Do You Steal My Model? Signature Diffusion Embedded Dual-Verification Watermarking for Protecting Intellectual Property of Hyperspectral Image Classification Models

Yufei Yang[1] , **Song Xiao**[2*] , **Lixiang Li**[1] , **Wenqian Dong**[3] and **Jiahui Qu**[3]
[1]School of Cyberspace Security, Beijing University of Posts and Telecommunications, Beijing, China
[2]Department of Electronic and Communication Engineering, Beijing Electronic Science and Technology Institute, Beijing, China
[3]State Key Laboratory of Integrated Service Network, Xidian University, Xi'an, China
{yfyang, lixiang}@bupt.edu.cn, xiaosong@mail.xidian.edu.cn, {wqdong, jhqu}@xidian.edu.cn

Abstract

Due to the high cost of data collection and training, the well-performed hyperspectral image (HSI) classification models are of great value and vulnerable to piracy threat during transmission and use. Model watermarking is a promising technology for intellectual property (IP) protection of models. However, the existing model watermarking methods for RGB image classification models ignore the complexity of ground objects and high dimension of HSIs, which makes trigger samples easy to be detected and forged. To address this problem, we propose a signature diffusion embedded dual-verification watermarking method, which generates imperceptible trigger samples with explicit owner information to achieve dual verification of both model ownership and legality of trigger set. Specifically, the subpixel-space owner signature diffusion incorporated imperceptible trigger set generation method is proposed to manipulate owner signature incorporated to the abundance matrix of seeds via diffusion model in subpixel space, thus balancing the perceptual quality of trigger samples and signature extraction capability. To resist ownership confusion, dual-stamp ownership verification is proposed to query the suspicious model with trigger samples for ownership verification, and further extracts signature from trigger samples to guarantee their legality. Extensive experiments demonstrate the proposed method can effectively protect IP of HSI classification models.

1 Introduction

Deep learning (DL) techniques have tremendously developed and achieved great success in various hyperspectral image (HSI) processing tasks, e.g. classification [Yang *et al.*, 2024a], change detection [Luo *et al.*, 2024; Qu *et al.*, 2025a], super-resolution [Qu *et al.*, 2025b] and so on. HSI classification models, which accurately identify objects on the Earth surface, play a crucial role in environmental monitoring and military [Li *et al.*, 2019a]. To fully mine the capability of HSI classification models, network architectures need to be delicately designed, and substantial computational resources is required. Additionally, considering the specialized equipment for HSI acquisition and expert knowledge for data notation, it is hard to obtain sufficient training data. Therefore, a well-performed HSI classification model is of great values. The illegal use of HSI classification model may result in the leakage of sensitive information, thereby threatening to the national security. The intellectual property (IP) protection of HSI classification models has emerged as a pressing demand.

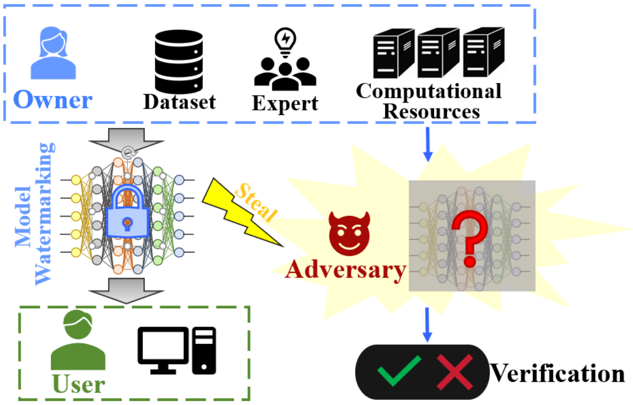

Figure 1: The illustration of model watermarking for intellectual property protection.

Inspired by the idea of digital watermarking [Fang *et al.*, 2021], model watermarking methods are developed to protect the IP of deep models (shown in Figure 1). Generally, according to the information accessible to the verifier, the model watermarking methods are divided into white-box watermarking and black-box watermarking. White-box watermarking methods embed the watermark information into the weights or the activations of the model [Rouhani *et al.*, 2019; Li *et al.*, 2021]. However, the adversaries may not willing to provide the internal parameters and structure of model to verifiers, which limits the practicability in real-world scenarios. Black-box watermarking can be verified easier, which

[*] Corresponding author

only requires API access to the suspicious model. Black-box watermarking methods train the model with a set of carefully crafted trigger samples and the designated labels to embed a secret watermark [Li *et al.*, 2023; Kim *et al.*, 2023]. The ownership of the model is verified by comparing the outputs of the model with the pre-defined labels of the trigger samples.

However, mainstream black-box methods construct trigger set by embedding a noticeable pattern to clean samples, which is easily detected by human inspection. Although some researches try to generate imperceptible trigger samples, they ignore the data characteristic of HSIs, which may result in spectral distortion of trigger samples. Moreover, the existing model watermarking methods fail to verify ownership when trigger sets generated by different parties provide the same mapping relationship, leading to a successful forgery attack.

To address these problems, we propose a signature diffusion embedded dual-verification watermarking for IP protection of HSI classification models, which constructs imperceptible trigger set with owner signature to verify the ownership of the model against forgery attack. Specifically, the seeds near decision boundary are selected to capture the information of the protected model while avoiding significant impact to the classification task. The owner signature is incorporated to the abundance matrix of the selected seeds via diffusion model in subpixel space, thus ensuring the imperceptibility of trigger samples. To resist the forgery attack, a dual-stamp ownership verification strategy is proposed to make the watermark traceable. The contributions of this paper are:

- We propose a signature diffusion embedded dual-verification watermarking method, which is the first to introduce the IP protection problem into HSI classification tasks.
- We propose a subpixel-space owner signature diffusion incorporated imperceptible trigger set generation method to hide the owner signature into the abundance of the seeds in subpixel space via diffusion model, thus ensuring the imperceptibility of trigger samples with explicit owner information.
- We design a dual-stamp ownership verification strategy to verify both the ownership of the model and the legality of trigger set, effectively resisting forgery attack.

2 Related Work

Model Watermarking

Various model watermarking methods have been proposed for the IP protection of deep models, which can be divided into white-box watermarking and black-box watermarking.

White-box watermarking methods embed the watermark into the parameters of the target model. [Uchida *et al.*, 2017] first proposed watermarking deep models by embedding a bit string into a middle layer. [Rouhani *et al.*, 2019] embedded the watermark information in the probability density function of activation map. [Liu *et al.*, 2021] greedily selected a few and important model parameters for watermarking embedding to improve the robustness. [Fan *et al.*, 2022] added a special passport layer to the model, which performs unsatisfactory when the weights of passport are incorrect. [Zhao *et al.*, 2021] pruned the internal channels of the model with pruning rates controlled by watermark. The verification process of white-box watermarking methods requires the internal details of suspicious model, severely limiting its applicability.

Black-box watermarking methods facilitate ownership verification without detailed information of the parameters or structures of the model [Li *et al.*, 2024]. [Adi *et al.*, 2018] embedded abstract images as backdoor watermark into the model to verify the ownership. A blind watermark framework [Li *et al.*, 2019b] guaranteed the key samples with similar distribution of the original samples. [Li *et al.*, 2022] verified the ownership of suspicious model with the knowledge of defender-specified external features. CosWM [Charette *et al.*, 2022] embedded a cosine signal into the output of teacher model to defend against model distillation. [Lin *et al.*, 2024] utilized logistic chaos mapping to chunk and dislocate trigger samples with original labels. An unambiguous backdoor watermarking method [Hua *et al.*, 2023a] increased the cost of ambiguity attacks to exponential complexity. [Liu *et al.*, 2024] modified clean samples in frequency domain to generate trigger sample without noticeable artifacts. However, IP protection for HSI classification models is unexplored. How to generate imperceptible trigger set to protect HSI classification models remains an important open question.

Diffusion Model

Diffusion models [Sohl-Dickstein *et al.*, 2015] are proposed as generative model with high flexibility, which destroy the structure in a data distribution, and recover the data via a gradually denoising process.[Ho *et al.*, 2020] proposed denoising diffusion probabilistic models (DDPMs) to simplify this process. DDPMs use a Markov chain to convert the input image into Gaussian noise in the forward process, and adopt a denoising network to predict the added noise in the reverse process. Denoising diffusion implicit models (DDIMs) [Song *et al.*, 2020] provide an alternative noising process without Markov chain constraints, enabling faster sampling than DDPMs. Diffusion models have shown remarkable results in various computer vision tasks including image translation [Zhao *et al.*, 2022], steganography [Yu *et al.*, 2024; Yang *et al.*, 2024b], etc. In this paper, the potential of diffusion model is explored to embed the owner signature into HSI to generate imperceptible trigger samples.

3 Proposed Method

3.1 Threat Model

In the threat model, we consider a attack-defense setting with two parties: model owner $\mathcal{O}$ and adversary $\mathcal{A}$. The model owner aims to train a well-performed HSI classification model $F_{\mathcal{O}}$ and provide API for clients. The adversary may illegally construct a stolen model $F_{\mathcal{A}}$ that has similar performance ($Acc(F_{\mathcal{A}}) \approx Acc(F_{\mathcal{O}})$). Therefore, the model owner intends to design an effective model IP protection method to verify the ownership of suspicious model.

Adversary's Capability. The adversary steals a model and removes its watermark without affecting its functionality, but has limited computational resources and data. Moreover, aware of the input-output relationships of trigger set, the adversary may forge a trigger set and claim ownership.

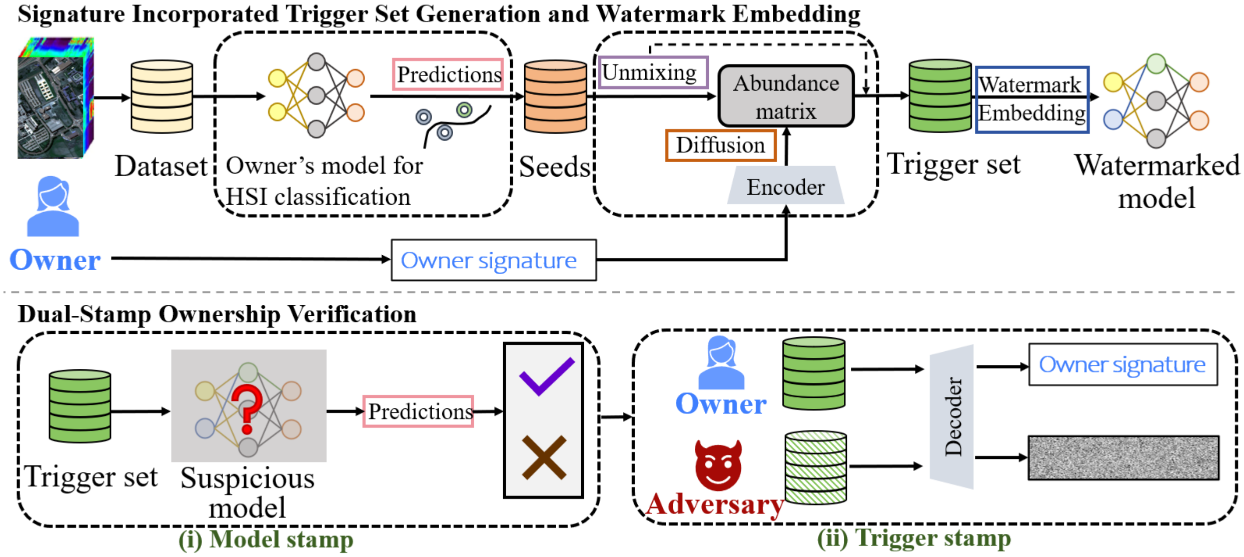

Figure 2: The framework of the proposed signature diffusion embedded dual-verification watermarking method for IP protection of HSI classification model.

Defender's Capability. The defender, who is the legitimate developer and owner of the model, embeds a watermark to assert ownership for unauthorized use. The defender has full access to the inner details of the model, while only API of the suspicious model is available for ownership verification.

3.2 Overview

We assume the HSI classification model $F_{\mathcal{O}}$ is trained with N labeled samples, and learns a function $F_{\mathcal{O}} : \mathcal{X} \rightarrow \mathcal{Y}$ to classify each pixel in HSI $\mathbf{X}_i \in \mathcal{X}$ into corresponding class $y_i \in \mathcal{Y} = \{1, 2, ..., c\}$. The problem of HSI classification model IP protection is defined as: given the model $F_{\mathcal{O}}$, the watermark is embedded into $F_{\mathcal{O}}$ to obtain watermarked model $F_{\mathcal{O}}^{\mathcal{WM}}$, and the owner can extract the watermark to verify whether the suspicious model $F_{\mathcal{SP}}$ is stolen from $F_{\mathcal{O}}^{\mathcal{WM}}$.

In this paper, the subpixel-space signature diffusion embedded dual-verification watermarking method is proposed for HSI classification model protection. The core idea is to generate imperceptible trigger samples that encode owner signature in the subpixel space for model watermarking, thus enabling ownership verification with both model stamp and trigger stamp to resist the forgery attack. As shown in Figure 2, the proposed method can be formalized into two steps:
1) Signature Incorporated Trigger Set Generation and Watermark Embedding: The model owner selects seeds $\mathcal{S} = \{(\mathbf{X}_s, y_s)\}_{s=1}^{S}$ near decision boundary and incorporates owner signature $\mathbf{Sig}$ into the abundance of seeds in the subpixel space (shown in Figure 3) to obtain trigger set, which harnesses the diffusion model to ensure stealthiness. The generated trigger set with pre-defined labels $\mathcal{R} = \{(\widehat{\mathbf{X}}_r, y_r)\}_{r=1}^{R}$ are embedded into model $F_{\mathcal{O}}$ as model watermark:

$$F_{\mathcal{O}}^{\mathcal{WM}}(\theta^*) = \arg\min_{\theta_0} \sum_{r=1}^{R} L(F_{\mathcal{O}}(\theta_0; \widehat{\mathbf{X}}_r), y_r) \qquad (1)$$

where θ^* and θ_0 are the parameters of the watermarked model $F_{\mathcal{O}}^{\mathcal{WM}}$ and the model without watermark $F_{\mathcal{O}}$.
2) Dual-Stamp Ownership Verification: The owner first verifies the ownership of the suspicious model $F_{\mathcal{SP}}$ remotely with the generated trigger set $\mathcal{R}$. To avoid forgery attack, the owner further verifies if the signature extracted from trigger set $\widehat{\mathbf{Sig}}$ matches the embedded signature $\mathbf{Sig}$. The dual-stamp ownership verification can be formulated as:

$$\{Verify_1(F_{\mathcal{SP}},\mathcal{R}) \geq \varepsilon\} \& \left\{Verify_2\left(\mathbf{Sig}, \widehat{\mathbf{Sig}}\right) \geq \eta\right\} \qquad (2)$$

3.3 Subpixel-Space Owner Signature Diffusion Incorporated Imperceptible Trigger Set Generation

Seeds Selection

To avoid significant changes of decision boundaries of model $F_{\mathcal{O}}$ in watermark embedding process, the samples that be easily misclassified to the target class are selected as seeds $\mathcal{S} = \{(\mathbf{X}_s, y_s)\}_{s=1}^{S}$. This ensures the generated trigger samples with imperceptible perturbation $\mathcal{R} = \{(\widehat{\mathbf{X}}_s, y_r)\}_{r=1}^{R}$ are still near the decision boundaries. Since the samples with small difference between the top two predicted probability values are regarded to be near decision boundaries, we select seeds $\mathbf{X}_s$ according to the following rule shown in Figure 4:

$$\mathbf{X}_s = Min_{top}^{i}(Pro_{y_n}(\mathbf{X}_n) - Pro_{y_{sec}}(\mathbf{X}_n)), y_r = y_{sec} \qquad (3)$$

where $Min_{top}^{i}(\cdot)$ selects the top i samples belong to class c_i with the smallest difference, Pro_a denotes the probability of class a, and y_{sec} is the label of second-highest class.

Subpixel Information Diffusion Based Signature Hiding

A subpixel information diffusion based signature hiding method is proposed to embed signature $\mathbf{Sig}$ into selected seeds $\mathbf{X}_s$ in subpixel space, which adopts diffusion model

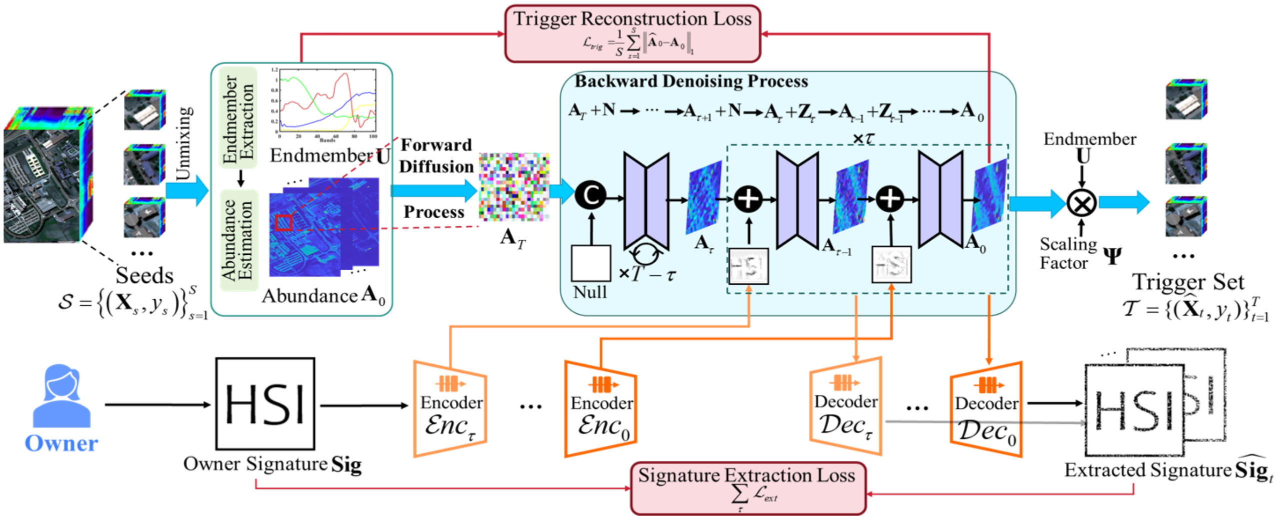
Trigger Reconstruction Loss
$\mathcal{L}_{trig}=\frac{1}{S}\sum_{s=1}^{S}\|\hat{\mathbf{A}}_0-\mathbf{A}_0\|_1$
Unmixing
Endmember Extraction
Endmember U
Abundance Estimation
Abundance $\mathbf{A}_0$
Seeds
$\mathcal{S}=\{(\mathbf{X}_s,y_s)\}_{s=1}^{S}$
Forward Diffusion Process
$\mathbf{A}_T$
Backward Denoising Process
$\mathbf{A}_T+\mathbf{N}\rightarrow\cdots\rightarrow\mathbf{A}_{\tau+1}+\mathbf{N}\rightarrow\mathbf{A}_\tau+\mathbf{Z}_\tau\rightarrow\mathbf{A}_{\tau-1}+\mathbf{Z}_{\tau-1}\rightarrow\cdots\rightarrow\mathbf{A}_0$
$\times\tau$
Null
$\times T-\tau$
$\mathbf{A}_\tau$
$\mathbf{A}_{\tau-1}$
$\mathbf{A}_0$
Endmember U
Scaling Factor Ψ
Trigger Set
$\mathcal{T}=\{(\hat{\mathbf{X}}_t,y_t)\}_{t=1}^{T}$
Owner
HSI
Owner Signature Sig
Encoder $\mathcal{E}nc_\tau$
Encoder $\mathcal{E}nc_0$
Decoder $\mathcal{D}ec_\tau$
Decoder $\mathcal{D}ec_0$
Extracted Signature $\widehat{\mathbf{Sig}}_t$
Signature Extraction Loss
$\sum_\tau \mathcal{L}_{ext}$

Figure 3: Overview of the subpixel-space owner signature diffusion incorporated imperceptible trigger set generation method.

to make trigger samples indistinguishable from seeds. The foundation lies in changing abundance of HSIs only involves partial spectral information, making perturbation in subpixel space subtler than in pixel space.

Specifically, the spectral signature of each pixel is decomposed into endmember signatures $\mathbf{U}$ and corresponding abundance in subpixel space. The endmember matrix is estimated by vertex component analysis (VCA), and under the extended linear mixture model assumption, the abundant matrix $\mathbf{A}$ of HSI $\mathbf{X}$ is obtained by solving the following problem:

$$\mathbf{X}=\mathbf{U}(\boldsymbol{\Psi}\circ\mathbf{A})+\mathbf{E} \tag{4}$$

where $\boldsymbol{\Psi}$ is the scaling factor, and $\mathbf{E}$ is the additive noise.

Given the abundant matrix of seeds $\mathbf{A}_{s,0}\sim q(\mathbf{A})$, the diffusion model is adopted to embed the signature to the abundant of the seeds in the subpixel space. The forward process gradually adds Gaussian noise to a real data distribution through T steps to obtain a noisy representation $\mathbf{A}_{s,T}$:

$$\mathbf{A}_{s,t}=\sqrt{\bar{\alpha}_t}\mathbf{A}_{s,0}+\sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon},\quad \boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I}) \tag{5}$$

where $\mathbf{A}_{s,t}$ is the noisy representation of the t-th step, $\bar{\alpha}_t$ is a noise scheduler, and $\boldsymbol{\epsilon}$ is Gaussian noise.

In the reverse process, the proposed method maps the Gaussian representation $\mathbf{A}_{s,T}$ to an abundance matrix $\hat{\mathbf{A}}_{s,0}\sim q(\mathbf{A})$ through DDIM, while injecting the signature to the estimated $\mathbf{A}_{s,t}$ via signature encoders. In this way, the signature is embedded to the generated abundance $\hat{\mathbf{A}}_{s,0}$, and can be extracted by the signature decoders to confirm the legality of the trigger sets. The signature encoder of the t-th step $\mathcal{E}nc_t$ is designed to learn the embedding signature $\mathbf{Z}_t$:

$$\mathbf{Z}_t=\mathcal{E}nc_t(\mathbf{Sig}),\quad t<\tau \tag{6}$$

For each step t, the input is obtained by concatenating the embedding signature with the estimated $\mathbf{A}_{s,t}$, denoted as $(\mathbf{A}_{s,t}\odot\mathbf{Z}_t)$, and the embedding signature is replaced with a null signature $\mathbf{N}$ when $t>\tau$ to minimize the influence on visual quality of the generated abundance $\hat{\mathbf{A}}_{s,0}$. The noise predictor $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}$ is adopted to estimate the noise added to $\mathbf{A}_{s,0}$, and the estimation of $\mathbf{A}_{s,0}$ is formulated as:

$$\hat{\mathbf{A}}_{s,0}^{t}=\begin{cases}\dfrac{(\mathbf{A}_{s,t}\odot\mathbf{N})-\sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{A}_{s,t}\odot\mathbf{N})}{\sqrt{\bar{\alpha}_t}},\ t\geq\tau\\[2ex]\dfrac{(\mathbf{A}_{s,t}\odot\mathbf{Z}_t)-\sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{A}_{s,t}\odot\mathbf{Z}_t)}{\sqrt{\bar{\alpha}_t}},\ t<\tau\end{cases} \tag{7}$$

The estimated noise is reintroduced to the approximated $\hat{\mathbf{A}}_{s,0}^{t}$ to obtain $\mathbf{A}_{s,t-1}$:

$$\mathbf{A}_{s,t-1}=\begin{cases}\sqrt{\bar{\alpha}_{t-1}}\hat{\mathbf{A}}_{s,0}^{t}+\sqrt{1-\bar{\alpha}_{t-1}}\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{A}_{s,t}\odot\mathbf{N}),t\geq\tau\\\sqrt{\bar{\alpha}_{t-1}}\hat{\mathbf{A}}_{s,0}^{t}+\sqrt{1-\bar{\alpha}_{t-1}}\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{A}_{s,t}\odot\mathbf{Z}_t),t<\tau\end{cases} \tag{8}$$

The signature decoder $\mathcal{D}ec_t$ is trained to extract the signature $\widehat{\mathbf{Sig}}_t$ concealed in $\mathbf{A}_{s,t-1}$:

$$\widehat{\mathbf{Sig}}_t=\mathcal{D}ec_t(\mathbf{A}_{s,t-1}),\quad t<\tau \tag{9}$$

The trigger sample $\hat{\mathbf{X}}_s$ is obtained by adjusting the generated abundance of seeds $\hat{\mathbf{A}}_{s,0}$ with the scaling factors $\boldsymbol{\Psi}$ and multiplying with the endmember $\mathbf{U}$:

$$\hat{\mathbf{X}}_s=\mathbf{U}(\boldsymbol{\Psi}\circ\hat{\mathbf{A}}_{s,0}) \tag{10}$$

Loss Function

Two key aspects warrant attention in the training process, i.e., visual quality of trigger samples and extraction accuracy of signature. Specifically, the trigger reconstruction loss $\mathcal{L}_{trig}$ minimizes the mean absolute error between trigger samples $\hat{\mathbf{A}}_{s,0}$ and seeds $\mathbf{A}_{s,0}$. The signature extraction loss constrains the cross entropy between the extracted signature $\widehat{\mathbf{Sig}}_t$ and original signature $\mathbf{Sig}$. The total loss $\mathcal{L}_{total}$ is defined as:

$$\begin{aligned}\mathcal{L}_{total}&=\alpha\mathcal{L}_{trig}+\beta\sum_{\tau}\mathcal{L}_{ext}\\&=\alpha\left\|\hat{\mathbf{A}}_{s,0}-\mathbf{A}_{s,0}\right\|_1+\beta\sum_{t=\tau}^{0}\Big(\mathbf{Sig}\log\widehat{\mathbf{Sig}}_t+(1-\mathbf{Sig})\log(1-\widehat{\mathbf{Sig}}_t)\Big)\end{aligned} \tag{11}$$

where α and β (set as 1 and 0.1) are loss weights.

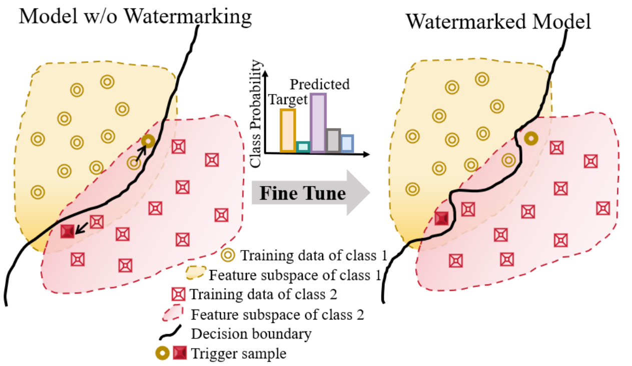

Figure 4: The illustration of the seeds selection rules.

3.4 Watermark Embedding

Once the trigger samples $\widehat{\mathbf{X}}_s$ with the signature information **Sig** is generated, the model owner can obtain the watermarked model $F_{\mathcal{O}}^{\mathcal{W}M}$ by training the model with clean samples and trigger samples. Based on the over-parameterization property of the owner model $F_{\mathcal{O}}$, the training process is able to force the model to learn the mapping between the trigger samples and the pre-defined labels. The watermark embedding process can be formulated as:

$$\widehat{\mathbf{X}}_s \mapsto y_t \neq F_{\mathcal{O}}(\widehat{\mathbf{X}}_s) \tag{12}$$

3.5 Dual-Stamp Ownership Verification

Assuming the watermarked model $F_{\mathcal{O}}^{\mathcal{W}M}$ is leaked, the owner can verify the ownership by extracting the watermarking from the suspicious model F_{SP} and comparing with the original watermarking. In this paper, a dual-stamp ownership verification strategy is proposed to confirm the owner of the suspicious model and further verify the generator of trigger set, which can be seen as a dual-stamp process to resist ownership confusion caused by forgery attack.

Definition 1. *Given the suspicious model F_{SP}, the model owner queries the predicted results of F_{SP} with the trigger samples. The model owner claims that the ownership of the suspicious model if the accuracy of F_{SP} on the trigger set $\mathcal{R} = \{(\widehat{\mathbf{X}}_s, y_r)\}_{r=1}^{R}$ is lager than the threshold ε:*

$$\frac{Num(F_{SP}(\widehat{\mathbf{X}}_s) = y_r)}{R} \geq \varepsilon \tag{13}$$

where $Num(\cdot)$ denotes the function calculating the number of trigger samples classified into the pre-defined class.

Definition 2. *Given the trigger samples to be verified as $\widehat{\mathbf{X}}'_s$, the model owner claims the legality of the trigger samples if the distance between the extracted signature* $\mathbf{Sig}'$ *and the original signature* $\mathbf{Sig}$ *is smaller than the threshold η:*

$$d\left(\mathbf{Sig}, \mathbf{Sig}'\right) = \left\| \mathcal{D}ec(\widehat{\mathbf{X}}'_s) - \mathbf{Sig} \right\|_2 \leq \eta \tag{14}$$

4 Experiment

4.1 Implementation Details

We evaluate the performance of the proposed method on three datasets. Pavia University dataset contains 9 classes and has 610×340 pixels with 103 spectral bands. Indian Pines dataset is composed of 224 bands and the spatial dimension is 145×145, which includes 16 land-cover types. Salinas dataset is of size 512×217×224 and contains 16 classes. We conduct experiments on three models, i.e., MCNN [Zheng *et al.*, 2021], DBDA [Li *et al.*, 2020], and SSFTT [Sun *et al.*, 2022] to comprehensively evaluate the performance across different backbones (2D-CNN, 3D-CNN, and Transformer).

In the experiments, HSI is divided into 27×27 and owner signature is fixed as the same spatial size. The DDIM sampler with 200 sampling steps is adopted. The batch size is 128 and learning rate of 1e-3. The number of epochs is set as 500.

OA and Kappa coefficient (KC) indicate the classification accuracy. Watermark success rate (WSR) is used to calculate the probability that the trigger samples are correctly classified into pre-defined label y_t. Bit error rate (BER) evaluates quality of extracted signature. PSNR, SSIM, and SAM evaluate visual quality as well as spectral distortion of trigger samples.

Dataset	Model	w/o watermark			watermarked		
		OA	KC	WSR	OA	KC	WSR
Pavia University	MCNN	99.02%	98.91%	0.0%	98.95%	98.84%	100%
	DBDA	99.01%	98.91%	2.0%	98.95%	98.84%	100%
	SSFTT	99.33%	99.26%	0.0%	99.02%	98.92%	100%
Indian Pines	MCNN	94.72%	94.39%	0.0%	95.16%	94.86%	100%
	DBDA	97.15%	96.97%	7.0%	97.10%	96.92%	100%
	SSFTT	98.07%	97.95%	0.0%	97.90%	97.78%	100%
Salinas	MCNN	99.92%	99.91%	0.0%	99.92%	99.91%	100%
	DBDA	99.76%	99.75%	0.0%	99.89%	99.88%	96.0%
	SSFTT	99.90%	99.89%	5.0%	99.82%	99.81%	100%

Table 1: The fidelity and effectiveness of the proposed method.

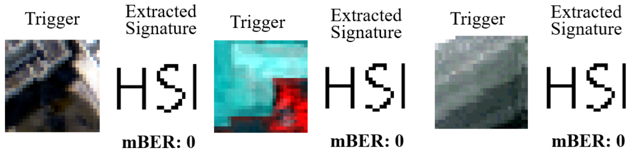

Figure 5: Visualization of the extracted signature on three datasets.

4.2 Results

Fidelity

Fidelity measures the impact of watermark embedding on the performance of the original model. OA and KC of clean model and watermarked model are compared on different datasets to evaluate the performance degradation in Table 1. The proposed method achievesan OA drop of less than 0.1% on Salinas dataset and even improves the classification accuracy of MCNN on Indian Pines dataset. The results show that the watermarked models achieve similar performance to clean model, indicating the good fidelity of the proposed method.

Effectiveness

Effectiveness measures whether the ownership of model can be verified successfully by the watermarking method. As shown in Table 1, WSR of watermarked model are higher

Rate (%)	MCNN						DBDA						SSFTT					
	UHFB		MEA		Proposed		UHFB		MEA		Proposed		UHFB		MEA		Proposed	
	OA	WSR	OA	WSR	OA	WSR	OA	WSR	OA	WSR	OA	WSR	OA	WSR	OA	WSR	OA	WSR
0	93.44%	100%	98.30%	100%	98.95%	100%	98.55%	100%	97.59%	100%	98.95%	100%	98.88%	100%	98.57%	100%	99.02%	100%
10	93.46%	100%	98.24%	100%	98.95%	100%	98.62%	100%	97.40%	100%	98.93%	100%	98.86%	100%	98.14%	100%	98.98%	95.0%
20	93.28%	100%	98.25%	100%	98.96%	100%	97.59%	99.0%	97.39%	100%	98.14%	95.0%	98.88%	100%	98.09%	100%	98.98%	95.0%
30	93.33%	100%	98.26%	100%	98.96%	100%	94.94%	100%	85.23%	100%	96.99%	68.0%	98.75%	100%	98.05%	100%	98.98%	95.0%
40	93.44%	100%	98.28%	100%	98.96%	100%	72.94%	98.0%	85.73%	100%	61.54%	20.0%	98.73%	100%	97.71%	100%	98.93%	95.0%
50	93.18%	100%	98.27%	100%	98.97%	100%	48.03%	89.0%	87.31%	100%	65.34%	9.0%	97.90%	100%	94.43%	100%	98.18%	98.0%
60	93.07%	100%	98.25%	100%	98.97%	100%	44.11%	62.0%	74.76%	99.0%	36.46%	10.0%	97.02%	81.0%	87.40%	97.0%	85.09%	78.0%
70	92.99%	100%	98.20%	100%	98.95%	100%	2.59%	29.0%	2.21%	0.0%	14.18%	10.0%	87.88%	30.0%	51.81%	61.0%	54.24%	27.0%
80	91.85%	100%	98.06%	100%	99.16%	100%	5.11%	10.0%	2.21%	0.0%	2.21%	19.0%	47.59%	10.0%	21.30%	4.0%	25.09%	2.0%
90	80.90%	100%	97.66%	100%	99.07%	100%	43.54%	10.0%	2.21%	0.0%	2.21%	19.0%	14.71%	10.0%	15.49%	0.0%	19.25%	2.0%

Table 2: The robustness of the different methods against model pruning attack.

Dataset	Model	Before Quantization			After Quantization		
		OA	KC	WSR	OA	KC	WSR
Pavia University	MCNN	98.95%	98.84%	100%	98.96%	98.85%	100%
	DBDA	98.95%	98.84%	100%	98.95%	98.84%	100%
	SSFTT	99.02%	98.92%	100%	99.04%	98.93%	100%
Indian Pines	MCNN	95.16%	94.86%	100%	95.16%	94.86%	100%
	DBDA	97.10%	96.92%	100%	97.11%	96.93%	100%
	SSFTT	97.90%	97.78%	100%	97.91%	97.78%	100%
Salinas	MCNN	99.91%	99.91%	100%	99.92%	99.91%	100%
	DBDA	99.89%	99.88%	96.0%	99.89%	99.88%	96.0%
	SSFTT	99.82%	99.81%	100%	99.82%	99.81%	100%

Table 3: The robustness of the proposed method against weight quantization on three datasets.

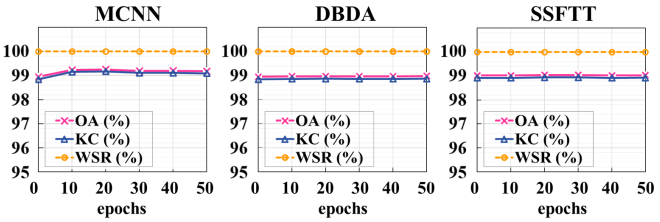

Figure 6: The robustness of the proposed method against fine-tuning attack.

than 95% on three datasets and three different models, while the clean models without watermarking have WSR lower than 10% for trigger samples. The results can verify the ownership without falsely claiming the ownership of clean models.

To verify the effectiveness for avoiding forgery attack, we evaluate both visual results and objective metrics of extracted owner signature (shown in Figure 5). The signature extracted from trigger samples generated by the owner exhibits high quality with low mean BER. The proposed method proves the unique relationship between the trigger samples and the owner, avoiding the confusion of model's ownership.

Robustness

We evaluate the robustness of the propose method against model fine-tuning, model pruning, and weight quantization.

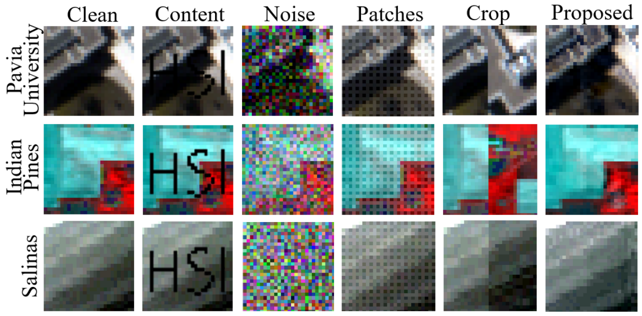

Figure 7: Visualization of the trigger samples generated by different methods.

Model Fine-Tuning: Model fine-tuning retrains the watermarked model to remove watermark by modifying the weights. We randomly select 1% samples in the test set to fine-tune the watermarked model, and the numbers of epochs are set from 10 to 50. As shown in Figure 6, WSR of fine-tuned model is still 100% for all models, and OA and KC of the watermarked model show minimal decline after fine-tuning. Therefore, the proposed method is robust to model fine-tuning attack.

Model Pruning: Model pruning removes redundant parameters of the model while maintaining the performance of primary task. We prune the watermarked model with prune rate from 10% to 90% to evaluate the performance and robustness. As can be observed in Table 2, until the pruning rate reaches a threshold that significantly degrades the classification performance, the proposed method can effectively extract the watermark, showing robustness to model pruning attack.

Weight Quantization: Weight quantization compresses the weights to lower bit representation to reduce the storage requirements. Table 3 shows that the values of WSR are still 100%, and the accuracy of classification can be maintained after quantization on all datasets, which shows that the proposed method is resistant to weight quantization attack.

Stealthiness

To evaluate the stealthiness of the proposed method, we compare the trigger samples generated by the proposed method with four representative methods, i.e., Content,

Method	Pavia University			Indian Pines			Salinas		
	mSSIM	mPSNR	mSAM	mSSIM	mPSNR	mSAM	mSSIM	mPSNR	mSAM
Content	0.6611	25.8246	17.7778	0.6707	29.3622	17.7778	0.7118	32.1098	17.7778
Noise	0.2534	20.7870	26.4717	0.1823	21.2097	18.1754	0.1555	21.4089	26.3584
Patches	0.7013	25.7970	**4.5160**	0.6375	27.3806	4.6016	0.6809	29.3188	5.3699
Crop	0.6431	24.6746	8.8856	0.7882	33.0796	5.4622	0.8381	33.5038	9.5902
Proposed	**0.8413**	**32.1444**	6.1126	**0.9411**	**42.8497**	**2.7704**	**0.9891**	**48.1443**	**1.8615**

Table 4: Comparison of stealthiness of different watermarking methods on three datasets.

Dataset	Method	w/o watermark			watermarked		
		OA	KC	WSR	OA	KC	WSR
Pavia University	UHFB	99.33%	99.26%	**0.0%**	98.88%	98.52%	**100%**
	MEA	-	-	**0.0%**	98.54%	98.54%	98.5%
	SSW	-	-	100%	98.22%	97.64%	**100%**
	Proposed	-	-	**0.0%**	**99.02%**	**98.92%**	**100%**
Indian Pines	UHFB	98.07%	97.95%	**0.0%**	95.25%	94.59%	**100%**
	MEA	-	-	**0.0%**	93.16%	92.21%	**100%**
	SSW	-	-	100%	95.47%	94.84%	**100%**
	Proposed	-	-	**0.0%**	**97.90%**	**97.78%**	**100%**
Salinas	UHFB	99.90%	99.89%	10.0%	98.77%	98.63%	**100%**
	MEA	-	-	**0.0%**	99.73%	99.70%	**100%**
	SSW	-	-	100%	99.06%	98.95%	**100%**
	Proposed	-	-	5.0%	**99.82%**	**99.81%**	**100%**

Table 5: Comparison of fidelity and effectiveness of different model watermarking methods.

Noise, Patches [Wang *et al.*, 2022], and Crop [Lv *et al.*, 2024]. As shown in Figure 7, the trigger patterns of Content and Crop is easily distinguished, which exposes the construction way of trigger set. The noticeable artifacts can be observed on trigger samples generated by Noise and Patches. As shown in Table 4, the proposed method can minimize the visual difference between clean samples and trigger samples, and achieve the best results in terms of mean PSNR, SSIM, and SAM, indicating good stealthiness.

4.3 Comparison to Other Methods

To verify the superiority of the proposed method, we conduct the experiments to compare it to the state-of-the-art methods [Hua *et al.*, 2023b; Lv *et al.*, 2024; Tan *et al.*, 2023]. The results of different methods on various datasets are shown in Table 5. MEA yields a 4.91% accuracy loss in terms of OA on India Pines dataset, while the dropout of OA and KC of the proposed method are less than 0.34% on all datasets, achieving better fidelity than UHFB, MEA, and SSW. This is because the proposed method selects the seeds near decision boundaries and injects imperceptible information to the seeds, ensuring small impact to the performance of the model.

As for the effectiveness, the WSR of all methods are larger than 98%, and the proposed method achieves the highest WSR. However, the main drawback of the competing methods is that they cannot resist forgery attack. The proposed method enables the signature information extracted from the trigger samples to claim the legality of the trigger set, improving the effectiveness for ownership verification.

Number	w/o watermark			watermarked		
	OA	KC	WSR	OA	KC	WSR
50	99.90%	99.89%	8.0%	99.89%	99.89%	100%
100	-	-	5.0%	99.82%	99.81%	100%
200	-	-	5.0%	99.79%	99.78%	100%

Table 6: Results of the proposed method with different numbers of trigger samples.

Time Step τ	mSSIM	mPSNR	mSAM	OA	KC
200	0.9459	42.3547	2.4932	99.51%	99.45%
100	0.9620	44.0146	2.3578	99.71%	99.68%
20	0.9891	48.1443	1.8615	99.82%	99.81%

Table 7: Results of the proposed method with different time steps that begins to inject the owner signature.

4.4 Ablation Study

To evaluate the impact of number of trigger samples, we conduct experiments on the proposed method with 50, 100, 200 trigger samples. As shown in Table 6, a larger trigger set leads to a slight decline in OA and KC, indicating a reduction in fidelity. Although WSR remains 100% for different numbers of trigger sets, smaller trigger sets lead to higher miss detection rates. In this paper, 100 trigger samples are adopted to balance fidelity and effectiveness.

We discuss the effect of time step that begins to inject the owner signature (threshold τ). The experiments on models with τ=200, 100, and 20 demonstrate that injecting the signature only in the final steps (τ=20) of diffusion model can minimize the visual influence of signature embedding to trigger sample as well as the impact to classification accuracy.

5 Conclusion

In this paper, we propose a signature diffusion embedded dual-verification watermarking to protect IP of HSI classification models for the first time, which verifies the ownership of model with imperceptible trigger set. The subpixel-space owner signature diffusion incorporated imperceptible trigger set generation method is designed to encode owner signature to the seeds in subpixel space via diffusion model. To ensure adversary cannot forge trigger samples to confuse ownership, dual-stamp ownership verification strategy is proposed to verify both model ownership and legality of trigger set. Experiments demonstrate the effectiveness of the proposed watermarking method in IP protection of HSI classification model.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62201423, 62471359, and 62476013, the Key Research and Development Program of Shaanxi under Grant 2025SF-YBXM-513, the Young Talent Fund of Association for Science and Technology in Shaanxi under Grant 20230117 and 20250133, the Young Talent Fund of Xi'an Association for Science and Technology under Grant 959202313052, and in part by the Fundamental Research Funds for the Central Universities under Grant QTZX25084.

References

[Adi *et al.*, 2018] Yossi Adi, Carsten Baum, Moustapha Cisse, Benny Pinkas, and Joseph Keshet. Turning your weakness into a strength: Watermarking deep neural networks by backdooring. In *27th USENIX Security Symposium (USENIX Security 18)*, pages 1615–1631, 2018.

[Charette *et al.*, 2022] Laurent Charette, Lingyang Chu, Yizhou Chen, Jian Pei, Lanjun Wang, and Yong Zhang. Cosine model watermarking against ensemble distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 9512–9520, 2022.

[Fan *et al.*, 2022] Lixin Fan, Kam Woh Ng, Chee Seng Chan, and Qiang Yang. DeepIPR: Deep neural network ownership verification with passports. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6122–6139, 2022.

[Fang *et al.*, 2021] Han Fang, Dongdong Chen, Qidong Huang, Jie Zhang, Zehua Ma, Weiming Zhang, and Nenghai Yu. Deep template-based watermarking. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(4):1436–1451, 2021.

[Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[Hua *et al.*, 2023a] Guang Hua, Andrew Beng Jin Teoh, Yong Xiang, and Hao Jiang. Unambiguous and high-fidelity backdoor watermarking for deep neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–14, 2023.

[Hua *et al.*, 2023b] Guang Hua, Andrew Beng Jin Teoh, Yong Xiang, and Hao Jiang. Unambiguous and high-fidelity backdoor watermarking for deep neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[Kim *et al.*, 2023] Byungjoo Kim, Suyoung Lee, Seanie Lee, Sooel Son, and Sung Ju Hwang. Margin-based neural network watermarking. In *International Conference on Machine Learning*, pages 16696–16711. PMLR, 2023.

[Li *et al.*, 2019a] Shutao Li, Weiwei Song, Leyuan Fang, Yushi Chen, Pedram Ghamisi, and Jón Atli Benediktsson. Deep learning for hyperspectral image classification: An overview. *IEEE Transactions on Geoscience and Remote Sensing*, 57(9):6690–6709, 2019.

[Li *et al.*, 2019b] Zheng Li, Chengyu Hu, Yang Zhang, and Shanqing Guo. How to prove your model belongs to you: a blind-watermark based framework to protect intellectual property of DNN. In *Proceedings of the 35th Annual Computer Security Applications Conference*, ACSAC '19, page 126–137, New York, NY, USA, 2019. Association for Computing Machinery.

[Li *et al.*, 2020] Rui Li, Shunyi Zheng, Chenxi Duan, Yang Yang, and Xiqi Wang. Classification of hyperspectral image based on double-branch dual-attention mechanism network. *Remote Sensing*, 12(3):582, 2020.

[Li *et al.*, 2021] Yue Li, Benedetta Tondi, and Mauro Barni. Spread-transform dither modulation watermarking of deep neural network. *Journal of Information Security and Applications*, 63:103004, 2021.

[Li *et al.*, 2022] Yiming Li, Linghui Zhu, Xiaojun Jia, Yong Jiang, Shu-Tao Xia, and Xiaochun Cao. Defending against model stealing via verifying embedded external features. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 1464–1472, 2022.

[Li *et al.*, 2023] Peixuan Li, Pengzhou Cheng, Fangqi Li, Wei Du, Haodong Zhao, and Gongshen Liu. Plmmark: a secure and robust black-box watermarking framework for pre-trained language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 14991–14999, 2023.

[Li *et al.*, 2024] Fangqi Li, Haodong Zhao, Wei Du, and Shilin Wang. Revisiting the information capacity of neural network watermarks: Upper bound estimation and beyond. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 21331–21339, 2024.

[Lin *et al.*, 2024] Huanjie Lin, Shuyuan Shen, and Haojie Lyu. Protecting IP of deep neural networks with watermarking using logistic disorder generation trigger sets. *Multimedia Tools and Applications*, 83(4):10735–10754, 2024.

[Liu *et al.*, 2021] Hanwen Liu, Zhenyu Weng, and Yuesheng Zhu. Watermarking deep neural networks with greedy residuals. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 6978–6988. PMLR, 18–24 Jul 2021.

[Liu *et al.*, 2024] Yong Liu, Hanzhou Wu, and Xinpeng Zhang. Robust and imperceptible black-box DNN watermarking based on fourier perturbation analysis and frequency sensitivity clustering. *IEEE Transactions on Dependable and Secure Computing*, 2024.

[Luo *et al.*, 2024] Fulin Luo, Tianyuan Zhou, Jiamin Liu, Tan Guo, Xiuwen Gong, and Xinbo Gao. DCENet: Diff-feature contrast enhancement network for semi-supervised hyperspectral change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–14, 2024.

[Lv *et al.*, 2024] P. Lv, H. Ma, K. Chen, J. Zhou, S. Zhang, R. Liang, S. Zhu, P. Li, and Y. Zhang. MEA-Defender: A robust watermark against model extraction attack. In *2024*

IEEE Symposium on Security and Privacy (SP), pages 102–102, Los Alamitos, CA, USA, may 2024. IEEE Computer Society.

[Qu *et al.*, 2025a] Jiahui Qu, Wenqian Dong, Qian Du, Yufei Yang, Yunshuang Xu, and Yunsong Li. Cyclic consistency constrained multiview graph matching network for unsupervised heterogeneous change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–15, 2025.

[Qu *et al.*, 2025b] Jiahui Qu, Xiaoyang Wu, Wenqian Dong, Jizhou Cui, and Yunsong Li. IR&ArF: Toward deep interpretable arbitrary resolution fusion of unregistered hyperspectral and multispectral images. *IEEE Transactions on Image Processing*, 34:1934–1949, 2025.

[Rouhani *et al.*, 2019] Bita Darvish Rouhani, Huili Chen, and Farinaz Koushanfar. Deepsigns: an end-to-end watermarking framework for protecting the ownership of deep neural networks. In *ACM International Conference on Architectural Support for Programming Languages and Operating Systems*, volume 3, 2019.

[Sohl-Dickstein *et al.*, 2015] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015.

[Song *et al.*, 2020] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[Sun *et al.*, 2022] Le Sun, Guangrui Zhao, Yuhui Zheng, and Zebin Wu. Spectral-spatial feature tokenization transformer for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2022.

[Tan *et al.*, 2023] Jingxuan Tan, Nan Zhong, Zhenxing Qian, Xinpeng Zhang, and Sheng Li. Deep neural network watermarking against model extraction attack. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 1588–1597, 2023.

[Uchida *et al.*, 2017] Yusuke Uchida, Yuki Nagai, Shigeyuki Sakazawa, and Shin'ichi Satoh. Embedding watermarks into deep neural networks. In *Proceedings of the 2017 ACM on international conference on multimedia retrieval*, pages 269–277, 2017.

[Wang *et al.*, 2022] Lixu Wang, Shichao Xu, Ruiqi Xu, Xiao Wang, and Qi Zhu. Non-transferable learning: A new approach for model ownership verification and applicability authorization. In *International Conference on Learning Representations*, 2022.

[Yang *et al.*, 2024a] Yueguang Yang, Jiahui Qu, Wenqian Dong, Tongzhen Zhang, Song Xiao, and Yunsong Li. TMCFN: Text-supervised multidimensional contrastive fusion network for hyperspectral and LiDAR classification. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–15, 2024.

[Yang *et al.*, 2024b] Zijin Yang, Kai Zeng, Kejiang Chen, Han Fang, Weiming Zhang, and Nenghai Yu. Gaussian shading: Provable performance-lossless image watermarking for diffusion models. *arXiv preprint arXiv:2404.04956*, 2024.

[Yu *et al.*, 2024] Jiwen Yu, Xuanyu Zhang, Youmin Xu, and Jian Zhang. CRoss: Diffusion model makes controllable, robust and secure image steganography. *Advances in Neural Information Processing Systems*, 36, 2024.

[Zhao *et al.*, 2021] Xiangyu Zhao, Yinzhe Yao, Hanzhou Wu, and Xinpeng Zhang. Structural watermarking to deep neural networks via network channel pruning. In *2021 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6. IEEE, 2021.

[Zhao *et al.*, 2022] Min Zhao, Fan Bao, Chongxuan Li, and Jun Zhu. Egsde: Unpaired image-to-image translation via energy-guided stochastic differential equations. *Advances in Neural Information Processing Systems*, 35:3609–3623, 2022.

[Zheng *et al.*, 2021] Jianwei Zheng, Yuchao Feng, Cong Bai, and Jinglin Zhang. Hyperspectral image classification using mixed convolutions and covariance pooling. *IEEE Transactions on Geoscience and Remote Sensing*, 59(1):522–534, 2021.