly works [17,28], meticulously applying Dropout, Cropout, Gaussian blurs, Crop, and JPEG compresses.
- Plagiarism action-incurred noises. Within each training batch, we execute a range of distortions: 80%–90% random cropping, 110%–120% random stretching, rotations through random angles between 5°–10°, the application of a 5 × 5 white patch at random locations, and color-changing with shifts randomly selected within the interval $(-5, +5)$ degrees. These adversarial samples are evenly apportioned to encompass all types of noise.

- Camera-shooting distortions processing. Addressing distortions typically introduced by camera capture, we follow protocols established in [48], which include Perspective distortion, Illumination distortion, and Moiré distortion processing to closely replicate real-world conditions.

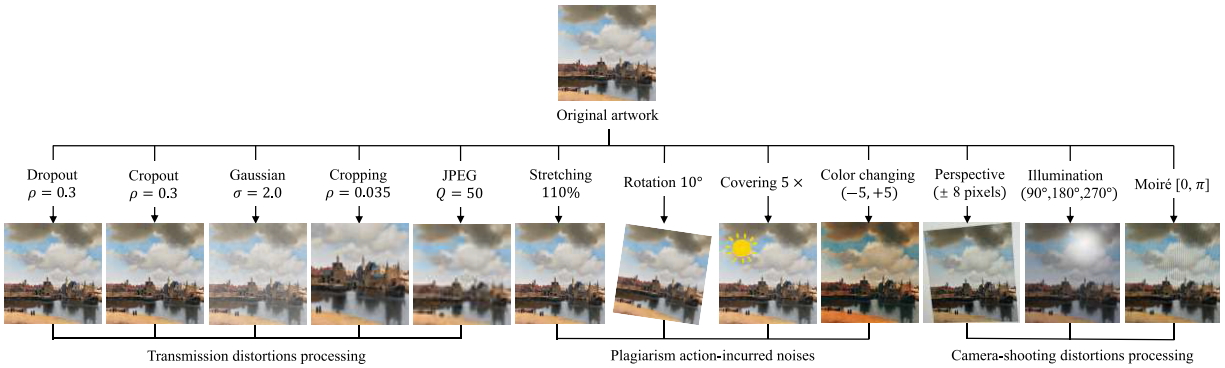

Figure 3: Illustration of noise layers.

4.3.3 Quality Enhancement Module (QEM)

In anticipation of the reverse operation, we have crafted a Quality Enhancement Module (QEM), meticulously designed to counteract the effects of distortions or minor perspective shifts in $\mathbf{I_{em}}$ that may arise from plagiarism/camera-shooting. As depicted in Fig. 4, this module incorporates two core components: a lightweight Spatial Transformer Network (STN) [49] and a simplified version of DnCNN. The STN includes a Localization Network, a Grid Generator, and a Sampler. The Localization Network uses convolutional and fully connected layers to learn the input's spatial transformation parameters. The Grid Generator produces a sampling grid, and the Sampler adjusts the input based on this grid to maintain spatial invariance for $\mathbf{I_{em}}$. DnCNN, a classic image denoising architecture, has been modified in our approach by removing its batch normalization layers and retaining only the Conv-ReLU cascade structure, effectively providing denoising for the distorted $\mathbf{I_{em}}$. By integrating QEM into the recovery process, $\mathbf{I_{em}}$ undergoes preprocessing before entering the backward reversible block, ensuring high-quality inputs for the backward pass.

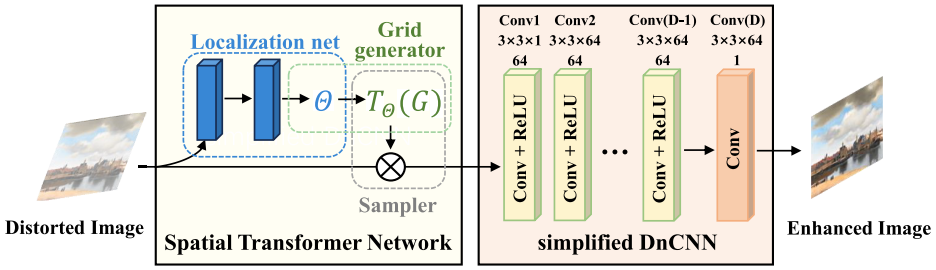

Figure 4: The architecture of our quality enhancement module (QEM).

4.4 Highlight-Guidance Embedding Strategy (HES)

Drawing upon the insights from *Remark 1* and *Remark 3*, we develop a Highlight-guidance Embedding Strategy (HES) to embed watermarks within areas of high frequency & interest, operationalized in two ways:

Wavelet domain embedding preference. Leveraging the perfect reconstruction and bidirectional symmetry inherent in wavelet theory, we employ the Haar wavelet kernel to perform DWT and IWT. During the forward pass, prior to engagement with the invertible blocks, the original cover image $\mathbf{I_{co}}$

undergoes DWT, decomposing it into low and high-frequency components. This transformation reshapes the feature map dimensions from (C, H, W) to $(4C, H/2, W/2)$, with C, H, and W representing the number of channels, height, and width, respectively. The network then targets the high-frequency sub-band for watermark embedding. Post the final invertible block, the IWT is invoked to synthesize the watermarked image $\mathbf{I_{em}}$, effectively reverting the feature map dimensions from $(4C, H/2, W/2)$ to (C, H, W).

Concatenated channel-spatial attention mechanism. As mentioned in Section 4.3.1, $\phi(\cdot)$, $\rho(\cdot)$ and $\eta(\cdot)$, founded on dense architecture, capture only rudimentary image features. Drawing inspiration from [39], we refine the DenseNet module with a concatenated channel-spatial attention layer to enhance the focus on image high-interest details. As illustrated in Fig. 5, in the channel attention part, we aggregate a feature map's spatial details through average and max pooling to obtain two distinct context descriptors: $\mathbf{F^c_{avg}}$ and $\mathbf{F^c_{max}}$, representing the mean and peak features, respectively. These descriptors are processed by a shared multi-layer perceptron (MLP) with a single hidden layer to generate the channel attention map $\mathbf{M_c} \in \mathbb{R}^{C\times1\times1}$:

$$\mathbf{M_c(F)} = \sigma(MLP(AvgPool(\mathbf{F})) + (MLP(MaxPool(\mathbf{F})). \tag{4}$$

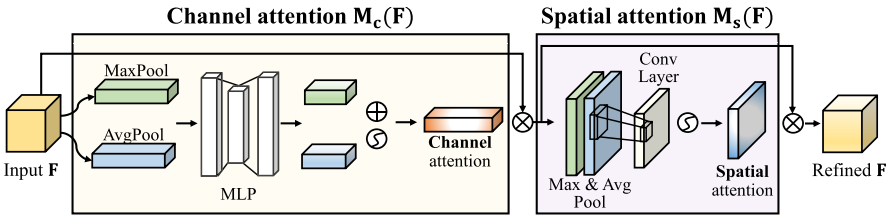

Figure 5: The detailed structure of concatenated channel-spatial attention.

Regarding the spatial attention part, channel information from a feature map is pooled using average and max operations to form $\mathbf{F^s_{avg}} \in \mathbb{R}^{1\times H\times W}$ and $\mathbf{F^s_{max}} \in \mathbb{R}^{1\times H\times W}$, each representing pooled features across channels. Concatenation and convolution with a standard layer produce the spatial attention map:

$$\mathbf{M_s(F)} = \sigma(Conv([AvgPool(\mathbf{F}); MaxPool(\mathbf{F})])), \tag{5}$$

where σ is the sigmoid function and $Conv$ is a convolution with a 7×7 kernel.

4.5 Loss Function

The overall loss function is decomposed into three key components: the embedding loss for watermark embedding quality, the recovery loss for assessing watermark recovery accuracy, and the anti-removal loss for the resilience of the cover image restoration.

Embedding loss. The forward process of ArtFlow usually requires that the watermark should be embedded covertly with high perceptual quality, i.e., the generated $\mathbf{I_{em}}$ is indistinguishable from $\mathbf{I_{co}}$. The corresponding loss $\mathcal{L}_E$ can be defined as:

$$\mathcal{L}_E = \sum_{n=1}^{N} \ell_e(\mathbf{I^{(n)}_{co}}, \mathbf{I^{(n)}_{em}}), \tag{6}$$

where N represents the number of training samples. ℓ_e quantifies the discrepancy between $\mathbf{I_{co}}$ and $\mathbf{I_{em}}$, incorporating a low-frequency wavelet loss ℓ_{freq} [50] to ensure high-frequency embedding:

$$\ell_{freq} = \ell_1(\mathcal{H}(\mathbf{I_{co}}), \mathcal{H}(\mathbf{I_{em}})) = (\mathcal{H}(\mathbf{I_{co}}) - \mathcal{H}(\mathbf{I_{em}}))^2, \tag{7}$$

where $\mathcal{H}(\cdot)$ means the operation of extracting low-frequency sub-bands after wavelet decomposition; a ℓ_2 norm to guide pixel-level reconstruction:

$$\ell_2 = \|\mathbf{I_{co}} - \mathbf{I_{em}}\|_2^2/(C \cdot H \cdot W)); \tag{8}$$

a perceptual loss ℓ_{lpips} [51] and a negative cosine similarity loss ℓ_{ncs} [52] to supervise perceptual improvement:

$$\ell_{lpips} = \sum_l \frac{1}{H_l W_l} \sum_{h,w} \left\| w_l \odot (\mathbf{I}_{\mathbf{co}}^{\hat{l}} - \mathbf{I}_{\mathbf{em}}^{\hat{l}}) \right\|_2^2, \tag{9}$$

$$\ell_{ncs} = -\frac{\mathcal{P}(\mathbf{I_{co}})^T \mathcal{P}(\mathbf{I_{em}})}{(\tau \|\mathcal{P}(\mathbf{I_{co}})\| \, \|\mathcal{P}(\mathbf{I_{em}})\|)}, \tag{10}$$

where $\mathbf{I}_{\mathbf{co}}^{\hat{l}}, \mathbf{I}_{\mathbf{em}}^{\hat{l}} \in \mathbb{R}^{H_l \times W_l \times C_l}$ for layer l are derived from the unit-normalized feature stack along the channel dimension, extracted across L layers of the VGG network. The vector $w^l \in \mathbb{R}^{C_l}$ scales the activations channel-wise. For each layer l, $w_l = 1$ is used to compute the cosine distance. The operation $\mathcal{P}(\cdot)$ denotes the generation of feature vectors via networks, and τ represents a temperature parameter.

Recovery loss. The backward process of ArtFlow usually requires that the watermark can be recovered using any sample of $\widetilde{z}$ from the Gaussian distribution $p(z)$, and the recovered $\mathbf{W_{re}}$ needs to closely match the original version. Thus, the recovery loss $\mathcal{L}_R$ is delineated as follows:

$$\mathcal{L}_R = \sum_{n=1}^{N} \mathbb{E}_{\widetilde{z} \sim p(\widetilde{z})} [\ell_r(\mathbf{W}_{\mathbf{or}}^{(\mathbf{n})}, \mathbf{W}_{\mathbf{re}}^{(\mathbf{n})}]. \tag{11}$$

Similar to ℓ_e, ℓ_r measures the difference between $\mathbf{W_{or}}$ and $\mathbf{W_{re}}$, which consists of ℓ_2 norm and ℓ_{ncs}.

Anti-removal loss. To prevent the removal of the embedded watermark, that is, to avoid retrieving an unmarked cover image during the reverse watermark recovery process, the recovered image $\mathbf{I_{re}}$ should substantially differ from the cover image $\mathbf{I_{co}}$, approaching the appearance of an entirely unrelated fake image $\mathbf{I_{fake}}$. This optimization goal is achieved using the specified contrastive loss:

$$\mathcal{L}_{AR} = \sum_{n=1}^{N} \left(-log \left(\frac{\exp(sim(\mathcal{P}(\mathbf{I}_{\mathbf{re}}^{(\mathbf{n})}), \mathcal{P}(\mathbf{I}_{\mathbf{fake}}^{(\mathbf{n})})/\tau)}{\exp(sim(\mathcal{P}(\mathbf{I}_{\mathbf{re}}^{(\mathbf{n})}), \mathcal{P}(\mathbf{I}_{\mathbf{co}}^{(\mathbf{n})})/\tau)} \right) \right). \tag{12}$$

In this setup, $\mathbf{I_{fake}}$ and $\mathbf{I_{co}}$ are treated as positive and negative samples of $\mathbf{I_{re}}$, respectively. The function $sim(\cdot, \cdot)$ measures the cosine similarity between the two feature vectors.

Total loss. By integrating these three types of distortion losses, we achieve our ultimate optimization objective:

$$\mathcal{L}_{total} = \lambda_1 \cdot \mathcal{L}_E + \lambda_2 \cdot \mathcal{L}_R + \lambda_3 \cdot \mathcal{L}_{AR}, \tag{13}$$

where λ controls the relative weights of the losses.

Note that the noise disturbances we added for robustness improvement are not entirely reversible. To tailor the model to handle irreversible transformations and to learn to counteract the impact of quantization errors and noise, we adopt a two-stage training approach inspired by previous works [10,33]. Initially, we conduct joint end-to-end training of the network without noise, focusing on minimizing $\mathcal{L}_{total}$. Following this, we concentrate solely on refining the backward pass under adversarial conditions by optimizing the

recovery loss $\mathcal{L}_R$, effectively setting λ_1 and λ_3 to 0. The training procedure of our ArtFlow is outlined in Algorithm 1.

Algorithm 1: The training process of our ArtFlow

Input: The images $\mathbf{I_{co}}$, $\mathbf{W_{or}}$ and $\mathbf{I_{fake}}$.
Output: Trained network $AG(\cdot)$.
Initialize $\Theta_{Fwd}, \Theta_{Bwd}$ with Gaussian initialization.
if Stage I training **then**
  while $Step < max\_steps$ **do**
    Compute $\mathbf{I_{em}} = AF_{Fwd}(\mathbf{I_{co}}, \mathbf{W_{or}})$.
    Compute $\mathbf{W_{re}} = AF_{Bwd}(\mathbf{I_{em}}, \widetilde{\mathbf{z}})$.
    Compute $\mathbf{I_{re}} = AF_{Bwd}(\mathbf{I_{em}}, \widetilde{\mathbf{z}})$.
    Update $\Theta_{Fwd} \leftarrow \Theta_{Fwd} + lr \times \text{Adam}(\mathcal{L}_{total})$.
    Update $\Theta_{Bwd} \leftarrow \Theta_{Bwd} + lr \times \text{Adam}(\mathcal{L}_{total})$.
  end while
end if
if Stage II training **then**
  while $Step < max\_steps$ **do**
    Attack $\mathbf{I_{em}}$.
    Enhance $\mathbf{I_{em}}$ via QEM.
    Compute $\mathbf{W_{re}} = AF_{Bwd}(\mathbf{I_{em}}, \widetilde{\mathbf{z}})$.
    Update $\Theta_{Bwd} \leftarrow \Theta_{Bwd} + lr \times \text{Adam}(\mathcal{L}_R)$.
  end while
end if
return $\Theta_{Fwd}, \Theta_{Bwd}$.

5 Experimental Setup

5.1 Datasets

The experiments are conducted across three image datasets, namely DIV2K [53], MS COCO, Wiki Art [44] and LOGO_IRWArt [20].

- **DIV2K**, consisting of 1000 high-resolution images split into 800 for training and 200 for validation and testing, is used for training ArtFlow. Half of the training images are randomly chosen as cover patches, while the rest are used for watermark and counterfeit patches. Model evaluation is performed on the validation and test sets using these images as cover images.
- **COCO**, vast in scale with over 330,000 images and 220,000 annotated. We have randomly selected 1800 images for testing as cover images.
- **Wiki Art** offers a rich collection of paintings from 195 distinct artists, totaling 42,129 images for training and 10,628 for testing. We randomly choose 6000 images as cover images for the purpose of testing.
- **LOGO_IRWArt** encompasses a collection of 8000 LOGOs sourced online. All of these images are selected to serve as test watermarks in our evaluation.

5.2 Implementation Details

ArtFlow is implemented with PyTorch 1.10.0 and leverages the computational power of an Nvidia GeForce GTX 3080 Ti GPU for accelerated processing. The training process employs the Adam optimizer, configured with hyperparameters $\beta_1 = 0.9$, $\beta_2 = 0.99$, a learning rate of $\beta = 10^{-5}$ and a mini-batch of size

16. The model processes images in patches sized $C \cdot H \cdot W = 3 \times 512 \times 512$, and undergoes a total of 10,000 iterations. For effective learning, three hyperparameters of the total loss function are set to $\lambda_1{:}\lambda_2{:}\lambda_3 = 2{:}10{:}1$ to balance the contribution of different error terms, optimizing performance across various facets of the training data.

For camera-shooting, we adhere to the mature paradigm described in [17,54], where the distance between camera and display range from 23 cm to 4.3 m, and the shooting angles included frontal and 45°. The camera equipment used is "iPhone 15 Pro Max", and the display is "AOC Q24P1".

5.3 Baselines

We benchmark ArtFlow against several open-source state-of-the-art (SOTA) methods to validate its performance. These include two CNN-based auto-encoder methods and two normalizing flow-based methods:

- **HiDDeN** [28], a standard auto-encoder that encodes both the cover image and watermark with one encoder.
- **Udh** [17], another auto-encoder method focusing on watermark encoding before integration with the cover image.
- **HiNet** [18], a pioneering framework using invertible neural networks (INNs) for joint encoding of cover images and watermarks.
- **IRWArt** [20], our previous work, employing a normalizing flow-based approach for multimedia watermarking.

All baseline models are utilized in their default settings. In our experiments, we made two key adjustments: 1) The HiDDeN model, originally for message embedding, was adapted to output images and retrained accordingly. 2) HiNet, initially not accounting for image distortion, was fine-tuned with our noise parameters, resulting in the **HiNet⁺** model. These changes ensure a fair comparison under consistent conditions.

5.4 Evaluation Metrics

To evaluate our method, we use two key metrics: *Visual Imperceptibility*, measured by the image distortion rate comparing original and watermarked images; and *Anti-Attack Robustness*, measured by the watermark distortion rate under noise conditions. The distortion rate includes PSNR, SSIM, and BER, detailed as follows:

- **PSNR** serves as an objective measure of image quality, defined as follows:

$$PSNR(x, y) = 10 \log_{10} \left(\frac{(MAX_I)^2}{MSE(x, y)} \right), \tag{14}$$

where MAX_I is the maximum possible pixel value of images x and y. $MSE(x, y)$ represents the Mean Squared Error (MSE) between images x and y:

$$MSE = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} \|\mathbf{X}(i, j) - \mathbf{Y}(i, j)\|^2 . \tag{15}$$

- **SSIM** quantifies the resemblance between **X** and **Y**, calculated as follows:

$$SSIM(x, y) = \frac{(2\mu_x \mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}, \tag{16}$$

where μ_x and μ_y indicate the average grayscale values, or means, of **X** and **Y**. Symbol σ_x and σ_y represent the variances of **X** and **Y**. Symbol σ_{xy} represents covariance. $C_1 = (k_1 L)^2$ and $C_2 = (k_2 L)^2$ are two constants which are used to maintain stability when either $\mu_x^2 + \mu_y^2$ or $\sigma_x^2 + \sigma_y^2$ is very close to 0, where $K_1 = 0.01$ and $K_2 = 0.03$. L is the dynamic range of the pixel values.

- **BER** indicates the frequency of bits received in error and is used to assess the extraction effectiveness of embedded binary sequences.

$$BER = \frac{n_{err}}{len(str)}, \tag{17}$$

where n_{err} is the number of error bits, $len(str)$ represents the length of hidden messages.

6 Experimental Results and Discussion

In this section, we first evaluate the visual imperceptibility and anti-attack robustness of our proposed network in Section 6.1 and Section 6.2, respectively, benchmarking it against several SOTA approaches. Subsequently, in Section 6.3, we test the anti-ablation capability of our proposed approach. Finally, in Section 6.4, we discuss the impact of different forms of cover images and watermarks on the performance of our approach.

6.1 Comparison of Visual Imperceptibility

Maintaining high visual imperceptibility is crucial, ensuring the embedded image closely resembles the original with minimal distortion. To assess both objective and subjective image quality, alongside the metrics in Section 5.4, we conducted a user study with 50 volunteers. Participants were presented with five watermarked images from HiDDeN, Udh, HiNet$^+$, IRWArt, and ArtFlow, plus one original image. Unaware of the images' details, they identified any altered images with a '1' and the rest with '0'. The resulting mean opinion scores (MOS) [55] are the final outcomes.

Fig. 6 and Table 2 visualise the results of these qualitative comparisons. It is evident that images embedded with ArtFlow closely resemble the original cover images, exhibiting no artifacts from texture replication. Compared to HiDDeN, Udh, and HiNet$^+$, ArtFlow achieves improvements of 1.15×, 1.17×, and 1.39× in the average values of PSNR and SSIM, respectively. These enhancements are attributed to our reversible embedding architecture and highlight-guidance embedding strategy, enhanced by optimized loss functions such as ℓ_{lpips} and ℓ_{ncs}, which notably improve the perceptual quality of the embedded images. Following closely is IRWArt, whose performance nearly matches that of ArtFlow, thanks to its symmetric embedding framework and the use of perceptual losses. However, it slightly lags in PSNR, as it does not fully leverage the intrinsic features of the artwork images during the watermark embedding process. Udh and HiDDeN, ranking third and fourth, respectively, utilize auto-encoders for watermark embedding, which unfortunately leads to some loss of image features during forward propagation, resulting in suboptimal embedding outcomes. HiNet$^+$, despite also using a reversible neural network, ranks fifth as its embedding performance is compromised by our specific noise settings that lead to asymmetric forward and backward inferences, adversely affecting its performance before fine-tuning. Finally, as expected, ArtFlow achieves the lowest MOS values, underscoring that watermarks embedded through this method are the most indistinguishable, affirming its superiority in maintaining high visual imperceptibility.

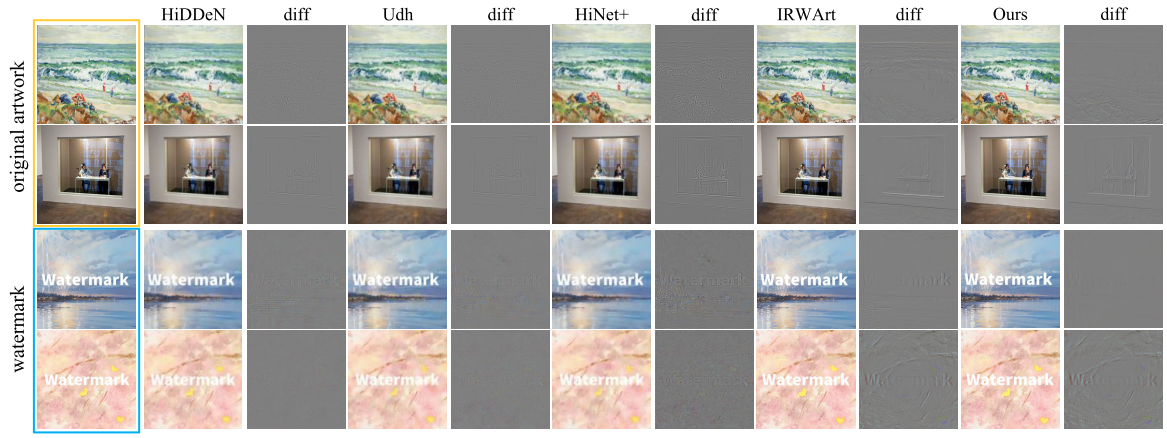

Figure 6: Visual comparisons of embedded images and recovered watermarks of different methods.

Table 2: Visual imperceptibility comparison.

	HiDDeN	Udh	HiNet$^+$	IRWArt	ArtFlow
PSNR(dB) ↑	38.4	37.6	31.8	49.7	49.8
SSIM ↑	0.991	0.974	0.812	0.999	0.999
MOS value↓	0.42	0.78	0.52	0.02	0.02

6.2 Comparison of Anti-Attack Robustness

Robustness is crucial for ensuring that a watermark, once recovered from a noisy environment, closely resembles its original form, thereby maintaining a low rate of watermark distortion. In real-world applications, artworks are frequently subjected to a variety of noise sources, such as transmission distortions, plagiarism, and camera-shooting distortions, all of which can significantly compromise the fidelity of watermark recovery. In this research, we conducted thorough testing of our model against a spectrum of distortions that we anticipated during the evaluation phase. Parameters for transmission distortions were set according to [17,28], while parameters for plagiarism were determined based on findings from our user study (*Remark 2*). For camera-shooting distortions, we followed established protocols outlined in [17,54]. It is worth noting that the watermarks used in our tests are lightly-colored logo images from LOGO_IRWArt, characterized by *a very low fault-tolerant rate*. Even slight distortions are conspicuously noticeable, which makes them particularly suitable for assessing the efficacy of our watermark extraction process.

The outcomes of these tests are detailed in Table 3, where "Identity" represents conditions without any introduced noise, and Fig. 6 displays the watermark recovery in such scenarios. Observations reveal that all five models exhibit commendable robustness against transmission noise. However, in cases involving plagiarism, the robustness of HiDDeN and Udh appears relatively weaker. This vulnerability is primarily due to these models being trained solely with conventional distortion processes. In camera-shooting scenarios, only Udh and our method demonstrated substantial robustness. Although HiNet$^+$ underwent fine-tuning specifically for our noise settings, its performance fell short of expectations. Overall, our ArtFlow model outperformed others across different test environments, with its average PSNR and SSIM values respectively showing improvements of 1.18×, 1.1×, 1.33×, and 1.04× compared to HiDDeN, Udh, HiNet$^+$, and IRWArt. These improvements are largely attributed to the rigorous training of ArtFlow within a carefully designed noise environment and the substantial enhancement provided by our Quality Enhancement Module (QEM).

Table 3: Robustness comparison.

Noise	HiDDeN		Udh		HiNet+		IRWArt		ArtFlow	
	PSNR (dB)↑	SSIM↑	PSNR (dB)↑	SSIM↑	PSNR (dB)↑	SSIM↑	PSNR (dB)↑	SSIM↑	PSNR (dB)↑	SSIM↑
Identity	34.7	0.992	35.9	0.987	28.1	0.872	36.5	0.991	34.6	0.989
Cropout	26.5	0.891	29.2	0.896	23.1	0.821	27.5	0.967	28.1	0.965
Dropout	26.4	0.868	27.9	0.861	22.8	0.812	26.8	0.902	27.5	0.932
Gaussian	28.2	0.871	29.7	0.871	17.5	0.632	27.2	0.897	26.9	0.815
JPEG	23.7	0.801	21.1	0.740	20.2	0.827	23.7	0.863	22.9	0.865
Gaussian Filtering	16.1	0.575	15.3	0.502	15.0	0.568	21.7	0.830	21.9	0.821
Cropping	14.1	0.582	13.4	0.547	15.2	0.639	22.6	0.816	23.0	0.819
Streching	20.9	0.875	20.5	0.875	16.9	0.597	22.9	0.835	22.6	0.897
Covering	20.7	0.939	23.3	0.868	12.4	0.694	24.1	0.878	24.5	0.879
Rotation	10.6	0.383	11.7	0.465	11.7	0.511	18.6	0.801	19.1	0.829
Coloring	22.9	0.911	22.3	0.855	15.7	0.724	23.6	0.870	23.9	0.914
Camera-shooting, frontal	14.3	0.574	24.5	0.871	17.5	0.612	13.9	0.506	24.7	0.869
Camera-shooting, 45°	11.2	0.479	21.7	0.860	15.6	0.583	11.7	0.499	21.7	0.849
Average	20.7	0.749	22.8	0.784	17.8	0.684	23.1	0.819	24.7	0.880

6.3 Ablation Study

The ablation experiments are performed on randomly selected 2700 images that are evenly divided to cover each form of noise for watermark recovery testing. In this discussion, we focus on the primary network architectures, including the Noise Layers and QEM, as well as the highlight-guidance embedding strategy, which integrates DWT/IWT and Attention mechanisms. Additionally, we examine the role of Contrastive Loss in influencing the final outcomes.

6.3.1 Effectiveness of Noise Layers

The integration of noise layers significantly boosts ArtFlow's resilience to noisy environments, as demonstrated by the data in the first and sixth rows of Table 4, which show an increase in the PSNR value by 7.1 dB for watermark distortion rates. This enhancement is a direct result of the noise layers forcing the model to develop encodings that are robust enough to endure distortions encountered during transmission. This feature ensures that ArtFlow not only adapts to but effectively counters the adverse effects of environmental noise.

6.3.2 Effectiveness of QEM

The QEM utilizes a STN along with a DnCNN-inspired network to pre-process and optimize the distorted embedded image, effectively mitigating the effects of plagiarism actions. The effectiveness of the QEM is underscored by the favorable outcomes reported in Table 4, which illustrate significant improvements in image quality and robustness due to its implementation. These results validate the crucial role of QEM in the watermark recovery process, emphasizing its contribution to improving the overall performance and reliability of the watermarking system.

Table 4: Ablation studies for network architecture design and training strategy. The sixth row represents our ArtFlow.

Network Architecture		HES		Contrastive	Image Distortion Rate		Watermark Distortion Rate		Watermark
Noise Layers	QEM	DWT/IWT	Attention	Loss	PSNR (dB) ↑	SSIM ↑	PSNR (dB) ↑	SSIM ↑	Removal-resistance?
✗	✓	✓	✓	✓	51.3	0.999	19.4	0.752	✓
✓	✗	✓	✓	✓	49.7	0.999	23.3	0.859	✓
✓	✓	✗	✓	✓	47.3	0.972	25.2	0.904	✓
✓	✓	✓	✗	✓	48.4	0.992	26.1	0.886	✓
✓	✓	✓	✓	✗	47.7	0.996	26.2	0.869	✗
✓	✓	✓	✓	✓	49.7	0.999	26.5	0.887	✓

6.3.3 Effectiveness of HES

The HES notably enhances the imperceptibility of the ArtFlow system. When employing DWT/IWT, the image's PSNR value increases by 2.4 dB. This enhancement is likely due to DWT/IWT's ability to effectively separate low-frequency and high-frequency sub-bands, thereby facilitating the embedding of watermarks into the more appropriate high-frequency domain. Additionally, when utilizing attention mechanisms, the PSNR value increases by 1.3 dB. This improvement may be attributed to the attention mechanism directing the watermark embedding into areas of the image that are more visually engaging and typically feature complex high-frequency textures. The Fourier analysis spectrum displayed in Fig. 7 corroborates our embedding preferences. HES not only optimizes the embedding process but also preserves the integrity of the original image.

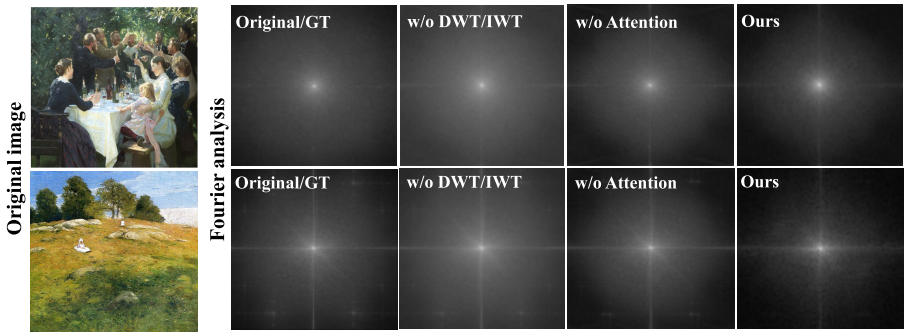

Figure 7: Fourier analysis of cover and embedded images demonstrates that the integration of HES components directs the watermark to be predominantly embedded in the high-frequency areas of cover images.

6.3.4 Effectiveness of Contrastive Loss

Contrastive loss is engineered to improve the visual quality of embedded images and the clarity of recovered watermarks. According to the data shown in the fifth and sixth rows of Table 4, the implementation of contrastive loss results in an increase of 2 and 0.3 dB in the PSNR values for image and watermark distortion rates, respectively. Additionally, contrastive loss plays a crucial role in preventing the effective removal of watermarks. Notably, the un-watermarked cover images recovered with contrastive loss show significant deviation from the original cover images, with a low PSNR value of 9.6 ± 0.2. This substantial disparity acts as a robust defense against unauthorized alterations, highlighting the critical role of contrastive loss in maintaining the integrity and authenticity of digital content.

6.4 Universality

6.4.1 Performance across Various Cover Images

Our watermarking model, developed by harnessing the frequency domain information of artwork images, was trained using the DIV2K dataset and exhibits outstanding performance across four distinct datasets. To further explore the adaptability of our model to various types of cover images, we conducted tests on a diverse array of challenging images, including two monochrome and two random noise images. As depicted in Fig. 8, despite the unique challenges posed by these cover images, our model adeptly embeds watermarks in a highly discreet manner and retrieves them with remarkable precision. This proficiency highlights the model's robust reconstruction capabilities, a key factor for its applicability in real-world scenarios, where the ability to handle a wide range of image types and conditions is essential.

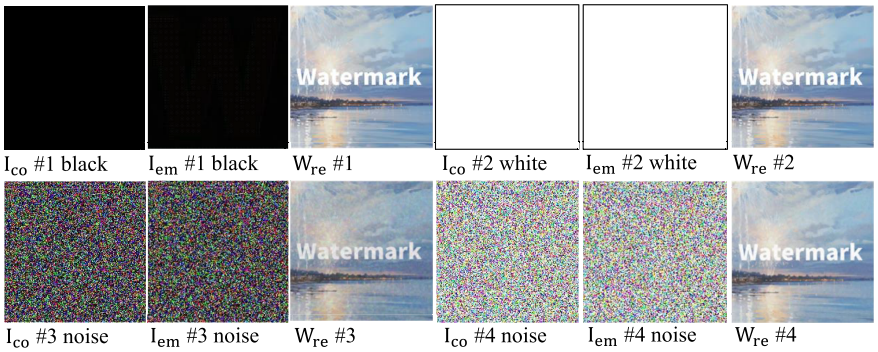

Figure 8: Visual outcomes for a selection of extreme cases, featuring two monochrome images and two images with random noise.

6.4.2 Performance across Various Watermarks

In our study, we selected logo images as watermarks to provide a clear and straightforward method for proving authorship. Recognizing the prevalence of binary messages like barcodes in watermarking, we extended our experimentation to include the embedding of pseudo-binary information using our developed technique. This method is elaborated in Fig. 9, where we depict pseudo-binary messages by dividing a barcode into $m \times n$ patches. Each patch is assigned a uniform value of either 0 or 255. We calculate the average value of each patch, assigning a bit value of 1 if this average surpasses 128, and 0 otherwise, thus encoding the pseudo-binary message into $m \times n$ bits of information.

To assess the robustness of our watermarking method, we conducted tests to measure the Bit Error Rate (BER) across different patch sizes and under various noise conditions. As outlined in Table 5, the BER tends to rise with an increase in the number of embedded bits. Despite this, our technique maintains a low BER against most types of plagiarism attacks, with a notable exception being rotation/shooting at 45°—a vulnerability due to the method's reduced stability to angular changes.

It is important to note that our model was initially trained using general images, not specifically barcodes. Therefore, retraining the model exclusively with barcode data could potentially refine its performance, enhancing its ability to handle specialized data types and further mitigating errors like those observed in specific orientations and conditions. This adaptation could lead to significant improvements in watermark robustness, particularly under challenging conditions that involve rotations and angular distortions.

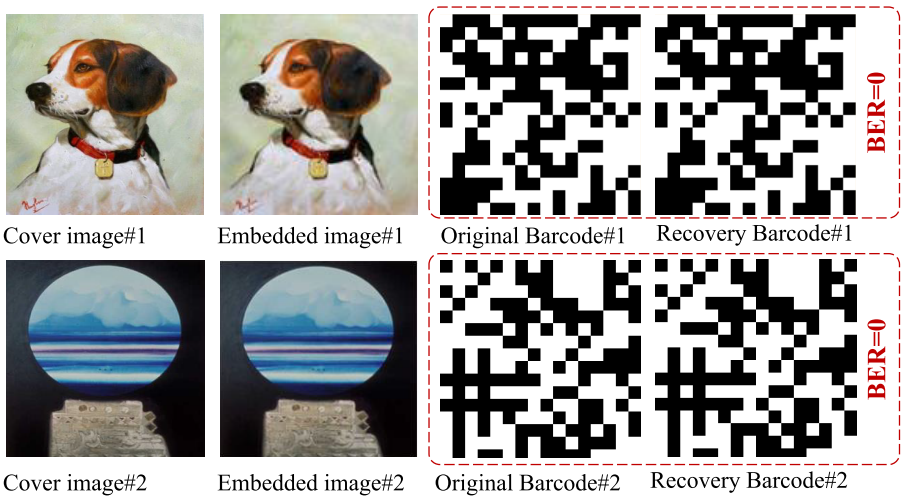

Figure 9: Embedding barcode (256 bits) as the watermark.

Table 5: Bit error rate (%) ↓ when embedding binary sequences.

Noise	Message Length (bits)					
	32	**64**	**128**	**256**	**512**	**1024**
Cropout	0	0	0	3.1	6.8	9.3
Dropout	0	0	0	2.4	9.5	12.7
Gaussian	0	0	0	2.6	12.4	17.5
JPEG	0	0	4.6	9.4	13.7	27.3
Croppiing	0	0	2.2	15.6	17.4	20.4
Streching	0	0	0	0	0	8.4
Covring	0	0	0	0	0	2.5
Rotation	0	16.3	24.8	47.7	50.7	56.7
Coloring	0	0	0	0	0	0
Camera-shooting, frontal	0	3.7	6.5	10.5	15.2	27.9
Camera-shooting, 45°	5.1	19.7	30.4	45.2	50	61.4

7 Conclusion

We introduce ArtFlow, a robust watermarking framework using INN to protect high-quality artworks according to an exploratory study of artworks. This system treats watermark embedding and recovery as inverse transformations, leveraging INN's forward and backward processes. The framework strategically embeds watermarks in high-interest areas with minimal artistic impact through HES. It also incorporates Noise Layers with various infringement scenarios and a QEM to bolster plagiarism-resistant ability. Experiments and visualization analysis demonstrate the superiority of ArtFlow, underscoring its effectiveness in copyright protection.

Acknowledgement: The authors would like to express our sincere gratitude and appreciation to each other for our combined efforts and contributions throughout the course of this research paper.

Funding Statement: This work was supported in part by the National Natural Science Foundation of China under Grants 62172155, 62402171, 62402505 and 62472434; in part by the Science and Technology Innovation Program of Hunan Province under Grant 2022RC3061.

Author Contributions: The authors confirm contribution to the paper as follows: study conception and design: Yuanjing Luo, Xichen Tan, Yinuo Jiang, and Zhiping Cai; data curation: Xichen Tan; formal analysis: Yuanjing Luo and Xichen Tan; writing—original draft: Yuanjing Luo and Xichen Tan; writing—review & editing: Xichen Tan and Yinuo Jiang. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The data and materials used in this study are derived from publicly accessible databases and previously published studies, which are cited throughout the text. References to these sources are provided in the bibliography.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Murray LJ. Plagiarism and copyright infringement. In: Originality, imitation, and plagiarism: teaching writing in the digital age. Ann Arbor, MI, USA: University of Michigan Press; 2008. p. 173–82.
2. Cui S, Liu F, Zhou T, Zhang M. Understanding and identifying artwork plagiarism with the wisdom of designers: a case study on poster artworks. In: MM '22: Proceedings of the 30th ACM International Conference on Multimedia. New York, NY, USA: ACM; 2022. p. 1117–27.
3. Bsteh S. From painting to pixel: understanding NFT artworks [master's thesis]. Rotterdam, The Netherland: Universidad Erasmo Disponible en Formato Digital Aquí; 2021.
4. Adler A. Why art does not need copyright. Geo Wash L Rev. 2018;86(2):313–75.
5. Lee SJ, Jung SH. A survey of watermarking techniques applied to multimedia. In: ISIE 2001. 2001 IEEE International Symposium on Industrial Electronics Proceedings. Piscataway, NJ, USA: IEEE; 2001. p. 272–7.
6. Luo Y, Tan X, Cai Z. Robust deep image watermarking: a survey. Comput Mater Contin. 2024;81(1):133–60. doi:10.32604/cmc.2024.055150.
7. Cox I, Miller M, Bloom J, Fridrich J, Kalker T. Digital watermarking and steganography. 2nd ed. Amsterdam, The Netherland: Elsevier Inc.; 2007.
8. Panetta KA, Wharton EJ, Agaian SS. Human visual system-based image enhancement and logarithmic contrast measure. IEEE Trans Syst Man Cybern B Cybern. 2008;38(1):174–88. doi:10.1109/tsmcb.2007.909440.
9. Berghel H, O'Gorman L. Protecting ownership rights through digital watermarking. Computer. 1996;29(7):101–3. doi:10.1109/2.511977.
10. Liu Y, Guo M, Zhang J, Zhu Y, Xie X. A novel two-stage separable deep learning framework for practical blind watermarking. In: MM '19: Proceedings of the 27th ACM International Conference on Multimedia. New York, NY, USA: ACM; 2019. p. 1509–17.
11. Luo X, Zhan R, Chang H, Yang F, Milanfar P. Distortion agnostic deep watermarking. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE; 2020. p. 13548–57.
12. Yu C. Attention based data hiding with generative adversarial networks. Proc AAAI Conf Artif Intell. 2020;34(1):1120–8. doi:10.1609/aaai.v34i01.5463.
13. Jia J, Gao Z, Chen K, Hu M, Min X, Zhai G, et al. RIHOOP: robust invisible hyperlinks in offline and online photographs. IEEE Trans Cybern. 2022;52(7):7094–7106. doi:10.1109/tcyb.2020.3037208.
14. Ahmadi M, Norouzi A, Karimi N, Samavi S, Emami A. ReDMark: framework for residual diffusion watermarking based on deep networks. Expert Syst Appl. 2020;146:113157.
15. Zhong X, Huang PC, Mastorakis S, Shih FY. An automated and robust image watermarking scheme based on deep neural networks. IEEE Trans Multim. 2020;23:1951–61. doi:10.1109/tmm.2020.3006415.

16. Zhang H, Wang H, Li Y, Cao Y, Shen C. Robust watermarking using inverse gradient attention. arXiv:2011.10850v1. 2020.
17. Zhang C, Benz P, Karjauv A, Sun G, Kweon IS. UDH: universal deep hiding for steganography, watermarking, and light field messaging. In: NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc.; 2020. p. 10223–34.
18. Jing J, Deng X, Xu M, Wang J, Guan Z. HiNet: deep image hiding by invertible network. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway, NJ, USA: IEEE; 2021. p. 4733–42.
19. Lu SP, Wang R, Zhong T, Rosin PL. Large-capacity image steganography based on invertible neural networks. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway, NJ, USA: IEEE; 2021. p. 10816–25.
20. Luo Y, Zhou T, Liu F, Cai Z. IRWArt: levering watermarking performance for protecting high-quality artwork images. In: WWW '23: Proceedings of the ACM Web Conference 2023. New York, NY, USA: ACM; 2023. p. 2340–8.
21. Cox I, Miller M, Bloom J, Honsinger C. Digital watermarking. J Electron Imaging. 2002;11(3):414–4. doi:10.1117/1.1494075.
22. O'Ruanaidh JJ, Dowling W, Boland FM. Watermarking digital images for copyright protection. IEE Proc Vis Image Signal Process. 1996;143(4):250–6. doi:10.1049/ip-vis:19960711.
23. Jia J, Gao Z, Zhu D, Min X, Hu M, Zhai G. RIVIE: robust inherent video information embedding. IEEE Trans Multimedia. 2023;25:7364–77. doi:10.1109/tmm.2022.3221894.
24. Li W, Wang H, Chen Y, Abdullahi SM, Luo J. Constructing immunized stego-image for secure steganography via artificial immune system. IEEE Trans Multimedia. 2023;25(2):8320–33. doi:10.1109/tmm.2023.3234812.
25. Hsu CT, Wu JL. Hidden digital watermarks in images. IEEE Trans Image Process. 1999;8(1):58–68. doi:10.1109/83.736686.
26. Barni M, Bartolini F, Piva A. Improved wavelet-based watermarking through pixel-wise masking. IEEE Trans Image Process. 2001;10(5):783–91. doi:10.1109/83.918570.
27. Mun SM, Nam SH, Jang H, Kim D, Lee HK. Finding robust domain from attacks: a learning framework for blind watermarking. Neurocomputing. 2019;337:191–202.
28. Zhu J, Kaplan R, Johnson J, Fei-Fei L. Hidden: hiding data with deep networks. In: Computer Vision—ECCV 2018: 15th European Conference. Cham, Switzerland: Springer; 2018. p. 657–72.
29. Zhang R, Dong S, Liu J. Invisible steganography via generative adversarial networks. Multimed Tools Appl. 2019;78(7):8559–75. doi:10.1007/s11042-018-6951-z.
30. Zheng Z, Hu Y, Bin Y, Xu X, Yang Y, Shen HT. Composition-aware image steganography through adversarial self-generated supervision. IEEE Trans Neural Netw Learn Syst. 2023;34(11):9451–65. doi:10.1109/tnnls.2022.3175627.
31. Gilbert AC, Zhang Y, Lee K, Zhang Y, Lee H. Towards understanding the invertibility of convolutional neural networks. In: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17). Palo Alto, CA, USA: AAAI Press; 2017. p. 1703–10.
32. van der Ouderaa TF, Worrall DE. Reversible GANs for memory-efficient image-to-image translation. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE; 2019. p. 4720–8.
33. Xu Y, Mou C, Hu Y, Xie J, Zhang J. Robust invertible image steganography. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE; 2022. p. 7875–84.
34. Ardizzone L, Kruse J, Wirkert S, Rahner D, Pellegrini EW, Klessen RS, et al. Analyzing inverse problems with invertible neural networks. arXiv:1808.04730v1. 2018.
35. Xiao M, Zheng S, Liu C, Wang Y, He D, Ke G, et al. Invertible image rescaling. In: Computer Vision—ECCV 2020: 16th European Conference. Cham, Switzerland: Springer; 2020. p. 126–44.
36. Song Y, Meng C, Ermon S. MintNet: building invertible neural networks with masked convolutions. arXiv:1907.07945. 2019.
37. Fang H, Qiu Y, Chen K, Zhang J, Zhang W, Chang EC. Flow-based robust watermarking with invertible noise layer for black-box distortions. In: Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto, CA, USA: AAAI Press; 2023. p. 5054–61.

38. Li F, Sheng Y, Zhang X, Qin C. iSCMIS: spatial-channel attention based deep invertible network for multi-image steganography. IEEE Trans Multimedia. 2024;26:3137–52. doi:10.1109/tmm.2023.3307970.
39. Woo S, Park J, Lee JY, Kweon IS. CBAM: convolutional block attention module. In: Computer Vision—ECCV 2018: 15th European Conference. Cham, Switzerland: Springer; 2018. p. 3–19.
40. Tan J, Liao X, Liu J, Cao Y, Jiang H. Channel attention image steganography with generative adversarial networks. IEEE Trans Netw Sci Eng. 2021;9(2):888–903. doi:10.1109/tnse.2021.3139671.
41. Cao F, Guo D, Wang T, Yao H, Li J, Qin C. Universal screen-shooting robust image watermarking with channel-attention in DCT domain. Expert Syst Appl. 2024;238(2):122062. doi:10.1016/j.eswa.2023.122062.
42. Huang J, Luo T, Li L, Yang G, Xu H, Chang CC. ARWGAN: attention-guided robust image watermarking model based on GAN. IEEE Trans Instrum Meas. 2023;72:5018417.
43. Weng X, Li Y, Chi L, Mu Y. High-capacity convolutional video steganography with temporal residual modeling. In: ICMR '19: Proceedings of the 2019 on International Conference on Multimedia Retrieval. New York, NY, USA: ACM; 2019. p. 87–95.
44. Mohammad S, Kiritchenko S. WikiArt emotions: an annotated dataset of emotions evoked by art. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). Paris, France: ELRA; 2018.
45. Lang Y, He Y, Yang F, Dong J, Xue H. Which is plagiarism: fashion image retrieval based on regional representation for design protection. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE; 2020. p. 2595–604.
46. Song C, Sudirman S, Merabti M, Llewellyn-Jones D. Analysis of digital image watermark attacks. In: CCNC'10: Proceedings of the 7th IEEE Conference on Consumer Communications and Networking Conference. Piscataway, NJ, USA: IEEE; 2010. p. 941–5.
47. Hayes J, Danezis G. Generating steganographic images via adversarial training. In: NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc.; 2017. p. 1951–60.
48. Fang H, Jia Z, Ma Z, Chang EC, Zhang W. PIMoG: an effective screen-shooting noise-layer simulation for deep-learning-based watermarking network. In: Proceedings of the 30th ACM International Conference on Multimedia. New York, NY, USA: ACM; 2022. p. 2267–75.
49. Jaderberg M, Simonyan K, Zisserman A, Kavukcuoglu K. Spatial transformer networks. In: NIPS'15: Proceedings of the 29th International Conference on Neural Information Processing Systems. Cambridge, MA, USA: MIT Press; 2015. p. 2017–25.
50. Baluja S. Hiding images in plain sight: deep steganography. In: NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc.; 2017. p. 2066–76.
51. Zhang R, Isola P, Efros AA, Shechtman E, Wang O. The unreasonable effectiveness of deep features as a perceptual metric. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE; 2018. p. 586–95.
52. Chen T, Kornblith S, Norouzi M, Hinton G. A simple framework for contrastive learning of visual representations. In: ICML'20: Proceedings of the 37th International Conference on Machine Learning. London, UK: PMLR; 2020. p. 1597–607.
53. Timofte R, Agustsson E, Gool LV, Yang MH, Zhang L. NTIRE 2017 challenge on single image super-resolution: methods and results. In: 2017 Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Piscataway, NJ, USA: IEEE; 2017. p. 114–25.
54. Wengrowski E, Dana K. Light field messaging with deep photographic steganography. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE; 2019. p. 1515–24.
55. Streijl RC, Winkler S, Hands DS. Mean opinion score (MOS) revisited: methods and applications, limitations and alternatives. Multimedia Syst. 2016;22(2):213–27. doi:10.1007/s00530-014-0446-1.