This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Robust Invertible Image Steganography

Youmin Xu[†], Chong Mou[†], Yujie Hu[†], Jingfen Xie[†], Jian Zhang[†,‡]
[†]Peking University Shenzhen Graduate School, Shenzhen, China
[‡]Peng Cheng Laboratory, Shenzhen, China
youmin.xu@stu.pku.edu.cn; eechongm@gmail.com; hhuyujie@stu.pku.edu.cn;
xiejf@stu.pku.edu.cn; zhangjian.sz@pku.edu.cn

Abstract

Image steganography aims to hide secret images into a container image, where the secret is hidden from human vision and can be restored when necessary. Previous image steganography methods are limited in hiding capacity and robustness, commonly vulnerable to distortion on container images such as Gaussian noise, Poisson noise, and lossy compression. This paper presents a novel flow-based framework for robust invertible image steganography, dubbed as RIIS. A conditional normalizing flow is introduced to model the distribution of the redundant high-frequency component with the condition of the container image. Moreover, a well-designed container enhancement module (CEM) also contributes to the robust reconstruction. To regulate the network parameters for different distortion levels, a distortion-guided modulation (DGM) is implemented over flow-based blocks to make it a one-size-fits-all model. In terms of both clean and distorted image steganography, extensive experiments reveal that the proposed RIIS efficiently improves the robustness while maintaining imperceptibility and capacity. As far as we know, we are the first to propose a learning-based scheme to enhance the robustness of image steganography in the literature. The guarantee of steganography robustness significantly broadens the application of steganography in real-world applications.

1. Introduction

Steganography is a widely studied topic [12], which aims to hide messages like audio, image, and hyperlink into one container in an undetected way. In Fig. 1, image steganography takes the secret and host image as input to produce the container image. In its reverse process, it is only possible for the receivers with a specific revealing network to reconstruct secret information from the container image, which is visually identical to the host image. Steganalysis techniques usually distinguish the container and host image by color, frequency, and other features. Thus the secret image should be hidden in the invisible domain of the container image. It is also valuable in applications to embed as much confidential data as possible into the host image, which is evaluated as payload capacity.

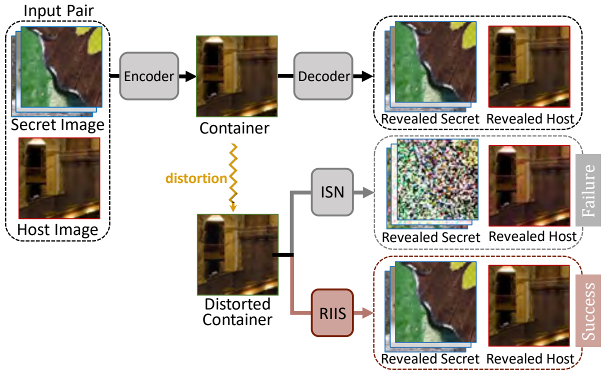

Figure 1. **The upper row depicts the universal pipeline of image steganography.** Previous steganography like ISN [35] gains poor revealed secret and revealed host image when the container is under slight distortion. On the contrary, our RIIS takes various distortion into consideration, which shows satisfactory robustness.

The image steganography is designed to keep the hiding capacity while considering security and imperceptibility against steganalysis. Existing steganography schemes [11, 43, 59] fail to strike a balance between imperceptibility and high payload capacity. Traditional methods transform the secret messages in the spatial or adaptive domains [29], achieving the capacities of $0.2 \sim 4$ bits per pixel (bpp). The secret data is usually embedded into fewer significance bits [11] or indistinguishable parts, limiting the amount of secret information capacity. Recent learning-based steganography methods [7, 8] make an effort to exploit the potential capacity of secret. Most of them take the pre-processing, concealing, and revealing as separate modules and design specific

This work was supported in part by Shenzhen Fundamental Research Program (No.GXWD20201231165807007-20200807164903001). (*Corresponding author: Jian Zhang.*)

networks with independent parameters to handle them.

Recent attempts [50] to introduce invertible neural networks (INN) into low-level inverse problems like denoising, rescaling, and colorization show impressive potential over auto-encoder, GAN [2, 52], and other learning-based architectures. The image steganography composed of concealing and revealing process can be considered as a pair of inverse problems. Thus flow-based INN is naturally suitable for this task. Besides, multiple secret images can be easily hidden into one container by increasing the number of channels of INN branches. That incredibly improves steganography capacity and makes ISN [35] the state-of-the-art image hiding technique in the literature.

Since the earlier image steganography works stress capacity and invisibility rather than robustness and ignores the noise and compression interference in practice, they are usually sensitive to the interference during the media spread of container. Due to the dependence on inherent invertible bijective transformation property, flows tend to be more vulnerable to intermediate distortion [27, 31, 41]. In Fig. 1, we take the state-off-the-art ISN [35] for example. Once a slight noise or lossy compression is implemented on the container, the secret revealed is barely recognizable and also the host image at the receiver end. Even if the network is specifically finetuned against pre-defined noise or JPEG compression level, the reconstruction quality and generalization are still limited.

In this paper, we design a conditional flow-based framework, dubbed as Robust Invertible Image Steganography (RIIS), to alleviate the distortion influence and improve robustness. Inspired by conditional normalizing flow, we simultaneously model container image distribution and disposable high-frequency information to keep valuable secret information implicitly. As the flow model is bijective, our corresponding enhancement module and optimization strategy handle irreversible processes like channel reduction and quantization. The main contributions are listed as follows:

- To solve the substantial performance drop under distortion in former learning-based steganography methods, we proposed a general and robust framework RIIS for image steganography under diverse distortion (Gaussian noise, Poisson noise, and JPEG compression).

- We introduce the conditional flow into the steganography framework by regulating the high-frequency distribution conditional on the container image to implicitly preserve essential information for the revealing.

- We propose a distortion-guided modulation (DGM) over flow-based blocks to modulate the parameters for different distortion levels. The modulation makes it a general, controllable model for various types and distortion levels, with just one copy of parameters.

- Whether in a lossless or a distorted environment, abundant experiments demonstrate the superior robustness of our proposed RIIS while maintaining the imperceptibility and capacity of steganography. The robustness of RIIS has been proved successful in broader applications like real-world steganography, face-swap detection, and grayscale colorization.

2. Related Work

Image Steganography. Unlike cryptography, Steganography is designed to hide secret data into a host to produce an information container. As for the image steganography task, the host image acts as the cover of the secret image, which is confidential. The **hiding network** first hides the secret into the host image to produce a container. Next, the container image can be restored to secret and host image by the **revealing network** at the receiver end.

Traditionally, spatial-based [25, 37, 40, 44] methods ultilize the Least Significant Bits (LSB), pixel value differencing (PVD) [40], histogram shifting [48], multiplebit-planes [37] and palettes [25, 39] to hide images. They usaually arise statistical suspicion and vulnerable to steganalysis methods. Adaptive methods [32, 43] decomposed the steganography into embedding distortion minimization and data coding, which is indistinguishable by apperance but limited in capacity. Various transform-based schemes [12, 29] including JSteg [44] and DCT steganography [22] also fail to offer high payload capacity.

Various deep learning-based schemes have been produced to solve image steganography recently. Generative adversarial networks (GANs) [45] are introduced to synthesize container images. Probability map methods focus on generating various cost functions satisfying minimal-distortion embedding [43, 47]. [51] proposed a generator with U-Net architecture. [46] presents an adversarial scheme under the distortion minimization framework. Three-player game methods like SteganoGAN [56] and HiDDeN [59] learn information embedding and recovery by auto-encoder architecture to adversarially resist steganalysis. Deep Steganography [8] involved a fully convolutional network consisting of preparation, hiding, and revealing parts. The previous schemes reveal the potential of image steganography in digital communication, copyright protection, information certification, e-commerce, and many other practical fields [13].

Normalizing Flow-based Model. Normalizing flow [18, 19, 30] is a kind of powerful generative model with the advantage of straightforward calculation of likelihood. They are designed to learn a bijective mapping between the input domain and the target domain. The invertible neural network (INN), which involves the forward and backpropagation operations in the same network, is taken as the backbone of normalizing flow. Pioneering researches such as

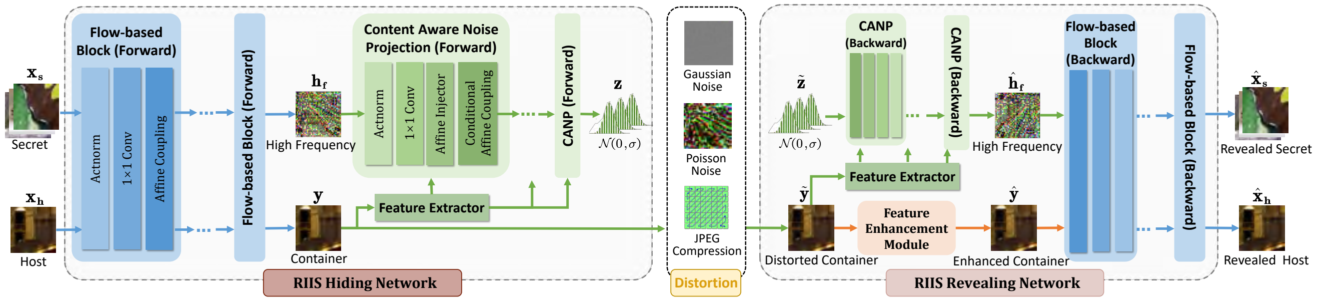

Figure 2. **Overview of our proposed RIIS stegenography framework.** The flow-based invertible blocks map the input pair $[\mathbf{x_s}, \mathbf{x_h}]$ into high frequency $\mathbf{h_f}$ and container image $\mathbf{y}$. The CANP project $\mathbf{h_f}$ to Gaussian-like noise $\mathbf{z}$ under the condition feature extracted from $\mathbf{y}$. Only $\mathbf{y}$ is transmitted by internet media and then receiver will get the distorted container image $\tilde{\mathbf{y}}$. Conversely, restored $\hat{\mathbf{h}}_\mathbf{f}$ along with enhanced container $\hat{\mathbf{y}}$ are taken into the reversed backward flow-based blocks, which is the same as the forward pass in parameters. Finally we get the revealed secret and host images $[\hat{\mathbf{x}}_\mathbf{s}, \hat{\mathbf{x}}_\mathbf{h}]$.

NICE [18] and RealNVP [19] mainly stress on the generation ability of flow-based model. In [20], a further explanation for the invertibility is explored. An unbiased flow-based generative model is introduced in [14]. Besides, glow [30] and i-RevNet [26] further improve coupling layers for density estimation achieving better generation results.

The theory of flow has recently attracted widespread attention in image processing, especially in low-level problems. Flow models have been proved to share the advantages in estimating the posterior of an inverse problem [6]. Recent flow-based models [3, 23, 34, 38] are capable of handling the image hiding and restoration problems. Due to the powerful representation ability, normalizing flow is also exploited in various inference tasks such as image rescaling [50], compression [53] and video super-resolution [60].

Although the existing steganography methods perform well in given application domains, they are not robust against distortion [6, 24]. There are also marvelous works about deblocking [54, 57] and denoising [16, 36]. However, it is impractical to apply these off-the-shelf methods into steganography tasks directly. The latest steganography method ISN [35] takes advantage of normalizing flow to perform probabilistic bijective construction for the steganography task. ISN [35] utilizes a single invertible network to hide and reveal images efficiently. HiNet [28] keeps almost the same architecture as ISN [35] despite the discrete wavelet transformation (DWT) channel squeezing. Flow-based methods [28, 35] show superiority over traditional schemes but severely rely on the reversibility of the framework. Since most steganography methods ignore the intermediate distortion, a subtle interference on the container usually causes a considerable drop in performance. Simply introducing distortion or simulation into model training can only handle a limited range of distortion and fail to produce satisfactory restoration and generalization.

3. Method

3.1. Overview

The major target of RIIS is to design a general and robust framework for image steganography under diverse distortion. It hides several secret images $\mathbf{x_s}$ into one informative container image $\mathbf{y}$, which is resistant to image distortion. For training stabilization, our framework directly learns the bijective mapping between secret images $\mathbf{x_s}$, host image $\mathbf{x_h}$ and container image $\mathbf{y}$ instead of explicitly modeling the distribution of inner latent. We mark the input as $\mathbf{x}$, composed of host $\mathbf{x_h}$ and secret image $\mathbf{x_s}$. The container image is capable of covering multiple images in it while keeping the appearance identical to the host image $\mathbf{x_h}$. Our robust model enables the receiver to restore revealed host $\hat{\mathbf{x}}_\mathbf{h}$ and secret images $\hat{\mathbf{x}}_\mathbf{s}$ from the distorted container image $\tilde{\mathbf{y}}$.

3.2. Flow-based Invertible Block

The flow-based network is naturally and intuitively suitable for the image steganography task because of its reversibility. The hiding and revealing procedure are ideally invertible with shared parameters and should be treated as the forward and backward processes of normalizing flow to enable end-to-end optimization. There are two main characteristics of the normalizing-flow model: the log-determinant of inference function $f_{\boldsymbol{\theta}}(\cdot)$ is simple to compute; the corresponding inverse function $f_{\boldsymbol{\theta}}^{-1}(\cdot)$ is tractable to solve.

In Fig. 2, we build up invertible blocks based on IRN [50]. For a input variable (*e.g.*, an image) $\mathbf{x}{=}[\mathbf{x_s}, \mathbf{x_h}]$ with distribution $\mathbf{x}{\sim}p(\mathbf{x})$ and a output variable $[\mathbf{h_f}, \mathbf{y}]$ with simple tractable distribution $[\mathbf{h_f}, \mathbf{y}]{\sim}p([\mathbf{h_f}, \mathbf{y}])$, flow models perform a bijective projection $f_{\boldsymbol{\theta}}$: $[\mathbf{h_f}, \mathbf{y}]{=}f_{\boldsymbol{\theta}}([\mathbf{x_s}, \mathbf{x_h}])$. Conversely, $\mathbf{x}$ can be recovered from $[\mathbf{h_f}, \mathbf{y}]$ by the inverse mapping $[\mathbf{x_s}, \mathbf{x_h}] = f_{\boldsymbol{\theta}}^{-1}([\mathbf{h_f}, \mathbf{y}])$. The input and output size of normalizing flow are exactly identical. $f_{\boldsymbol{\theta}}$ is composed of a series of invertible flow blocks: $f_{\boldsymbol{\theta}} =$

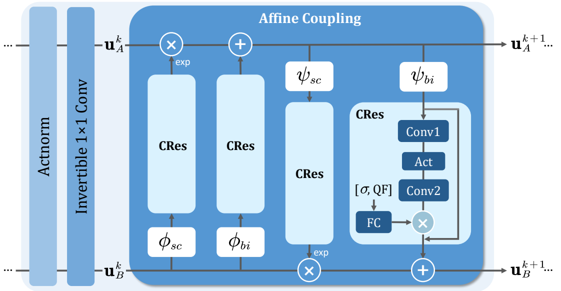

Figure 3. **The distortion-guided modulation (DGM) over flow-based invertible blocks.** The basic flow block is composed of act-norm, 1×1 conv, and affine coupling layer. ϕ_{bi}, ϕ_{sc}, ψ_{bi} and ψ_{sc} are built up with denseblocks to produce the affine-transformation factors (bias or scale). The CRes module takes the noise level σ or JPEG QF as an input condition feature to modulate the factors.

$f_{\boldsymbol{\theta}}^1 \circ f_{\boldsymbol{\theta}}^2 \circ \cdots \circ f_{\boldsymbol{\theta}}^N$. Concretely, block $f_{\boldsymbol{\theta}}^k$ for $k \in \{1, ..., N\}$ is composed of 1×1 convolution permuting, activation-normalization layer and affine coupling layer. The intermediate variables are defined as $\mathbf{u}^k = f_{\boldsymbol{\theta}}^k(\mathbf{u}^{k-1})$, where $\mathbf{u}^0$ represents $\mathbf{x}$ and $\mathbf{u}^N$ is output $[\mathbf{h_f}, \mathbf{y}]$.

In Fig. 3, $\mathbf{u}^k$ is split into $\mathbf{u}_A^k$ and $\mathbf{u}_B^k$ in the k-th block. Then, they will pass through affine couplings where ϕ and ψ constructed by denseblocks with ReLU activation that produce the scaling and bias on the $\mathbf{u}^k$:

$$\begin{aligned} \mathbf{u}_A^{k+1} &= \exp\left(\phi_{sc}\left(\mathbf{u}_B^k\right)\right) \odot \mathbf{u}_A^k + \phi_{bi}\left(\mathbf{u}_B^k\right), \\ \mathbf{u}_B^{k+1} &= \exp\left(\psi_{sc}\left(\mathbf{u}_A^{k+1}\right)\right) \odot \mathbf{u}_B^k + \psi_{bi}\left(\mathbf{u}_A^{k+1}\right). \end{aligned} \tag{1}$$

Obviously, the above affine coupling layer is mathematically invertible and has a triangular Jacobian matrix whose log-determinant is tractable to compute.

3.3. Content-Aware Noise Projection (CANP)

Earlier normalizing-flows always transform input into a target image and Gaussian noise distribution. However, due to the limited network depth and mapping ability of flow models, the direct target output of Gaussian distribution may lead to unsatisfactory results.

Inspired by the conditional flow [33], the high-frequency output $\mathbf{h_f}$ is assumed to rely on the container $\mathbf{y}$. Once trained, the forward process will squeeze the input host and secret images pair $[\mathbf{x_s}, \mathbf{x_h}]$ and transform it to container image $\mathbf{y}$ and high-frequency $\mathbf{h_f}$ as $p(\mathbf{x}|\mathbf{x_h}) \leftrightarrow p(\mathbf{y}, \mathbf{h_f}|\mathbf{x_h})$. $\mathbf{y}$ is constrained to approach host image $\mathbf{x_h}$ while containing information from $\mathbf{x_s}$. For the tacitness, the model is aimed to generate exactly the same container $\mathbf{y}$ as the input host image $\mathbf{x_h}$. The relation is presented as a Dirac delta function $\delta(\mathbf{x_h} - \mathbf{y})$ in Eq. (2):

$$\begin{aligned} p(\mathbf{y}|\mathbf{x_h})p(\mathbf{h_f}|\mathbf{y}, \mathbf{x_h}) &= \delta(\mathbf{x_h} - \mathbf{y})p(\mathbf{h_f}|\mathbf{y}) \\ &= \lim_{\boldsymbol{\Sigma} \to \mathbf{0}} \mathcal{N}(\mathbf{y}|\mathbf{x_h}, \boldsymbol{\Sigma})p(\mathbf{h_f}|\mathbf{y}), \end{aligned} \tag{2}$$

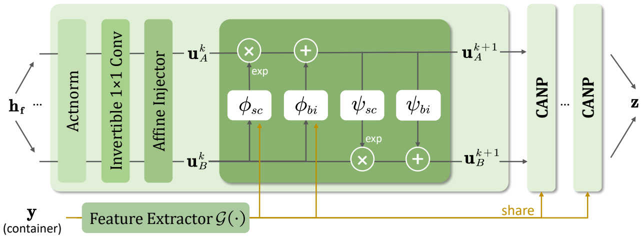

Figure 4. **Network architecture of the Content-Aware Noise Projection module based on conditional flow block.** It maps the high-frequency input part $\mathbf{h_f}$ to Gaussian distribution constrained $\mathbf{z}$ with the conditional feature extracted from containing $\mathbf{y}$.

$$p(\mathbf{x}|\mathbf{x_h}) \leftrightarrow \lim_{\boldsymbol{\Sigma} \to \mathbf{0}} \mathcal{N}(\mathbf{y}|\mathbf{x_h}, \boldsymbol{\Sigma})\mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbf{I}), \tag{3}$$

where the limit of Gaussian distribution is ultilized to model Dirac function. $\boldsymbol{\Sigma}$ is a covariance matrix where all diagonal elements are approximately zero. In this conditional flow, the dependent relation between high-frequency component $\mathbf{h_f}$ and container image $\mathbf{y}$ is deconstructed by the cascaded conditional mapping. Thus, $p(\mathbf{h_f}|\mathbf{y})$ is modeled as a Gaussian distribution $p(\mathbf{z}) \sim \mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbf{I})$ in Eq. (3).

This process is dubbed as content-aware noise projection(CANP) where $\mathbf{h_f}$ is projected to $\mathbf{z}$ with the conditional feature from $\mathbf{y}$. During the forward direction, CANP can transform the input image pair $[\mathbf{x_s}, \mathbf{x_h}]$ into container image $\mathbf{y}$ and nearly Gaussain-random variable $\mathbf{z}$. Given the container $\mathbf{y}$ and random samples $\tilde{\mathbf{z}}$ from $\mathcal{N}(\mathbf{0}, \mathbf{I})$, the bijective RIIS can generate $[\hat{\mathbf{x}}_\mathbf{s}, \hat{\mathbf{x}}_\mathbf{h}]$ in the backward pass. In Fig. 4, we build the CANP based on the conditional affine coupling that is powerful and still invertible. A branch of input dataflow $\mathbf{u}_B^k$ is merged with conditional features $\mathcal{G}(\mathbf{y})$ extracted from container image $\mathbf{y}$ and then be taken as the input of ϕ. Since there is permutation operation such as 1×1 convolution at the start of each flow block, the information in $\mathbf{u}_A$ and $\mathbf{u}_B$ are both affected by $\mathcal{G}(\mathbf{y})$ as:

$$\begin{aligned} \mathbf{u}_A^{k+1} &= \exp\left(\phi_{sc}\left(\mathbf{u}_B^k; \mathcal{G}(\mathbf{y})\right)\right) \odot \mathbf{u}_A^k + \phi_{bi}\left(\mathbf{u}_B^k; \mathcal{G}(\mathbf{y})\right) \\ \mathbf{u}_B^{k+1} &= \exp\left(\psi_{sc}\left(\mathbf{u}_A^{k+1}\right)\right) \odot \mathbf{u}_B^k + \psi_{bi}\left(\mathbf{u}_A^{k+1}\right). \end{aligned} \tag{4}$$

3.4. Container Enhancement Module (CEM)

A Container Enhancement Module (CEM) is utilized as the pre-processing module at the receiver end to eliminate the influence of image distortion like JPEG. In Fig. 2, We take a compromised step by integrating the CEM in the RIIS revealing network. The Batch-Normalization layer in the DnCNN [55] is removed for a lighter and suitable structure. Ahead of the flow blocks in the revealing network, the distorted image is firstly pre-processed by a simplified DnCNN network for denoising and JPEG deblocking. Given the distortion pattern, the CEM will process the container image to achieve a cleaner input for flow blocks.

3.5. End-to-End Optimization Strategy

Two-Stage Decoupled Tuning of Flow. To make the flow model adaptive to the distortion on the container image, we involve the decoupled training scheme into the latter half of our training process. In a flow-based model, inference forward and reverse passes are theoretically symmetrical and equal in parameters. However, there exist irreversible operations such as quantization in codecs, noise interference, and CEM network. These changes on intermediate container images require adjustment of the flow-based model. To a certain extent, the forward and backward parameters are proposed to be incompletely equal during the latter half stage of model training. The relaxation of parameters brings variance into the forward and backward passes. This strategy is named as Two-Stage Decoupled Tuning (**2DT**).

Loss Functions. It is required that the revealed host $\hat{\mathbf{x}}_\mathbf{h}$ and secret images $\hat{\mathbf{x}}_\mathbf{s}$ should be as close as possible to the input host $\mathbf{x_h}$ and secret $\mathbf{x_s}$. Here we employ the term $\mathcal{L}_{rev}$ to minimize the average distance among each pair of the restored and original images. The Container Enhancement Module (CEM) is meant to reconstruct clean container $\mathbf{y}$, from the distorted one $\tilde{\mathbf{y}}$ to restored $\hat{\mathbf{y}}$ with the term $\mathcal{L}_{CEM}$:

$$\mathcal{L}_{rev} = ||\mathbf{x}_s - \hat{\mathbf{x}}_\mathbf{s}||_2 + ||\mathbf{x_h} - \hat{\mathbf{x}}_\mathbf{h}||_2. \quad (5)$$

$$\mathcal{L}_{CEM} = ||\mathbf{y} - \tilde{\mathbf{y}}||_2. \quad (6)$$

$$\mathcal{L}_{distr} = \ell_{\mathcal{CE}}(p(\mathbf{z}), \mathcal{N}(\mathbf{0}, \mathbf{I})). \quad (7)$$

$$\mathcal{L}_{con} = ||\mathbf{x_h} - \mathbf{y}||_2 + ||\mathrm{FFT}(\mathbf{x_h}) - \mathrm{FFT}(\mathbf{y})||_2. \quad (8)$$

Concretely, since the distribution can be tractably depicted in flow-based models, CANP encourages the $p(\mathbf{z})$ to be independent from $p(\mathbf{h_f})$ and $p(\mathbf{y})$ and approximate to Gaussian distribution by distribution loss $\mathcal{L}_{distr}$ in Eq. (7). We depict the distribution distance by cross-entropy (CE) on $\mathbf{z}$. In order to guide the container image $\mathbf{y}$ to be approximately identical to the host image $\mathbf{x_h}$ both in spatial and frequency domain, we further apply fast fourier transform (FFT) [10] to extract frequency component in Eq. (8).

In summary, The total loss function in Eq. (9) considers the following four components: embedded image revealing, container invisibility, distortion enhancement, and noise distribution distance:

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{rev} + \lambda_2 \mathcal{L}_{con} + \lambda_3 \mathcal{L}_{CEM} + \lambda_4 \mathcal{L}_{distr}, \quad (9)$$

JPEG Simulation. The tolerance against JPEG compression is an essential concern in RIIS. JPEG pipeline consists of four main steps: color space transformation, discrete cosine transformation (DCT), quantization, and entropy encoding [42]. In fact, quantization is a lossy and non-differentiable step in JPEG compression. Thus, JPEG is not suitable for direct end-to-end optimization. To enable training over JPEG operations, a differentiable simulator module for JPEG compression is introduced in RIIS by replacing the quantization with fourier transformations(FT) [10].

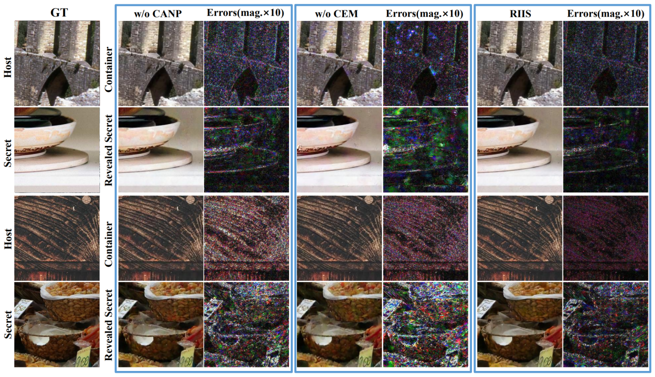

Figure 5. **Visual results of ablation study on CEM and CANP.** Container images here are distorted by the Gaussian noise ($\sigma = 10$). It reveals that the participation of CEM enhances the reconstruction and the CANP evidently adjusts the distribution.

3.6. Distortion-Guided Modulation (DGM)

It is not practical to train a specific network for every type and level of distortion. For general image steganography, we should make the RIIS parameter controllable with the strength of distortion. Here we propose a distortion-guided modulation (DGM) to control the affine coupling layer to handle container images corrupted with Gaussian noise or JPEG compression artifact. Concretely, given the distortion level (σ for the Gaussian noise and QF for the JPEG compression), the parameters in the DGM module will change with the distortion level through the affine transformation.

In Fig. 3, our DGM is constructed by deploying CResMD [21] into the affine coupling layer. Specifically, given the noise level or quality factor, the condition network produces the weight α to modulate the features by multiplying. The condition network is composed of several fully-connection layers. In this way, our method can handle various distortion levels by a single model. The unified framework built up with DGM is marked as **RIIS***.

4. Experimental Results

4.1. Implementation and Setup Details

Our proposed framework RIIS successfully maintains the payload capacity by hiding multiple images in one container image. RIIS is implemented with the NVIDIA Tesla V100 GPU for acceleration. We implement the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.99$. The learning rate is set to be 0.0001, and the batch size is set to be 16 for training. The dataset for training and testing is DIV2K [4], if not specified. For the loss, the corresponding weight factors are $\lambda_1 = 1$, $\lambda_2 = 16$, $\lambda_3 = 1$ and $\lambda_4 = 0.5$. The PSNR (Peak Signal to Noise Ratio) metric is utilized to evaluate the performance.

Table 1. Ablation studies for every model design, including 2DT, CANP and CEM in RIIS under various distortion. The results are evaluated on the pair of revealed secret and original secret images, by PSNR metric. It proves that all the modules make sense in our robust framework, among which the CANP is the most indispensable.

2DT	CANP	CEM	Gaussian $\sigma = 10$	Poisson Noise	JPEG Q=40	JPEG Q=80	JPEG Q=90	Clean
×	✓	✓	27.61	27.42	26.77	27.55	27.88	42.91
✓	×	✓	27.38	27.35	26.40	27.28	27.72	42.74
✓	✓	×	27.82	27.62	26.64	27.46	27.96	43.21
✓	✓	✓	28.08	28.01	27.32	28.25	28.71	44.19

Table 2. The reveal secret image PSNR of different schemes to producing $\mathbf{h_f}$ input during the backward pass. The experiments are evaluated under noise $\sigma = 10$ to hide one secret into a container image. The CANP is proved to be the best mapping of $\mathbf{h_f}$.

$\mathbf{h_f}$ Source	$\tilde{\mathbf{z}} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$	Copy from $\mathbf{y}$	Resblock from $\mathbf{y}$	CANP from $\tilde{\mathbf{z}}$ (Ours)
Secret	42.53	43.18	43.56	44.19

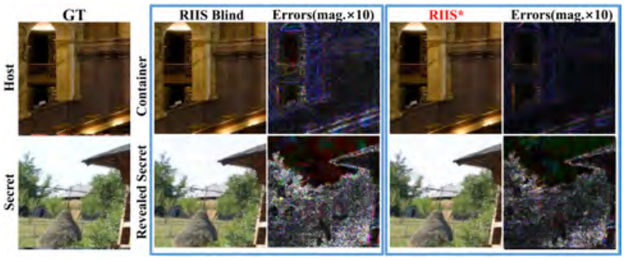

Figure 6. **Visual results of ablation study on DGM.** It reveals that the parameters in the unified framework RIIS* are efficiently modulated by DGM, evidently better than RIIS-Blind.

4.2. Ablation Study

Effect of CANP. When we assume the high-frequency component $\mathbf{h_f}$ is independent with the container image $\mathbf{y}$ and like IRN [50], the performance remains limited as it fails to keep an informative prior for image steganography. Experiments in Tab. 2 show that relating $\mathbf{h_f}$ with the container image $\mathbf{y}$ as conditional prior effectively improves revealing quality. The simplest modulation of backward input $\mathbf{h_f}$ is the copy of container image $\mathbf{y}$, which does not contain any useful information for revealing except for $\mathbf{y}$. We also implement a residual block to extract feature from the container $\mathbf{y}$ as input to produce $\mathbf{h_f}$. The performance growth from these two simple types of modulation over $\mathbf{h_f}$ proves the effectiveness of condition input from container $\mathbf{y}$.

The CANP is proposed due to the discovery that the high-frequency component $\mathbf{h_f}$ is highly dependent on the container (low-frequency component $\mathbf{y}$). Tab. 1 and Fig. 5 show that the conditional mechanism CANP efficiently models the relation between high-frequency $\mathbf{h_f}$ and low-frequency components $\mathbf{y}$. Though half the forward output $\mathbf{y}$ is evacuated, the high-frequency part of the input image pair is implictly stored. When the Gaussian sampling $\tilde{\mathbf{z}}$ is mapped into $\hat{\mathbf{h}}_\mathbf{f}$ with the condition of $\mathbf{y}$ in CANP, we get a approximate reconstruction $\hat{\mathbf{h}}_\mathbf{f}$ of original signal.

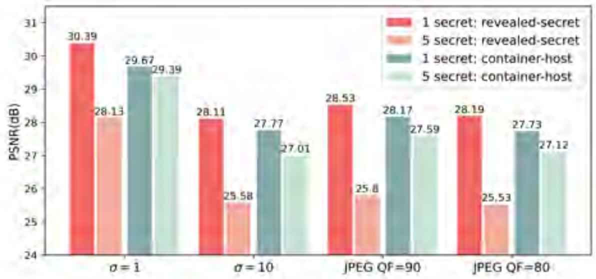

Figure 7. **The capacity performance for hiding multiple or single secret images into one container under distortion on ImageNet [17]** (1000 random samples). With the increse of secret number, The reconstruction quality drops but still maintains acceptable fidelity.

Effect of 2DT. We involve two-stage decoupled tunning into the flow-based model to adapt it against distortion and irreversible operations. According to our experimental result in Tab. 1, the decoupling evidently improves the reconstruction performance. After end-to-end training, our decoupled model learns to mitigate the loss of quantization and noise interference.

Effect of CEM. The CEM employs the DnCNN-like network to perform pre-processing over the distorted container image $\tilde{\mathbf{y}}$ to eliminate the effect of distortion. Results in Tab. 1 and Fig. 5 show that the participation of CEM plays an indispensable role in the total robust steganography.

4.3. Comparison with SOTA

In the image steganography task, the most common concern is the fidelity of two pairs: revealed secret $\hat{\mathbf{x}}_\mathbf{s}$ and origin $\mathbf{x_s}$, container $\mathbf{y}$ and host image $\mathbf{x_h}$. For the comparison with the latest method, we reproduce the State-of-the-art ISN [35] and reach the performance itself claimed on the DIV2K [5].

Container Image under distortion. Our model mainly focuses on the image steganography with the container image under various distortion. In Fig. 8, even under slight interference on container image, the secret restoration of HiNet [28] witnesses a substantial drop in performance. It shows that the previous methods ignorant of distortion are vulnerable and fragile, limiting their application in practice. Since the performance of the original ISN model ignorant of distortion [35] is too poor, we finetuned the ISN network

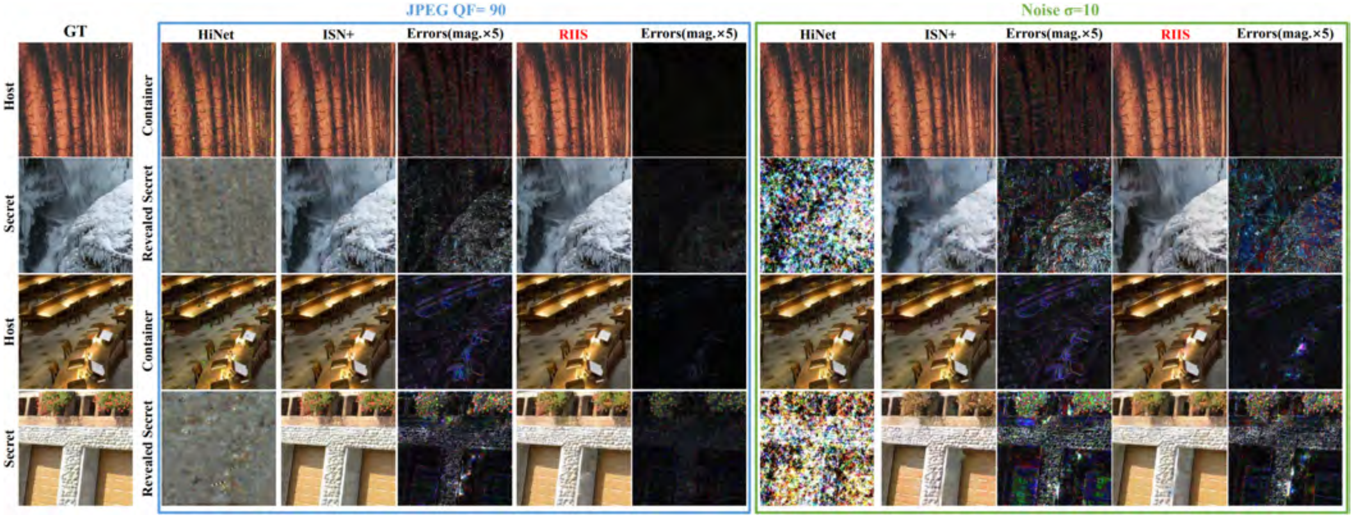

Figure 8. **Visual comparison of latest HiNet [28], finetuned ISN$^+$ [35] and our RIIS under the same JPEG QF=90 (blue border) or Gaussian noise $\sigma = 10$ (green border).** Under both distortion, the left-most column shows the failure of the latest HiNet [28], especially the revealed secret image. The reconstruction quality of our RIIS shows substantial superiority compared with the latest ISN$^+$ [35].

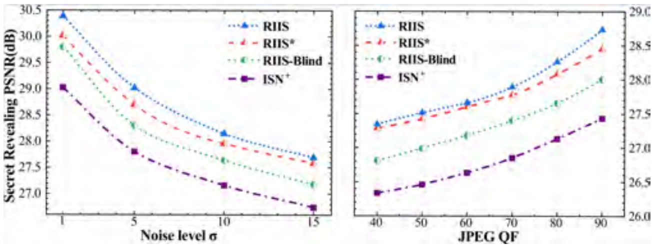

Figure 9. **The PSNR curves with different Gaussian noise σ and JPEG QF in our experiment settings.** RIIS* with distortion-guided (DGM) modulation achieves a subtle performance gap with RIIS and only requires one controllable network for all distortion. The improvement from RIIS-Blind to RIIS* proves that the DGM successfully adjust RIIS for different distortion levels.

separately for every distortion level in our experiment settings as the baseline. In contrast to the original ISN model, we name the finetuned ISN as **ISN**$^+$. The ISN$^+$ is also finetuned for every specific noise or compression level, but it still fails to offer satisfactory performance. Despite the variety of colors and structures of the images, RIIS can restore them with no viewable artifacts. The performance of hiding images with the container image stained by noise or JPEG compression is shown in Tab. 3. The results reveal that our proposed method RIIS successfully maintains higher reconstruction quality compared with the latest methods.

To prove the payload capacity of our method, we increase the channel of RIIS for hiding multiple secret images into one container. Fig. 7 shows the model performance for hiding single or multiple secret images into one container under different distortion. Since RIIS is the first model to hide multiple images under distortion, no other previous method is available for comparison.

Table 3. Comparison of secret image restoration quality when container image is under diverse distortion. Our RIIS is the highest under every distortion. The unified model RIIS* also shows evident performance superiority compared with previous methods.

Method	Gaussian Noise		Poisson Noise	JPEG	
	$\sigma = 10$	$\sigma = 1$		QF=40	QF=90
HiNet [28]	9.98	26.93	21.23	11.52	12.59
ISN [35]	8.55	25.19	19.38	10.11	11.25
ISN$^+$ [35]	27.12	28.98	26.71	26.25	27.48
RIIS*	28.03	30.01	27.23	27.18	28.44
RIIS	28.22	30.32	27.47	27.32	28.71

Discussion on unified RIIS* with DGM. We introduce distortion-guided modulation (DGM) to make network parameters vary with different distortion levels. The unified framework for all distortion built up with DGM is marked as **RIIS***. DGM allows RIIS to handle all the distortion levels with the shared base parameters. We also evaluate the RIIS-Blind model without DGM as the baseline, also trained under random types and levels of distortion. In Fig. 9, there is only a subtle gap between the unified RIIS* with DGM and separately tuned RIIS. In terms of the unified network for all distortion, the DGM gains substantial performance growth compared with RIIS-Blind in Fig. 6. The DGM scheme makes RIIS the first general steganography framework applicable under various distortions in practice.

Container Image without distortion. Comparison tests with the latest steganography methods [28, 35] with clean container images are conducted here. The results for hiding 1 or 5 secret images into a container image are numerically compared in Tab. 4. Our method shows superior perfor-

Table 4. Steganography performance comparison for hiding 5 or 1 secret images in a container on ImageNet [17]. Our RIIS shows the best performance under both circumstances.

Method	Multi Secrets		Single Secret	
	Container	Secret	Container	Secret
AutoEncoder [49]	32.35	31.21	-	-
ISN [35]	33.77	36.02	42.53	43.58
IICNet [15]	35.64	37.94	-	-
HiNet [28]	-	-	44.16	46.48
RIIS	35.92	38.13	43.97	46.71

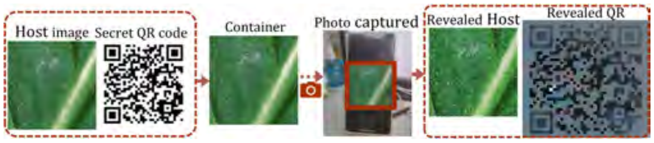

Figure 10. **Real-world steganography to take photo as container.** The right-most QR-code is the example of the secret revealed from the container photo, which can still be scanned out. It can be used for hiding messages on the screen or presswork.

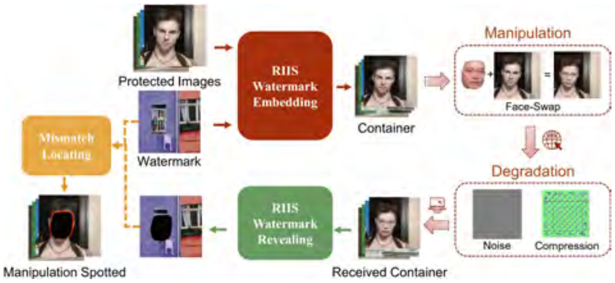

Figure 11. **Process of and face-swap detection.** The public watermark image is hidden into the container image by RIIS to protect a series of images. When a face-swap is deployed on the container, the revealed watermark will mismatch with the public-distributed watermark copy, which helps to locate manipulation.

mance compared with the latest methods when there is no distortion on the container image. As the HiNet [28] ignores multiple images steganography, it is not listed. The average PSNR for our restoration of images is evidently higher than IICNet [15], HiNet [28] and ISN [35]. The results show that our method achieves better performance even when hiding 5 images into one container, demonstrating the superior capacity and generality of our method.

4.4. Applications Derived from Robustness

Hiding secret in the real world. If we put the container image on display or print it on paper and capture it by a CMOS sensor, it suffers from transformation, sensor noise, motion blur, etc. In Fig. 10, our method can even reveal the secret from photos due to incredible robustness. This would implicitly bridge the cyberspace and real-world vision, which is potential in the construction of the meta-verse industry. It also makes sense in the protection of copyright and integrity of digital assets and artworks.

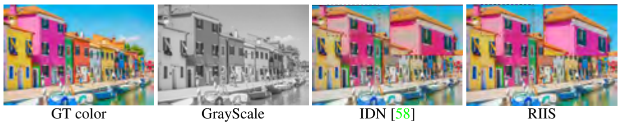

Figure 12. **Visual comparison of RIIS and IDN [58] with grayscale image as container, under JPEG compression QF = 80.** Notice the areas like sky and wall, the compression significantly harms the restored color image by IDN [58]. In contrast, the image produced by RIIS remains vivid color and fidelity due to reliable robustness.

Face-Swap Detection. Fig. 11 demonstrates our scheme of face-swap detection. On receiving the attacked version of the protected image produced by the RIIS hiding network, the watermark will be extracted, and the feature matching operation [9] is conducted between the original and revealed watermark to determine where the manipulation is. Due to the robustness of RIIS, the detection is effective and accurate under compression. We also extend our work to detect stretching and trimming.

Invertible Colorization. Since our framework can efficiently embed multiple images into a single container image, the YUV-channel color image can be embeded into a single grayscale channel in the same way. In the backward pass, RIIS reconstructs the color image from the grayscale container. Previous flow-based SOTA colorization method IDN [58] highly relies on the distribution of the synthetic grayscale. IDN also declares it still suffers from the JPEG compression on grayscale container images. Fig. 12 shows our superiority under the usual lossy compression situation. Our robustness against distortion addresses the application problem of image colorization in practice.

5. Conclusion and Discussion

The image steganography tasks are challenging when a great capacity of secret images need to be hidden in a limited size of a container image, especially under noise or JPEG interference. In this paper, we present a general and novel robust invertible image steganography (RIIS) framework, where the proposed CANP and CEM module, along with a well-designed training strategy are leveraged to prevent the container image during steganography from distortion such as Poisson noise, Gaussian noise, and JPEG compression. Experiments prove that our model design guarantees us the highest performance. The improvement of steganography robustness significantly broadens the application of information steganography in real-world applications. The efficiency of our model design is also proved on other low-level inverse problems like decolorization. Our future work will support RIIS on MindSpore [1], which is a new deep learning computing framework.

References

[1] Mindspore. https://www.mindspore.cn/, 2020. 8

[2] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4432–4441, 2019. 2

[3] Abdelrahman Abdelhamed, Marcus A Brubaker, and Michael S Brown. Noise flow: Noise modeling with conditional normalizing flows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3165–3173, 2019. 3

[4] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 126–135, 2017. 5

[5] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *IEEE Conference on International Conference on Computer Vision Workshops (CVPRW)*, 2017. 6

[6] Lynton Ardizzone, Jakob Kruse, Carsten Rother, and Ullrich Köthe. Analyzing inverse problems with invertible neural networks. In *International Conference on Learning Representations (ICLR)*, 2018. 3

[7] Shumeet Baluja. Hiding images in plain sight: Deep steganography. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 1

[8] Shumeet Baluja. Hiding images within images. *TPAMI*, 2019. 1, 2

[9] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *European Conference on Computer Vision (ECCV)*, 2006. 8

[10] E Oran Brigham. *The fast Fourier transform and its applications*. 1988. 5

[11] Chi-Kwong Chan and Lee-Ming Cheng. Hiding data in images by simple lsb substitution. *Pattern recognition*, 37(3):469–474, 2004. 1

[12] Yambem Jina Chanu, Kh Manglem Singh, and Themrichon Tuithung. Image steganography and steganalysis: A survey. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2012. 1, 2

[13] Abbas Cheddad, Joan Condell, Kevin Curran, and Paul Mc Kevitt. Digital image steganography: Survey and analysis of current methods. *IEEE Transactions on Signal Processing (TSP)*, 2010. 2

[14] Ricky TQ Chen, Jens Behrmann, David K Duvenaud, and Joern-Henrik Jacobsen. Residual flows for invertible generative modeling. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 3

[15] Ka Leong Cheng, Yueqi Xie, and Qifeng Chen. Iicnet: A generic framework for reversible image conversion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 8

[16] Mou Chong, Zhang Jian, Fan Xiaopeng, Liu Hangfan, and Wang Ronggang. Cola-net: Collaborative attention network for image restoration. *IEEE Transactions on Multimedia (TMM)*, 2021. 3

[17] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 6, 8

[18] Laurent Dinh, David Krueger, and Yoshua Bengio. Nice: Non-linear independent components estimation. In *International Conference on Learning Representations (ICLR)*, 2014. 2, 3

[19] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. In *International Conference on Learning Representations (ICLR)*, 2016. 2, 3

[20] Anna C Gilbert, Yi Zhang, Kibok Lee, Yuting Zhang, and Honglak Lee. Towards understanding the invertibility of convolutional neural networks. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2017. 3

[21] Jingwen He, Chao Dong, and Yu Qiao. Interactive multi-dimension modulation with dynamic controllable residual learning for image restoration. In *European Conference on Computer Vision (ECCV)*, 2020. 5

[22] Stefan Hetzl and Petra Mutzel. A graph–theoretic approach to steganography. In *IFIP International Conference on Communications and Multimedia Security*, 2005. 2

[23] Jonathan Ho, Xi Chen, Aravind Srinivas, Yan Duan, and Pieter Abbeel. Flow++: Improving flow-based generative models with variational dequantization and architecture design. In *International Conference on Machine Learning (ICML)*, 2019. 3

[24] Chin-Wei Huang, David Krueger, Alexandre Lacoste, and Aaron Courville. Neural autoregressive flows. In *International Conference on Machine Learning (ICML)*, 2018. 3

[25] Shoko Imaizumi and Kei Ozawa. Multibit embedding algorithm for steganography of palette-based images. In *Pacific-Rim Symposium on Image and Video Technology (PSIVT)*, pages 99–110. Springer, 2013. 2

[26] Jörn-Henrik Jacobsen, Arnold Smeulders, and Edouard Oyallon. i-revnet: Deep invertible networks. In *International Conference on Learning Representations (ICLR)*, 2018. 3

[27] Priyank Jaini, Kira A Selby, and Yaoliang Yu. Sum-of-squares polynomial flow. In *International Conference on Machine Learning (ICML)*, 2019. 2

[28] Junpeng Jing, Xin Deng, Mai Xu, Jianyi Wang, and Zhenyu Guan. Hinet: Deep image hiding by invertible network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 3, 6, 7, 8

[29] Inas Jawad Kadhim, Prashan Premaratne, Peter James Vial, and Brendan Halloran. Comprehensive survey of image steganography: Techniques, evaluations, and trends in future research. *Neurocomputing*, 2019. 1, 2

[30] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 2, 3

[31] Durk P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improved variational inference with inverse autoregressive flow. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016. 2

[32] Bin Li, Ming Wang, Jiwu Huang, and Xiaolong Li. A new cost function for spatial image steganography. In *IEEE International Conference on Image Processing (ICIP)*, 2014. 2

[33] Jingyun Liang, Andreas Lugmayr, Kai Zhang, Martin Danelljan, Luc Van Gool, and Radu Timofte. Hierarchical conditional flow: A unified framework for image super-resolution and image rescaling. In *IEEE Conference on International Conference on Computer Vision*, 2021. 4

[34] Jingyun Liang, Kai Zhang, Shuhang Gu, Luc Van Gool, and Radu Timofte. Flow-based kernel prior with application to blind super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[35] Shao-Ping Lu, Rong Wang, Tao Zhong, and Paul L Rosin. Large-capacity image steganography based on invertible neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 2, 3, 6, 7, 8

[36] Chong Mou, Jian Zhang, and Zhuoyuan Wu. Dynamic attentive graph learning for image restoration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4328–4337, 2021. 3

[37] Bui Cong Nguyen, Sang Moon Yoon, and Heung-Kyu Lee. Multi bit plane image steganography. In *International Workshop on Digital Watermarking*, 2006. 2

[38] Didrik Nielsen, Priyank Jaini, Emiel Hoogeboom, Ole Winther, and Max Welling. Survae flows: Surjections to bridge the gap between vaes and flows. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 3

[39] Michiharu Niimi, Hideki Noda, Eiji Kawaguchi, and Richard O Eason. High capacity and secure digital steganography to palette-based images. In *IEEE International Conference on Image Processing (ICIP)*, 2002. 2

[40] Feng Pan, Jun Li, and Xiaoyuan Yang. Image steganography method based on pvd and modulus function. In *International Conference on Electronics, Communications and Control(ICECC)*, 2011. 2

[41] George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for density estimation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 2

[42] William B Pennebaker and Joan L Mitchell. *JPEG: Still image data compression standard*. 1992. 5

[43] Tomáš Pevnỳ, Tomáš Filler, and Patrick Bas. Using high-dimensional image models to perform highly undetectable steganography. In *International Workshop on Information Hiding (IHIP)*, 2010. 1, 2

[44] Niels Provos and Peter Honeyman. Hide and seek: An introduction to steganography. *IEEE Symposium on Security and Privacy*, 2003. 2

[45] Haichao Shi, Jing Dong, Wei Wang, Yinlong Qian, and Xiaoyu Zhang. Ssgan: secure steganography based on generative adversarial networks. In *Pacific Rim Conference on Multimedia*, 2017. 2

[46] Weixuan Tang, Bin Li, Shunquan Tan, Mauro Barni, and Jiwu Huang. Cnn-based adversarial embedding for image steganography. *IEEE Transactions on Information Forensics and Security (TIFS)*, 2019. 2

[47] Weixuan Tang, Shunquan Tan, Bin Li, and Jiwu Huang. Automatic steganographic distortion learning using a generative adversarial network. *IEEE Signal Processing Letters (SPL)*, 2017. 2

[48] Piyu Tsai, Yu-Chen Hu, and Hsiu-Lien Yeh. Reversible image hiding scheme using predictive coding and histogram shifting. *Signal Processing*, 2009. 2

[49] Menghan Xia, Xueting Liu, and Tien-Tsin Wong. Invertible grayscale. *TOG*, 2018. 8

[50] Mingqing Xiao, Shuxin Zheng, Chang Liu, Yaolong Wang, Di He, Guolin Ke, Jiang Bian, Zhouchen Lin, and Tie-Yan Liu. Invertible image rescaling. In *European Conference on Computer Vision (ECCV)*, 2020. 2, 3, 6

[51] Jianhua Yang, Danyang Ruan, Jiwu Huang, Xiangui Kang, and Yun-Qing Shi. An embedding cost learning framework using gan. *IEEE Transactions on Information Forensics and Security (TIFS)*, 2019. 2

[52] Jian Zhang Youmin Xu. Expressive and compressive gan inversion network. In *IEEE International Conference on Image Processing (ICIP)*, 2021. 2

[53] Jian Zhang Youmin Xu. Invertible resampling-based layered image compression. In *Data Compression Conference (DCC)*, 2021. 3

[54] Jian Zhang, Ruiqin Xiong, Chen Zhao, Yongbing Zhang, Siwei Ma, and Wen Gao. Concolor: Constrained non-convex low-rank model for image deblocking. *IEEE Transactions on Image Processing (TIP)*, 25(3):1246–1259, 2016. 3

[55] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE Transactions on Image Processing (TIP)*, 2017. 4

[56] Kevin Alex Zhang, Alfredo Cuesta-Infante, Lei Xu, and Kalyan Veeramachaneni. Steganogan: High capacity image steganography with gans. *arXiv:1901.03892*, 2019. 2

[57] Chen Zhao, Jian Zhang, Siwei Ma, Xiaopeng Fan, Yongbing Zhang, and Wen Gao. Reducing image compression artifacts by structural sparse representation and quantization constraint prior. *IEEE Transactions on Circuits and Systems for Video Technology(TCSVT)*, 27(10):2057–2071, 2016. 3

[58] Rui Zhao, Tianshan Liu, Jun Xiao, Daniel PK Lun, and Kin-Man Lam. Invertible image decolorization. *IEEE Transactions on Image Processing (TIP)*, 2021. 8

[59] Jiren Zhu, Russell Kaplan, Justin Johnson, and Li Fei-Fei. Hidden: Hiding data with deep networks. In *European Conference on Computer Vision (ECCV)*, 2018. 1, 2

[60] Xiaobin Zhu, Zhuangzi Li, Xiao-Yu Zhang, Changsheng Li, Yaqi Liu, and Ziyu Xue. Residual invertible spatio-temporal network for video super-resolution. In *Proceedings of the AAAI conference on artificial intelligence (AAAI)*, 2019. 3