Robust Audio Watermarking Based on Invertible Neural Network

Jiji Zhu*

School of Information Science and Engineering, Chongqing Jiaotong University, Chongqing, China
**Corresponding author: zhujiji200007@163.com*

Abstract: *Audio watermarking technology leverages the characteristics of the human auditory system and the original audio carrier to imperceptibly embed watermark information into the audio. Traditional watermarking algorithms employ signal processing techniques and are limited by the experience of the model designer. In contrast, deep learning–based neural network audio watermarking algorithms offer greater adaptability and versatility, and their robustness can be enhanced through simulated attacks, marking an important direction for future development in audio watermarking technology. Related research primarily focuses on balancing the imperceptibility, robustness, and embedding capacity of watermark information. The audio watermarking model designed in this paper emphasizes imperceptibility and robustness. Imperceptibility is enhanced by designing a discriminator that ensures the human ear cannot distinguish between the original and watermarked audio. Robustness is improved by developing a simulated attack block, which provides strong resistance against multiple types of attacks, and by mitigating the damage caused by the attack layer through the invertible design of a neural network assisted by a balancing block. This study achieves high imperceptibility and strong robustness based on an invertible neural network. The experimental results demonstrate that the model performs well in terms of both watermark embedding and extraction accuracy, as well as anti-attack performance.*

Keywords: *Audio Watermarking, Invertible Neural Network, Imperceptibility, Balance Block*

1. Introduction

In recent years, advancements in the Internet and multimedia technologies have significantly facilitated our lives by enabling the easier transmission of digital audio. However, this progress introduces new challenges for data copyright protection. Digital audio watermarking is a technique that imperceptibly embeds watermark information into audio data to ensure secure transmission. By extracting the watermark at the receiving end, the authenticity of the audio data can be verified [1]. The extracted watermark information serves to establish copyright ownership in the event of disputes.

Embedding watermarks in digital audio signals is more challenging compared to digital images because the human auditory system is more sensitive than the visual system, making embedded watermark messages more easily detectable [2]. Therefore, research on digital audio watermarking is highly significant. Imperceptibility and robustness are the two most critical metrics of digital watermarking systems. Imperceptibility, a key feature of audio watermarking, requires that the added watermarked audio signal remains undetectable to the human ear. Achieving this is difficult due to significant individual differences in auditory perception and the advancements in digital technologies, data storage devices, and high-quality audio systems, which facilitate the detection of watermarking information. Robustness in an audio watermarking system generally requires that the watermark-containing signal can be accurately extracted from the watermarked message even after corruption or malicious attacks. Typically, there is a trade-off between robustness and imperceptibility; the imperceptibility of a system may need to be compromised to enhance its robustness, thereby ensuring its effectiveness, especially against sophisticated attacks. [3]

Audio watermarking algorithms can be categorized into time-domain and frequency-domain methods based on their watermarking domain. Time-domain algorithms are characterized by their straightforward implementation, achieved by directly modifying the audio data; however, they typically lack robustness against processing attacks on audio signals. In contrast, frequency-domain algorithms embed watermarks by altering the frequency-domain coefficients of the audio. Research in this area generally utilizes techniques such as the Discrete Cosine Transform (DCT), Discrete Wavelet

Transform (DWT), and Discrete Fourier Transform (DFT). Traditional audio watermarking techniques often rely on the expert knowledge of designers, imposing high demands on the designer and introducing subjective limitations in the designed audio algorithms. Conversely, the rapid advancement of deep learning in recent years offers novel solutions for digital watermarking techniques [4][5]. In these approaches, end-to-end watermarking models perform embedding and extraction operations during each training iteration, constrain the imperceptibility and completeness of the watermarking process through carefully designed training objectives, and enhance robustness by integrating a simulated attack layer.

In this study, we propose a robust audio watermarking model based on invertible neural networks. This model combines locator codes with watermarked messages, enabling embedding and extraction via invertible neural networks. Additionally, it incorporates a discriminator, a simulated attack layer, and a balancing block to enhance both imperceptibility and robustness of the watermarking model.

Our main contributions are summarized as follows:

(1) We design a combination of locator codes and watermark messages to achieve embedding and localized extraction of watermark information, optimizing the localization capability through a localization loss function.

(2) We introduce a discriminator to enhance the imperceptibility of audio watermarking by training it to differentiate between original and watermarked audio.

(3) The proposed watermarking model improves robustness and imperceptibility while maintaining a sufficient embedding capacity.

2. Related Work

2.1 Audio Watermarking

Digital watermarking technology is an effective method for digital multimedia copyright protection and content authentication. It leverages redundancies inherent in multimedia data, such as those found in images and audio, and employs specific time-domain or frequency-domain algorithms to embed watermark information into the multimedia content. In audio digital watermarking, the technology utilizes audio signal redundancies and the masking effects of human auditory perception to covertly embed digital information into audio media. This process ensures that the auditory quality of the audio carrier remains unaffected while enabling covert transmission of information, copyright protection, content authentication, tracking, and monitoring. Effective audio digital watermarking must guarantee that the embedded watermark does not degrade listening quality or interfere with the normal use of the audio. Additionally, it must ensure that the watermark can be accurately extracted even after the audio containing the watermark has been subjected to attacks or modifications.

Audio watermarking algorithms can be categorized into time-domain and frequency-domain methods based on their embedding domain. Time-domain audio watermarking algorithms include techniques such as Least Significant Bit (LSB) substitution and echo concealment. These algorithms are straightforward to implement, as they directly modify audio data values. However, they generally lack robustness against interference and processing attacks on audio signals. Watermarks embedded using time-domain methods are susceptible to modifications and tampering, making them more suitable for fragile watermarking applications, such as audio integrity verification, where the watermark indicates whether the audio has been altered.

In contrast, frequency-domain audio watermarking algorithms, which encompass methods such as the Discrete Wavelet Transform (DWT), the Discrete Cosine Transform (DCT), and Singular Value Decomposition (SVD), offer greater resistance to interference. By modifying frequency-domain coefficients, these algorithms render the watermark less perceptible and more difficult to detect, thereby expanding their range of applications.

In recent years, deep learning has been increasingly applied to watermarking[6][7][8], demonstrating superior performance compared to traditional methods, particularly in terms of imperceptibility and robustness. Watermarking models developed in related studies typically adopt an encoder-decoder architecture. Furthermore, the introduction of an attack layer has evolved these models into an encoder-attack layer-decoder structure, significantly enhancing the robustness of audio watermarking systems. However, most current audio watermarking models insufficiently address the localization of watermarks, complicating the extraction process. Some literature suggests using combined pattern bits and payloads to construct encoded messages as a means to resolve the watermark localization problem. This approach effectively relies on brute-force detection methods to identify the watermark within the audio content.

2.2 Invertible Neural Network

Invertible neural networks are the first learning-based framework for modeling complex high-dimensional densities of canonical flows, proposed by Dinh et al.[9] in 2014. The affine coupling layer[10] is the basic building block of a invertible neural network (INN). The encoding and decoding processes share the same parameters, making the model lightweight. Since invertible network is theoretically information-lossless, it can retain as much detail as possible about the inputs. Due to these remarkable properties, many works with invertible architectures achieve more satisfactory performance than traditional autoencoder frameworks[11]. They focus on learning the forward process, using additional potential output variables to capture information that would otherwise be lost, unlike classical neural networks that attempt to solve fuzzy inverse problems directly. While autoencoders are very capable of selecting important information for reconstruction, a certain amount of information is lost altogether; whereas encoding and decoding using an INN helps to retain the information. The specific features of an INN are that it has a bijective mapping between the inputs and outputs and the existence of its inverse mapping; that it can efficiently compute both forward and inverse mappings; and that both mappings have an easy-to-handle Jacobian matrix, which makes it possible to compute a posteriori probabilities explicitly[12]. Due to its flexibility and effectiveness, an INN has also been used for image super-resolution[13] and video super-resolution[14]. INN is also used for image-to-video synthesis, image compression, and image denoising[15]. Although INN has great potential for information embedding and extraction, it is less robust to lossy data compression and other distortions, which are key issues in digital watermarking.

3. Model Architecture

Our model architecture enhances the model proposed in [6], as illustrated in Figure 1. The architecture adopts an embedder-attack layer-extractor structure, where both the embedder and extractor are implemented using invertible neural networks. During the embedding process, the objective is to incorporate the watermark message into the original audio while maintaining the imperceptibility and robustness of the watermark. We integrate the localization code with the watermarked message for embedding, convert the original time-domain audio to the frequency domain using the Short-Time Fourier Transform (STFT), and simultaneously expand the watermarked message to the desired dimension in the frequency domain through a linear transformation. Subsequently, we apply STFT to the extended watermarked message to obtain its frequency domain representation. The watermark information is embedded into the audio spectrum within the invertible neural network block. Finally, the spectrum after embedding the watermark is converted back to the time domain using the Inverse Short-Time Fourier Transform (ISTFT) to produce the watermarked audio. The block processes the data as depicted in Equation 1 and Equation 2.

$$x^{i+1} = x^{i} \odot \exp(\alpha(\psi(m^{i}))) + \phi(m^{i}) \tag{1}$$

$$m^{i+1} = m^{i} \odot \exp(\alpha(\rho(x^{i+1}))) + \eta(x^{i+1}) \tag{2}$$

Where x denotes the audio data input to the invertible neural network and the processed audio data, and mmm denotes the watermarked information (including the localization code and watermarked message) input to the invertible neural network. $\alpha(.)$is a sigmoid function, and $\psi(.)$, $\phi(.)$, $\rho(.)$, and $\eta(.)$ are sub-networks composed of dense blocks. Dense blocks are the basic building blocks in invertible neural networks, which can effectively utilize features to enhance the expressive ability and information transfer efficiency of the network. At its core, it is capable of feature fusion and transfer, and the input of each layer includes not only the output of the previous layer but also the outputs of all previous layers to achieve feature fusion. The LeakyReLU activation function is applied after each convolutional layer to enhance the nonlinear expressiveness of the network, and the last layer of convolution maps all the fused features to the number of output channels to ensure the consistency of the data flow.

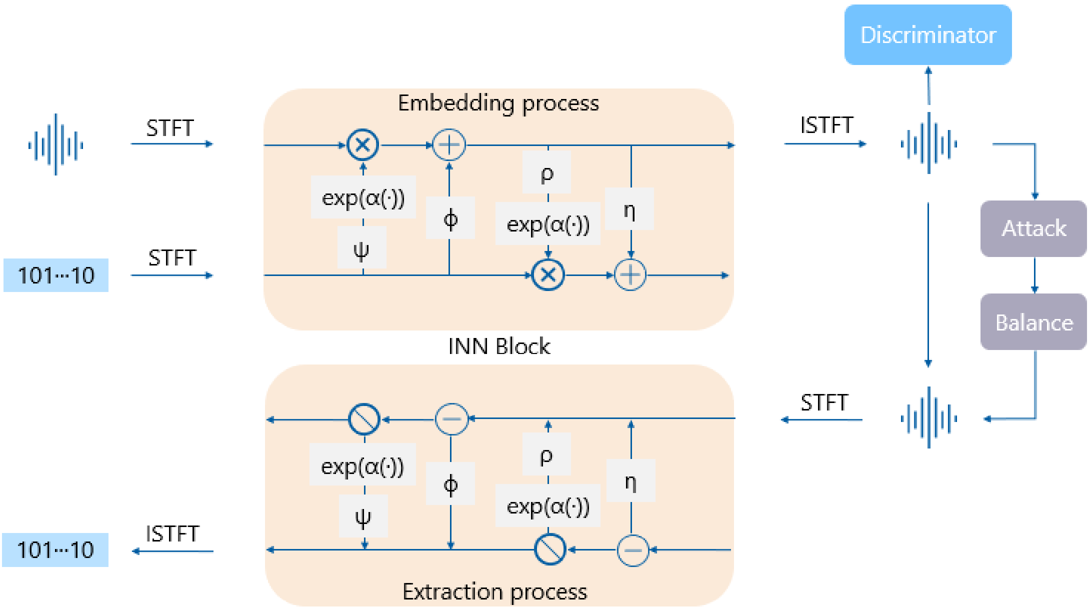

Figure 1: Model structure: contains watermarked messages, raw audio signals, invertible neural network blocks, discriminator, simulated attack layer and balance blocks

We employ a staged training approach, wherein a simulated attack layer is applied during the second half of training to emulate multiple random attacks on watermarked audio. For the attacked audio, we incorporate a balancing block to adjust its spectrum and mitigate the symmetry breaking induced by the attacks. In the extraction process, the objective is to accurately retrieve the watermark information from the watermarked audio. Since the extraction process is the inverse of the embedding process, we utilize an invertible neural network to perform an inverse transform and extract the watermark information from its frequency domain representation. The extracted watermark extension information is obtained by converting the frequency domain representation back to the time domain through the Inverse Short-Time Fourier Transform (ISTFT). The watermark extension information is then reduced to the original watermark size via a fully connected layer. Finally, the original watermark message is reconstructed by separating and intercepting the information based on the dimensions of the localization code and the watermark data.

To optimize our watermarking model, we perform backpropagation and optimization by calculating the total loss, which comprises integrity loss, perception loss, discrimination loss, and identification loss. Integrity loss is used to measure the difference between the extracted watermark information and localization code and the embedded watermark information and localization code. The loss function formula is as follows.

$$\mathcal{L}_1 = ||m' - m||_2 + ||c' - c||_2 \tag{3}$$

Here, m denotes the watermark information and c denotes the localization code. The perceptual loss is used to measure the difference between the original audio spectrum and the watermarked audio spectrum, and the loss function formula is as follows.

$$\mathcal{L}_2 = ||x' - x||_2 \tag{4}$$

Where x denotes the audio spectral data. The discrimination loss is used to measure the accuracy of the discriminator in classifying the original audio versus the watermarked audio, and the identification loss is used for adversarial training to optimize the network model in order to confuse the discriminator, the loss function formula is as follows.

$$\mathcal{L}3 = -y \cdot \log(D(x)) - (1 - y) \cdot \log(1 - D(x)) \tag{5}$$

$$\mathcal{L}4 = -\log(1 - D(x')) \tag{6}$$

where y denotes a label of 0 or 1 for classification by the discriminator. The total loss function is as follows, where λ_1, λ_2 and λ_3 are the weights of the components.

$$\mathcal{L}_{\text{total}} = \lambda_1\mathcal{L}_1 + \lambda_2\mathcal{L}_2 + \lambda_3\mathcal{L}_4 \tag{7}$$

Imperceptibility is an important metric for evaluating the performance of watermarking, and the main role of the discriminator block in this model is to ensure that the effect of the watermarked embedding on the original audio is undetectable. By training the discriminator to distinguish between the original audio and the watermarked audio, the embedding network is forced to learn to hide the watermark information in locations that are difficult to detect, thus improving the imperceptibility of the watermark. The discriminator is a neural network consisting of multiple fully connected layers and activation functions that ultimately output a probability value between 0 and 1, indicating whether the audio contains a watermark or not. Input adjustment is first performed to spread the spectrogram into a one-dimensional vector. Then, the feature dimensions are reduced layer by layer in the fully connected network by means of multiple linear layers and ReLU activation functions. Finally, a probability set is output in the output layer using a sigmoid activation function, which is used to determine whether or not the audio contains a watermark. To improve the accuracy of the discriminator, we adjust the input data to be the spectrogram of the audio data instead of the original audio data.

To make the neural audio watermark robust against various watermark removal attacks, we introduce a simulated attack layer. The main role of the attack block in this model is to simulate various audio attacks, such as adding noise, filtering, and compression, during the training process in order to improve the model's robustness against these attacks in real applications. By introducing attacks during training, the model is able to learn to extract watermark information accurately even after being attacked. In this study, the types of attacks we use include Gaussian noise, bandpass filtering, random erasure, random discard, resampling, amplitude adjustment, MP3 compression, and time stretching. For each audio sample, one attack type is randomly selected.

The introduction of the attack layer destroys the symmetry of the whole embedding and extraction process, which affects the training of the INN. In order to maintain the parameter sharing of the INN and the symmetry of the INN training at the same time, a balancing block is used to alleviate the asymmetry caused by the attack layer and stabilize the symmetric structure of the model. The main role of the balancing block in this model is to mitigate the symmetry breaking introduced by the attack layer, maintain the symmetric structure of the invertible neural network (INN), and ensure that the model can still be stable for watermark extraction in the face of attacks. By using the balancing block, the model can effectively adjust the audio spectrum after an attack to make it close to the distribution when it is unattacked, thus improving the accuracy and robustness of watermark extraction. It consists of a neural network with multiple dense blocks, a LeakyReLU activation function, convolutional layers, LayerNorm, and a residual connection.

4. Experiments

4.1 Experimental Setup

Our experiments were conducted on the AutoDL server. The selected configuration parameters include: the server’s CPU is a 16-core Xeon® Platinum 8481C, the GPU is an Nvidia RTX 4090D, and the GPU driver version is 550.78. The detailed parameters are shown in Table 1.

Table 1: Configuration parameters

Item	Setting	Item	Setting
System	Ubuntu20.04	CPU	16-core Xeon(R) Platinum 8481C
GPU Driver	550.78	GPU	Nvidia RTX 4090D
Memory	80GB	Cuda	11.8
Torch	2.0	Learning rate	0.001
Python	3.8	Decay	1e-5
Optimizer	Adam	Epochs	100

We train on the VCTK and FMA public datasets. The VCTK dataset contains approximately 44 hours of speech data, consisting of around 400 sentences read aloud by 110 English speakers with different accents. The FMA dataset contains a substantial amount of music audio. These two types of datasets are commonly used in audio watermarking scenarios. The training, validation, and test sets are divided into an 8:1:1 ratio.

We divide the training process into two phases. The first phase prioritizes imperceptibility and watermark integrity to ensure that the watermarking model can embed imperceptible watermarks and accurately extract watermarks without attacks. In the second phase, we introduce a simulated attack layer and balancing blocks, which ensure that the model is robust against common attacks and meets the needs of real-world scenarios based on the training from the first phase by processing randomly simulated attacks on each audio sample. During the training process, we divide all audio data into 1-second segments and resample them to 16 kHz. The Fourier transform size (FFT Length) used is 1000, Hop Length is 250, and Window Length is 1000 for the Short Time Fourier Transform (STFT), calculated using a Hann window. Both phases of the model were trained using the Adam optimizer with a learning rate of 1e-5. We fixed the length of the locator code to 10 bits and the length of the watermarked message to 22 bits, and in each batch, the locator code and the watermarked message were randomly generated to ensure that the model could handle any combination of 0-1 sequences. Considering the energy differences across audio data, we perform normalization during audio data preprocessing to reduce the impact of different carrier energies.

4.2 Result Analysis

We choose signal-to-noise ratio (SNR), bit error rate (BER), and watermark capacity as the measures of the model. SNR is used to measure the impact of the watermarked message on the original audio data, and the larger its value, the stronger the imperceptibility of the watermark. BER is used to measure the difference between the extracted watermarked message and the original watermarked message, which reflects the robustness of the watermarking model. The closer its value is to 0, the lower the degree of difference between the two. In this study, we set the watermarking capacity of the model to 32 bits, with the length of the locator code being 10 bits. The results of comparing the method proposed in this paper with the methods in the comparative literature without attacks are shown in Table 2. We test it on two classical public audio datasets. From the results, our model excels in both imperceptibility and robustness while ensuring high capacity.

Table 2: The test results of this algorithm and the comparison algorithm on different test sets (without attacks)

DataSet	Model	SNR	BER	Capacity
VCTK	[6]	38.55	0.0065	32
	[7]	40.43	0.0036	20
	[8]	26.18	0.0039	8.8
	Ours	39.57	0.0048	32
FMA	[6]	35.78	0.0081	32
	[7]	37.72	0.0048	20
	[8]	24.28	0.0045	8.8
	Ours	40.13	0.0051	32

In order to test the robustness of our proposed method against common attacks and synchronization attacks, several attacks are performed on watermarked signals, including Gaussian Added Noise (GN): Gaussian noise is added to the watermarked audio, and the signal-to-noise ratio is kept at approximately 35 dB; Low-Pass Filter (LF): a low-pass filter of 5 kHz is used; MP3 Compression (CP): the waveform is compressed into the 64 kbps MP3 format and then converted back to the original format; Quantization (QZ): the samples of the watermarked audio waveform are quantized to 29 levels; Random Discard (RD): 0.1% of the sample points in the watermarked audio are randomly discarded; Resample (RS): the audio is resampled to 200% of the original sample rate and then resampled back to the original sample rate; Amplitude Adjustment (AM): the overall amplitude of the audio is adjusted to 90% of the original; Time Stretch (TS): the audio is first compressed in the time domain to 90% of the original length and then stretched to maintain the original length. The results are shown in Table 3, and we find that our proposed method outperforms the comparative literature methods in general based on higher embedding capacity, which we believe is related to the simulated attack layer and balancing block introduced during our training process, ensuring the robustness of the watermarking model against attacks.

For our proposed method, we use ablation tests to verify the effectiveness of individual components. As shown in Table 4, removing all components negatively affects the model's performance. The experimental results demonstrate that the discriminator component improves the model's imperceptibility; without the discriminator, the signal-to-noise ratio decreases, although robustness remains high. After removing the balancing block component, we find that the BER of the model increases, which we believe is due to the absence of the balancing block to mitigate the asymmetry introduced by the attack layer. Conversely, the signal-to-noise ratio of the model remains comparable to that of the proposed method. Overall, the discriminator and balancing block designed in our model play important roles in enhancing the imperceptibility and robustness of the model, ensuring the quality of the watermark and the stability of the watermarking system.

Table 3: Robustness test results under various attacks (1 - BER values)

	GN	**LF**	**CP**	**QZ**	**RD**	**RS**	**AM**	**TS**
[6]	0.9784	0.9854	0.9881	0.9660	0.9868	0.9842	0.9929	0.9535
[7]	0.9947	0.9862	0.9941	0.9886	0.9960	0.9911	0.9949	0.9895
[8]	0.9961	0.9904	0.9955	0.9963	0.9961	0.9962	0.9961	0.9931
Ours	0.9931	0.9855	0.9920	0.9882	0.9921	0.9896	0.9950	0.9880

Table 4: Ablation Study

Model	SNR	BER
Ours	40.13	0.0051
- Discriminator	36.42	0.0057
- Balance Block	39.76	0.0086

5. Conclusion

The audio watermarking model based on the invertible neural network proposed in this paper improves watermark localization efficiency by combining the localization code and the watermarking message. It introduces a discriminator and a balancing block to enhance the imperceptibility and robustness of the watermarking model, ensuring the quality and stability of the watermarking under high embedding capacity conditions. The model exhibits high imperceptibility and watermark robustness in the absence of attacks. Even after multiple common and synchronization attacks, the overall performance still outperforms that of the comparative literature. Individual component tests demonstrate that our proposed methods effectively improve the imperceptibility and robustness of watermarking systems. Future directions include enhancing the localization efficiency of our algorithms, such as customizing the localization code length or designing more efficient watermark localization schemes, as well as improving the robustness of our algorithms in the presence of multiple hybrid attacks.

References

[1] Kumar K P, Kanhe A. An adaptive embedding approach for high imperceptible and robust audio watermarking using framelet transform and SVD[J]. Circuits, Systems, and Signal Processing, 2023, 42(9): 5684-5713.

[2] Liu X, Li X, Shi C, et al. A novel SVD-based adaptive robust audio watermarking algorithm[J]. Multimedia Tools and Applications, 2024: 1-23.

[3] Hua G, Huang J, Shi Y Q, et al. Twenty years of digital audio watermarking—a comprehensive review[J]. Signal processing, 2016, 128: 222-242.

[4] Jing J, Deng X, Xu M, et al. Hinet: Deep image hiding by invertible network[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 4733-4742.

[5] Moritz M, Olán T, Virtanen T. Noise-to-Mask Ratio Loss for Deep Neural Network Based Audio Watermarking[C]//2024 IEEE 5th International Symposium on the Internet of Sounds (IS2). IEEE, 2024: 1-6.

[6] Chen G, Wu Y, Liu S, et al. Wavmark: Watermarking for audio generation[J]. arXiv preprint arXiv:2308.12770, 2023.

[7] Li P, Zhang X, Xiao J, et al. IDEAW: Robust Neural Audio Watermarking with Invertible Dual-Embedding[J]. arXiv preprint arXiv:2409.19627, 2024.

[8] Liu C, Zhang J, Fang H, et al. Dear: A deep-learning-based audio re-recording resilient watermarking[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2023, 37(11): 13201-13209.

[9] Dinh L, Krueger D, Bengio Y. Nice: Non-linear independent components estimation[J]. arXiv preprint arXiv:1410.8516, 2014.

[10] Kingma D P, Dhariwal P. Glow: Generative flow with invertible 1x1 convolutions[J]. Advances in neural information processing systems, 2018, 31.

[11] Lan Y, Shang F, Yang J, et al. Robust image steganography: hiding messages in frequency coefficients[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2023, 37(12): 14955-14963.

[12] Montajabi Z, Ghassab V K, Bouguila N. Invertible Neural Network-Based Video Compression[C]//ICPRAM. 2023: 558-564.

[13] Lugmayr A, Danelljan M, Van Gool L, et al. Srflow: Learning the super-resolution space with normalizing flow[C]//Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16. Springer International Publishing, 2020: 715-732.

[14] Zhu X, Li Z, Zhang X Y, et al. Residual invertible spatio-temporal network for video super-resolution[C]//Proceedings of the AAAI conference on artificial intelligence. 2019, 33(01): 5981-5988.

[15] Liu Y, Qin Z, Anwar S, et al. Invertible denoising network: A light solution for real noise removal[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 13365-13374.