



Proactive Schemes: A Survey of Adversarial Attacks for Social Good

Vishal Asnani^{1,2} · Xi Yin³ · Xiaoming Liu¹

Received: 25 September 2024 / Accepted: 1 November 2025
© The Author(s) 2026

Abstract

Adversarial attacks in computer vision exploit the vulnerabilities of machine learning models by introducing subtle perturbations to input data, often leading to incorrect predictions or classifications. These attacks have evolved in sophistication with the advent of deep learning, presenting significant challenges in critical applications, which can be harmful for society. However, there is also a rich line of research from a transformative perspective that leverages adversarial techniques for social good. Specifically, we examine the rise of proactive schemes, methods that encrypt input data using additional signals termed templates, to enhance the performance of deep learning models. By embedding these imperceptible templates into digital media, proactive schemes are applied across various applications, from simple image enhancements to complicated deep learning frameworks to aid performance, as compared to the passive schemes, which don't change the input data distribution for their framework. The survey delves into the methodologies behind these proactive schemes, the data perturbation and template learning processes, and their application to modern computer vision and natural language processing applications. Additionally, it discusses the challenges, potential vulnerabilities, and future directions for proactive schemes, ultimately highlighting their potential to foster the responsible and secure advancement of deep learning technologies.

Keywords Proactive schemes · Templates · Perturbation · Social good · Adversarial attacks

1 Introduction

Adversarial attacks in computer vision exploit vulnerabilities in machine learning models by introducing subtle, often imperceptible perturbations to input data, leading to incorrect predictions or classifications. The subtle manipulations used in these attacks can lead to misinterpretations by AI systems, potentially causing widespread harm in critical applications such as security surveillance, healthcare diagnostics, and autonomous transportation (Dong et al., 2018; Huang et al.,

2017). Deep learning has been the main reason for significant development in different computer vision tasks, as shown in Sect. 1 (Table 1).

In the pre-deep learning era, traditional CV applications relied on handcrafted features and basic algorithms for object detection, image classification, and facial recognition, utilizing techniques such as edge detection, texture, and color (Chapelle et al., 1999; Haralick et al., 2007). Adversarial attacks during this period were less sophisticated, primarily involving manipulations like introducing noise or performing basic operations such as blurring and compression (Petitcolas et al., 1998; Westfeld & Pfitzmann, 1999).

In the deep learning era, traditional CV applications have evolved significantly with the advent of deep learning models like CNNs and transformers (He et al., 2016; Simonyan & Zisserman, 2014; Vaswani et al., 2017). These advancements have enhanced applications such as real-time object detection, advanced image classification, vision and large language models, and facial recognition, leading to substantial improvements in accuracy and efficiency. Adversarial attacks have also become more sophisticated, exploiting deep neural networks' vulnerabilities to create misleading inputs

Communicated by Yasuyuki Matsushita.

The literature review was done at Michigan State University.

✉ Vishal Asnani
asnani@msu.edu

Xi Yin
yinx@meta.com

Xiaoming Liu
liuxm@msu.edu

¹ Michigan State University, East Lansing, USA

² Adobe Research, San Jose, CA, USA

³ Meta AI, Menlo Park, USA

Table 1 Evolution of visual computing techniques across eras. We distinguish among traditional CV tasks, adversarial attacks, and proactive schemes. While earlier techniques such as steganography or signal enhancement shared goals with modern proactive schemes, they lacked

adversarial optimization and model awareness. This survey focuses on proactive schemes, which are adversarially optimized perturbations intentionally designed for constructive use in the deep learning era

Aspect	Pre-deep learning era	Deep learning era
Traditional CV Applications	<ul style="list-style-type: none"> - Handcrafted features: edge detection, color, texture. - Rule-based processing 	<ul style="list-style-type: none"> - Learned features from CNNs, transformers. - Application-specific deep architectures
Adversarial Attacks	<ul style="list-style-type: none"> - Rare and unsystematic use. - Visible noise or compression-based distortions 	<ul style="list-style-type: none"> - Optimized perturbations targeting neural decision boundaries. - Model-specific evasion or manipulation objectives
Proactive Schemes for Social Good (This Survey)	<ul style="list-style-type: none"> - Steganography and signal embedding without optimization. - Goals implicit in enhancement or obfuscation 	<ul style="list-style-type: none"> - Learnable perturbation templates with task-specific objectives. - Applications: GenAI/LLM protection, ownership, privacy, robust inference, content provenance

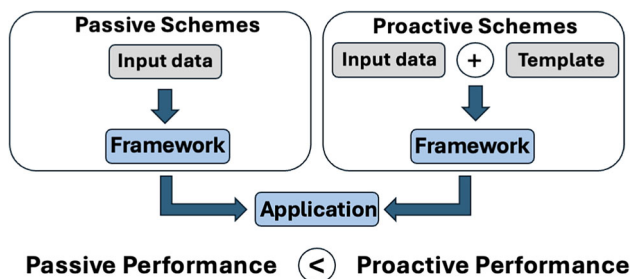


Fig. 1 Passive versus proactive schemes: Passive schemes take an input as is for their methods, while proactive schemes use templates to encrypt the input and use the encrypted data as the input. The advantage of the proactive schemes comes from their improved performance in downstream applications compared to the passive schemes

that appear normal to humans (Carlini & Wagner, 2017; Goodfellow et al., 2014; Madry et al., 2017).

The challenge of adversarial attacks extends beyond technical hurdles, posing ethical, legal, and safety concerns that society must address to ensure the responsible and secure advancement of computer vision applications. While adversarial attacks in computer vision are often viewed through the lens of their potential for harm, there exists a transformative perspective that leverages these techniques for social good (Asnani et al., 2024, 2022, 2023a). By understanding and harnessing the principles behind adversarial perturbations, researchers have innovated protective measures that utilize techniques that enhance various computer vision applications using imperceptible signals added onto the original media, known as *templates* (Asnani et al., 2024,

2022, 2023a), as shown in Fig. 1. The methods that encrypt input data using templates, allowing the encrypted data to enhance the performance for an application, are referred to as *proactive schemes*. In contrast, all methods that operate on the input data without modification are treated as *passive schemes* (Asnani et al., 2024, 2022, 2023a, c). We use the term *social good* to refer to beneficial outcomes that are commonly valued by the society, such as privacy preservation, content attribution, and robustness against misuse. In contrast to traditional adversarial attacks, the techniques surveyed here leverage perturbations proactively to support responsible and ethical uses of AI.

Proactive schemes have been used for a long time, using different methodologies. In the pre-deep learning era, proactive schemes focused on simple enhancements in image processing, with applications like steganography, encryption, and security surveillance (Kanai et al., 1998; Ohbuchi et al., 1998b). Proactive schemes also share a similar idea from approaches using stochastic resonance in signal processing (Gammaitoni et al., 1998) and non-linear systems (Semenov & Zakharova, 2022). Stochastic resonance occurs when a weak signal that is too faint to be detected by a system is enhanced by the addition of noise, allowing the system to cross a detection threshold. This happens because the noise helps to push the weak signal above the threshold intermittently, making it detectable by the system. The interplay between the noise and the signal can amplify the signal’s effects at certain points, leading to an overall improvement in the system’s ability to process or detect the signal. The noise level is tuned to an optimal range; too little noise won’t help

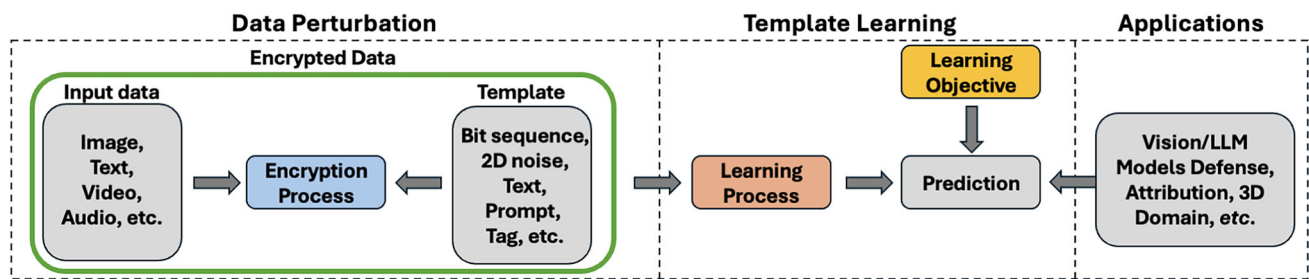


Fig. 2 A general overview of the proactive framework. The method starts by encrypting the input data with a template. This is known as data perturbation. The framework passes through a template learning process and is evaluated based on certain learning objectives. Finally,

every method is associated with a specific application. In the survey paper, we discuss all three stages in a sequential way, with each section focusing on several techniques and aspects of the respective stage

the signal, and too much noise will overwhelm it. However, deep learning has opened up the door for utilizing stochastic resonance in improving the performance by thresholding neural networks (Chen et al., 2022), noise-booster activation functions (Ren et al., 2024b), non-linear stochastic dynamics (Shen et al., 2022), Fourier domain (Rallabandi & Roy, 2010) *etc.* Similarly, many works inject noise in the data or labels as augmentations, to improve the robustness of the deep learning networks (Li et al., 2021a; Nishi et al., 2021; Yoshimura et al., 2023; Yin et al., 2019). Although the above methods resemble proactive schemes, the focus of this survey is on the usage of these schemes for social good in the deep learning era for a variety of applications in the realm of computer vision and natural language processing.

A general framework for proactive schemes is shown in Fig. 2. Each method has a specific **data perturbation** and **template learning** associated with it, which depends on the **application**. Firstly, the perturbation process is a critical component in the design of proactive schemes. This process involves the use of various innovative methods or operations to embed template information within digital media. The templates used for perturbation can take the form of many different types of signals like bit sequences, 2D noises, texts, visual prompts, predefined tags, audio, *etc.* The templates are added onto different types of media, such as images, texts, videos, audios, *etc.* The goal of the perturbation process is to create a secure framework that can withstand potential attacks while maintaining the quality of the encrypted media compared to the original. As technology evolves, so do the techniques used for perturbation, making it an ever-growing area of research.

Next, the template learning process involves training models to recognize and incorporate these templates, whether they are bit sequences, 2D templates, text signals, or visual prompts into various forms of digital content. This integration is achieved through specialized learning paradigms, *eg.* encoder-decoder frameworks, learning via objective functions, adversarial learning, specialized architectures like

GANs, transformers, *etc.*, tailored to the unique characteristics of each template type. The effectiveness of the template learning process is constrained, optimized, and evaluated using a range of objective functions and metrics. This encompasses the stage of learning objectives, which govern the efficacy of the proactive schemes for various applications. The learning objectives are heavily dependent on the application for which the method is being used.

These schemes are used for a plethora of applications, including encryption, GenAI and LLM defense, preservation of authorship rights, ownership verification, improving CV applications, and privacy protection. Based on each application, the researchers have explored various combinations of respective modules of proactive schemes, *i.e.*, type of template, perturbation process, and learning process. This survey comprehensively examines various combinations adopted by the researchers for proactive schemes across different deep learning applications in computer vision and natural language processing.

The survey begins with an overview of the types of templates used in proactive schemes, such as bit sequences, 2D templates, text signals, prompts, and others, supported by the discussion on the perturbation process for each type. The discussion then delves into the template learning process along with the learning objectives associated with each template type. Various applications of these techniques are explored, including defense strategies for vision models and large language models, methods for attribution and preservation of authorship rights, privacy preservation, and techniques specific to the 3D domain. Additionally, the survey covers advancements in improving generative models and other computer vision applications. Following this, the challenges associated with developing these templates, potential attacks against proactive schemes, and the current limitations are critically analyzed. By addressing these topics, the survey aims to enhance the field of computer vision by exploring a realm of proactive learning for social good.

1.1 Comparison with Existing Surveys

While a number of surveys have explored adversarial machine learning and its associated challenges, our work provides a distinct perspective by reframing adversarial perturbations as proactive tools for enabling social good. Table 2 summarizes how our survey complements and differs from existing literature across several key domains.

In contrast to domain-specific or defense-centric reviews, our work synthesizes a growing body of research that uses adversarial perturbations for positive impact, which we refer to as proactive schemes for social good.

2 Our Proposed Taxonomy

2.1 Overview

Figure 2 illustrates our proposed taxonomy, which serves as the foundational framework for understanding proactive schemes that utilize adversarial perturbations for social good. The taxonomy is structured around three key components: the Data Perturbation, the Template Learning, and the Applications. Each of these components corresponds to a major section in this survey and reflects the functional pipeline of proactive adversarial methods, from signal generation to optimization and eventual real-world usage.

Data Perturbation: This component captures the techniques used to generate and embed perturbations, also referred to in this paper as templates, into digital content. These perturbations may be universal or data-dependent and are crafted to be purpose-driven rather than malicious. Unlike conventional adversarial attacks aimed at fooling models, these perturbations are designed to achieve constructive goals such as traceability, attribution, or control over downstream behaviour. This section covers various perturbation strategies, such as pixel-space and latent-space perturbations, adversarial watermarking, and structured signal injection.

Template Learning: The learning process governs how these encrypted signals are optimized to achieve their intended objectives. This includes the training paradigms, loss functions, and co-learning strategies used to associate perturbations with model behavior. Depending on the use case, the perturbations may be learned jointly with the model (end-to-end) or separately (post-hoc). The section explores how this learning takes place, how imperceptibility and robustness are balanced during training, and the role of supervision, ranging from fully supervised to self-supervised or even zero-shot settings.

Applications: This component refers to the real-world contexts in which proactive schemes are applied. The paper identifies five broad application domains: media encryption and tamper detection, generative AI and LLM protection,

model watermarking and authorship attribution, visual forensics and anti-deepfake measures, and privacy-preserving data generation. Each domain imposes unique constraints, such as robustness to transformations, transferability across tasks, or minimal perceptual interference. This section showcases how the techniques developed in the perturbation and learning stages are adapted and deployed to achieve social good across diverse settings.

By organizing the survey around this taxonomy, we aim to provide a cohesive and structured understanding of proactive schemes, highlighting both common patterns and domain-specific adaptations. The following sections expand on each of these components in detail.

2.2 General Formulation of Proactive Schemes

Proactive schemes can be conceptually divided into three stages as defined in our taxonomy: (1) the *data perturbation*, which generates and embeds the perturbation or template; (2) the *Template Learning*, which jointly or independently optimizes the perturbation with respect to task goals; and (3) the *Applications*, where the perturbation is embedded to enable downstream functionality such as attribution, privacy, or robustness.

Baseline Optimization Objective. Let:

- $x \in \mathcal{X}$ be the original input (e.g., image, text, audio),
- \mathcal{D} be the data distribution over \mathcal{X} from which input samples x are drawn,
- $T_\theta(x)$ be the perturbation template generated via the data perturbation process (parameterized by θ),
- $E(x, T_\theta(x))$ be a general embedding function that combines x and the template,
- $f_\phi(\cdot)$ be the downstream model (parameterized by ϕ),
- y be the desired task-specific label or target,
- $\mathcal{L}_{\text{task}}$ be the task-specific loss function.

The general optimization objective becomes:

$$\min_{\theta, \phi} \mathbb{E}_{x \sim \mathcal{D}} [\mathcal{L}_{\text{task}}(f_\phi(E(x, T_\theta(x))), y)]. \quad (1)$$

Data Perturbations stage generates a perturbation template $T_\theta(x)$, which may be universal or input-specific. This template may encode task-specific signals like identity, ownership, or control. The template is added to the input using an embedding function $E(x, T_\theta(x))$. Template Learning Process optimizes the parameters θ and/or ϕ to minimize task loss over the embedded input $E(x, T_\theta(x))$. This process can vary depending on whether the embedding is differentiable or occurs post-hoc. At test time, the learned perturbation is embedded using $E(\cdot)$ into new inputs for downstream use in secure, private, or verifiable settings.

Table 2 Comparison of our survey with prior works across adjacent domains

Survey domain	Typical focus	How our work differs
Adversarial Attacks and Defenses (Akhtar et al., 2021; Costa et al., 2024; Khamaiseh et al., 2022; Liang et al., 2022a; Ozdag, 2018)	Perturbation-based attacks (e.g., FGSM, PGD) and defenses to evaluate or improve model robustness	We shift the focus from adversarial evaluation to constructive usage, using adversarial signals for traceability, privacy, or control
Watermarking in Deep Learning (Potdar et al., 2005; Liu et al., 2024; Lee & Jung, 2001; Mousavi et al., 2014; Singh & Chadha, 2013)	Media or model watermarking for IP protection and tamper detection	We integrate watermarking as one among several applications, framed under proactive design
Secure and Private ML (Xie et al., 2024; Aziz et al., 2023; Sayyad et al., 2024; Akinsiku, 2025; Han et al., 2024)	Federated learning, differential privacy, homomorphic encryption, and SMPC	We explore privacy via embedded templates, offering lightweight, data-level control rather than protocol-level security
Data Provenance and Attribution (Pan et al., 2023; Herschel et al., 2017; Simmhan et al., 2005)	Techniques for tracking data lineage and ownership using cryptographic or system-level tools	We highlight how learned perturbations can serve as embedded provenance or authorship cues within the data itself
Text/NLP Adversarial Attacks (Dhivya et al., 2025; Philip & Minhas, 2022; Sperduti & Moreo, 2025; Zhang et al., 2020)	Generating adversarial text examples for robustness testing in NLP tasks	Our framework is modality-agnostic and focuses on shared principles across text, vision, and audio
Visual Forensics and Deepfake Detection (Mohiuddin et al., 2023; Verdoliva, 2020; Qureshi et al., 2024; Mubarak et al., 2023; Zhang, 2022; Malik et al., 2022)	Detecting manipulated media using classification or forensic signals	Rather than passive detection, we focus on embedding perturbations proactively to facilitate traceability or resistance to misuse

Constraints and Deployment Trade-offs. While Eq. (1) provides a generic objective for proactive schemes, in practice it is optimized under perceptual, capacity, and deployment constraints. These constraints jointly determine the trade-offs between imperceptibility, robustness, and computational efficiency. Perceptual budgets restrict the magnitude of the embedded template, typically enforcing bounds such as $\|T_\theta(x)\|_p \leq \epsilon$ or $\text{LPIPS}(x, E(x, T_\theta(x))) \leq \tau$, to ensure that the perturbation remains imperceptible while preserving task relevance. Many approaches incorporate an expectation-over-transformation (EOT) objective, $\mathbb{E}_{\mathcal{T}}[\mathcal{L}(f_\phi(E(\mathcal{T}(x, T_\theta(x))), y))]$, to promote robustness against transformations including cropping, compression, or re-generation. Capacity constraints further regulate embedding complexity and model size, trading storage and inference cost against payload fidelity. Recent theoretical studies (He et al., 2024; Pang et al., 2024) analyze similar trade-offs between detectability, imperceptibility, and robustness in large-language-model watermarking, while adaptive optimization methods (Liu & Bu, 2024) extend these ideas by learning input-adaptive watermark strengths that preserve downstream task utility. Finally, proactive templates can be designed as *universal*, where a shared template T_u is applied across inputs for efficiency, or as *input-conditioned*, where $T_\theta(x)$ adapts to each input for improved fidelity but higher inference cost.

Unified Expectation-Constrained Objective. Building upon Eq. 1, we incorporate practical constraints that commonly appear across proactive formulations. These include perceptual budgets to limit distortion, an expectation-over-transformation (EOT) component to ensure robustness to post-processing, and regularization terms for perceptual and capacity control. A compact generalized formulation is:

$$\begin{aligned} & \left\{ \min_{\theta, \phi} \mathbb{E}_{x \sim \mathcal{D}} \mathbb{E}_{\mathcal{T} \sim \mathbb{T}} \left[\mathcal{L}_{\text{task}}(f_\phi(E(\mathcal{T}(x, T_\theta(x))), y)) \right] \right. \\ & \left. + \lambda_p \mathcal{R}_p(T_\theta) + \lambda_c \mathcal{R}_c(E) \right\} \\ \text{s.t. } & \|T_\theta(x)\|_p \leq \epsilon, \text{LPIPS}(x, E(x, T_\theta(x))) \leq \tau, \\ & T_\theta(x) = \begin{cases} T_u, & \text{universal template,} \\ T_\theta(x), & \text{input-conditioned template.} \end{cases} \quad (2) \end{aligned}$$

Here, $\mathcal{T} \sim \mathbb{T}$ denotes random transformations (e.g., resize, compression, regeneration) used for augmentation expectation, while \mathcal{R}_p and \mathcal{R}_c regularize perceptual quality and embedding capacity, respectively. This compact form unifies perceptual fidelity, transformation robustness, and deployment trade-offs between universal and input-conditioned templates.

3 Data Perturbation

In the realm of proactive learning and digital security, a variety of innovative methods have been developed to enhance the robustness, authenticity, and ownership protection of digital content. These methods employ different types of templates to embed and verify information across a wide range of applications, including vision models, large language models, and 3D applications. Each template type, whether it involves bit sequences, 2D noises, text signals, visual prompts, or other specialized forms, offers unique advantages and is tailored to specific challenges in the field. The remainder of this section will delve into these template types in detail, exploring their perturbation methodologies and innovative techniques used to embed them effectively.

3.1 Bit Sequences

Bit sequences, represented as one-hot encoding as shown in Fig. 3, are fundamental to many perturbation strategies, particularly in encryption and digital signatures. These sequences are embedded at a binary level, creating a secure, imperceptible layer that effectively detects unauthorized modifications while preserving content fidelity. A summary of techniques for embedding bit sequences is provided in Table 3, with detailed discussions below.

Neural Network Based Techniques Neural network-based embedding techniques leverage the power of neural networks to integrate random binary strings into images. Binary sequences combined with positional values are used by Sun et al. (2023). This method involves two templates: (a) “Strace”, using an encoder-decoder network to input an image and a binary sequence, outputting a template, and (b) “Etrace”, embedding a predefined value in the blue channel of the image, making it imperceptible. “Strace” identifies encrypted images, while “Etrace” detects fake images. Similarly, Meng et al. (2022) embed multiple binary sequences for authentication and traceability using a neural network, DINN, which injects the template at the feature level for verification and origin tracking. Darvish Rouhani et al. (2019) embed random binary strings into datasets using probability distributions of target neural networks for encryption or enhancing model robustness against adversarial attacks. Asnani et al. (2023c) propose to convert bit sequences to spatial noises, and then add those to the input data. Zhang et al. (2024d) conceal a “localization template” and a bit sequence for a template within images using a hiding module and a bit encryption module, respectively. Yu et al. (2021) utilize the stegastamp technique to encrypt binary sequences generated by message generators and embedded by encoders into image data. Zhu et al. (2018) incorporates an encoder-decoder process with image distortion and adversarial losses to guide training. Wu et al. (2020) use multiple encoder networks to convert image



Fig. 3 Bit sequences and 2D noises as a type of templates **a** Yu et al. (2021), **b** Yu et al. (2021); Zeng et al. (2023). Bit sequences templates are a one-hot encoding, which are then embedded into the input data according to different techniques, while 2D templates are spatial noises embedded into the input data

and template features for perturbation. Finally, Wu et al. (2023b) employ multiple binary sequences, one robust and the other non-robust, for source tracing and manipulation localization.

Neural Network Weights and Filters Embedding templates within neural network models, by utilizing their weights and filters, integrates authentication bits into the model’s learning process, enhancing security and performance. Some methods embed the template within the layers of neural network models by utilizing their weights (Chen et al., 2019a; Uchida et al., 2017). Nagai et al. (2018) further proposes using the filters of the convolution layer to embed the template. Additionally, Liu et al. (2021) propose embedding multiple authenticating bits in the objective function of the model to indirectly embed the bit sequence onto images. By incorporating this penalty term, the model’s learning behavior can be fine-tuned to prioritize certain features or patterns, potentially improving performance on specific applications or optimizing convergence during training.

Least Significant Bit (LSB) and Most Significant Bit (MSB) Techniques LSB and MSB techniques manipulate the least or most significant bits of image data to embed binary sequences with minimal visual impact while maintaining template integrity. Haghighi et al. (2018) use Lifting Wavelet Transform (LWT) and LSB rounding to embed bit sequences, preserving the host image’s visual quality. Other works (Dadkhah et al., 2014; Hsu & Tu, 2016; Qin et al., 2017; Cao et al., 2017; Hsu & Tu, 2010) also rely on LSB/MSB rounding for image tampering detection. Paruchuri (2009) conceal template information in selective DCT coefficients for authentication. Zhao et al. (2023b) propose identity-dependent fixed bit encoding, enhancing facial identity features with personalized sequences. Earlier methods like (Lu & Liao, 2001; Lee & Lin, 2008) use wavelet quantization and predefined mapping patterns to embed sequences.

Advanced Embedding Techniques Advanced embedding techniques employ innovative methods, such as AST-based intermediate representation, ASCII character conversion, isotropic unit vectors, and context-aware lexical substitution, which are diverse methods used to enhance the robustness and security of template embedding in digital content. AST-

Table 3 Summary of works which utilize bit sequences as the template

Category	Description	Keywords	References
Neural Network-Based Techniques	Techniques leveraging neural networks to embed random binary strings into images	neural networks, embedding, binary sequences, perturbation, authentication, traceability, encryption	Sun et al. (2023), Meng et al. (2022), Darvish Rouhani et al. (2019), Zhang et al. (2024d), Yu et al. (2021), Zhu et al. (2018), Wu et al. (2020), Wu et al. (2023b)
Neural Network Weights and Filters	Embedding templates within neural network models using their weights and filters	embedding, neural network models, weights, filters, authentication, bits, objective function	Uchida et al. (2017), Chen et al. (2019a), Nagai et al. (2018), Liu et al. (2021)
LSB/MSB Techniques	Manipulating the least or most significant bits of image data to embed binary sequences	LSB, MSB, minimal visual impact, template integrity, DCT coefficients, fixed bit encoding	Haghighi et al. (2018), Dadkhah et al. (2014), Hsu and Tu (2016), Qin et al. (2017), Cao et al. (2017), Hsu and Tu (2010), Lin et al. (2017), Singh and Singh (2017), Zhang et al. (2010), Kiatpapan and Kondo (2015), Singh and Singh (2016), Paruchuri (2009), Zhao et al. (2023b), Lu and Liao (2001), Lee and Lin (2008)
Advanced Embedding Techniques	Innovative methods for embedding templates, including transformations and character conversions	AST representation, ASCII characters, isotropic unit vectors, rare identifiers, lexical substitution	Yang et al. (2023a), Cui et al. (2025), Sablayrolles et al. (2020), Zhao et al. (2023c), Yang et al. (2022b), Fernandez et al. (2023), Furon and Desobieux (2014), He et al. (2022a)
3D Domain	Techniques for embedding binary sequences into 3D models and point clouds	3D models, point clouds, mesh vertices, hash functions, neural networks, spectral applications, wavelet transforms	Venugopal et al. (2011), Wang et al. (2022a), Zhang et al. (2021), Jang et al. (2024), Chen et al. (2024a), Wang et al. (2020), Wang and Kerschbaum (2021), Guo and Potkonjak (2018), Zhang et al. (2024f), Liu et al. (2019), Hamidi et al. (2019), Yeung and Yeo (1998), Wang et al. (2022b), Peng et al. (2022), Zhu et al. (2024), Chou and Tseng (2006), Chou and Tseng (2009), Luo et al. (2023), Molaie et al. (2016), Al-Khatfaji and Abhayaratne (2019), Ohbuchi et al. (2001), Cotting et al. (2004), Kuo et al. (2009), Wang et al. (2008), Mun et al. (2015), Kanai et al. (1998), Uccheddu et al. (2004), Wang et al. (2008), Kim et al. (2005)

based methods customize template embedding based on the structural and content-related aspects of images, tailoring the process to each image's unique characteristics (Yang et al., 2023a). Similarly, converting ASCII characters into binary sequences to create template patches provides a method for spatially integrating robust and semi-robust templates into images. These patches, representing parts of the binary sequence, are directly added to the image, enabling source tracing and manipulation localization (Cui et al., 2025). Both techniques emphasize the need for templates that are not only resistant to attacks but also adaptable to the content they protect.

In classification and identification applications, embedding isotropic unit vectors within the feature space introduces a layer of security and robustness through the incorporation of randomized vectors (Sablayrolles et al., 2020). Each vector, defined for every class, enhances the security of the model by embedding these randomized vectors into data representations. On the other hand, context-aware lexical substitution adapts perturbation to the semantic content of both text and images, ensuring that embedded binary sequences remain contextually relevant and preserving the authenticity of the content (Fernandez et al., 2023; Yang et al., 2022b; Zhu et al., 2018). By leveraging lexical knowledge, techniques like generating interchangeable lexicons embed templates in text outputs from generation APIs, ensuring both the integrity and the identification of the generated text, thus providing protection against plagiarism and unauthorized use (He et al., 2022a). Together, these approaches illustrate the evolving strategies in perturbation that balance robustness, adaptability, and contextual relevance across different types of digital content.

3D Domain Extending perturbation techniques to the 3D domain involves embedding binary sequences into 3D models, point clouds, and mesh vertices using innovative methods like hash functions and neural networks. Bit sequences are utilized in methods involving hash functions (Venugopal et al., 2011), neural networks (Wang et al., 2022a; Zhang et al., 2021), and modifications in 3D point clouds (Hamidi et al., 2019; Liu et al., 2019; Yeung & Yeo, 1998) and mesh vertices (Peng et al., 2022; Wang et al., 2022b). These sequences are also embedded in innovative ways, such as through graph Fourier coefficients (Al-Khafaji & Abhayaratne, 2019; Zhu et al., 2024), spectral applications (Cotting et al., 2004; Ohbuchi et al., 2001), and vertex distributions (Chou & Tseng, 2006; Kuo et al., 2009; Wang et al., 2008; Zhu et al., 2024). Other methods include embedding using neural networks, local deformations to SDF partitions (Mun et al., 2015), and wavelet analysis of 3D objects (Kanai et al., 1998; Uccheddu et al., 2004).

Specific techniques for 3D models, such as modifying spectral coefficients and using wavelet transforms, address the unique challenges of embedding templates in three-

dimensional data. Several specific techniques have been developed to address different aspects of template embedding in 3D models. For instance, the embedding of templates in point clouds involves calculating Root Mean Square Curvature (RMSC) values and modulating radial radii of vertices within ball rings (Liu et al., 2019). In mesh spectral applications, the algorithm modifies spectral coefficients to embed templates resistant to various transformations and attacks (Cotting et al., 2004; Ohbuchi et al., 2001). Wavelet-based approaches, such as those using hierarchical wavelet transforms, enable embedding in semi-regular meshes by modifying wavelet coefficient vectors (Kanai et al., 1998; Wang et al., 2008; Uccheddu et al., 2004).

In conclusion, the various techniques for embedding digital templates into images and 3D models highlight substantial progress in digital perturbation. Methods like fixed-bit encoding, neural network embedding, encoder-decoder architectures, and steganography provide secure and robust integration of binary sequences. Advanced approaches, including AST-based representations and isotropic unit vectors, further improve flexibility and security. Expanding these techniques to the 3D domain emphasizes their versatility and adaptability.

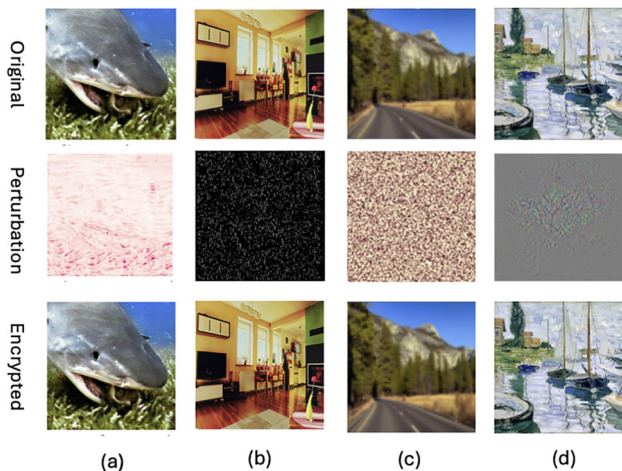
3.2 2D Templates

2D templates, as shown in Fig. 3, embed patterns or noise into 2D spaces, such as images or video frames. These templates are visually imperceptible but detectable, as demonstrated in Fig. 4. Commonly used in image and video perturbation, 2D templates maintain content quality while offering robust protection against tampering. Techniques like perturbation, masking, and spatial transformations seamlessly integrate these templates into digital media. 2D templates can be either fixed or learnable perturbations. Fixed perturbations are predefined and applied to the entire dataset, while learnable perturbations are optimized during training. These perturbations are tailored through various perturbation processes depending on the application. A summary of techniques for embedding 2D noise is provided in Table 4, with detailed discussions below.

Mathematical Operators These methods focus on using mathematical operators for injecting noise or specific patterns directly onto images to obscure signals and enhance security (Asnani et al., 2022, 2023a; Huang et al., 2022; Li et al., 2021; Ruiz et al., 2020; Van Le et al., 2023). Most of the methods involve adding noise or a specific pattern directly onto the image, potentially obscuring the signal. Other mathematical operators like multiplication (Asnani et al., 2024) are also used for the perturbation process. In Li and Lin (2019), direct addition methods use classifiers to learn perturbations that ensure robust template embedding. Universal adversarial perturbations exploit similar cluster centers

Table 4 Summary of works that utilize 2D noises as the template

Category	Description	Keywords	References
Mathematical Operators	Adding noise or patterns to images	direct operations, noise, patterns, security, learnable perturbations	Van Le et al. (2023), Huang et al. (2022), Ruiz et al. (2020), Li et al. (2021), Yang et al. (2021a), Zhong and Deng (2020), Asnani et al. (2022), Asnani et al. (2023a), Fong and Vedaldi (2017), Song et al. (2023), Asnani et al. (2024), Li and Lin (2019), Tang et al. (2024a), Li et al. (2024b), He et al. (2023)
Adversarial Attacks	Creating disruptive or protective perturbations	adversarial attacks, PGD, FGSM, optimization, disruptive, beneficial perturbations	Yeh et al. (2020), Van Le et al. (2023), Huang et al. (2022), Ruiz et al. (2020), Yang et al. (2021a), Zhong and Deng (2020), Peng et al. (2022), Dong et al. (2019), Zhang et al. (2021b), Zhang et al. (2022), Shi and Sagduyu (2017), Liu et al. (2020), Kitada and Iyatomi (2021), Tang et al. (2024a), Ducoffe and Precioso (2018), Xu et al. (2023), Xue et al. (2022), Wu et al. (2024), Madry et al. (2017), Wang et al. (2023a), Wu et al. (2023a)
Autoencoder-Based Learning	Learning image-dependent templates	autoencoder, image-dependent templates, networks, optimization	Segalis and Galili (2020), Wu et al. (2024), Mirjalili et al. (2018), Hu et al. (2022), Xiong et al. (2020), Wu et al. (2019), Xiao et al. (2021), Rajabi et al. (2021), Zeng et al. (2023)
Latent Space Perturbations	Learning perturbation vectors in latent space	latent space, perturbation vectors, 2D space, CVAE, coordinate shifts	Liang et al. (2022b), Lei et al. (2024), Meng et al. (2025), Ding et al. (2021), Wong and Kolter (2020), Shan et al. (2020), Xiao et al. (2021)
Miscellaneous Techniques	Various advanced methods	randomization, camera noise, backdoor perturbation, geometric perturbations, Fourier transform	Dhillon et al. (2018), Xie et al. (2017), Cui et al. (2023), Zhao et al. (2024), Kitada and Iyatomi (2021), Yu et al. (2019), Cozzolino and Verdoliva (2019), Tekgul et al. (2021), Wang et al. (2022c), Molaei et al. (2013), Wen et al. (2023), Le Merrer et al. (2020), Othman and Ross (2014), Mirjalili and Ross (2017)

**Fig. 4** Various examples of input-encrypted input pairs after adding the 2D noise templates into the original input images. **a** Zeng et al. (2023), **b** Asnani et al. (2022), **c** Asnani et al. (2023a) and **d** Cui et al. (2023)

in different models, updating perturbations based on sub-task gradients (Tang et al., 2024a). In deep active learning, random perturbations are added to model parameters (Li et al., 2024b), sampled from Gaussian distributions (He et al., 2023).

Adversarial Attacks Adversarial attacks are employed to create perturbations that can disrupt or protect templates by reversing their typical objective. These attacks, such as PGD and FGSM, usually aim to harm a victim model by creat-

ing disruptive perturbations. However, their objective can be reversed to benefit perturbation algorithms (Huang et al., 2022; Ruiz et al., 2020; Van Le et al., 2023; Yeh et al., 2020; Yang et al., 2021a; Zhong & Deng, 2020). These attacks involve iterative optimization to create perturbations that are particularly disruptive to templates, or conversely, to enhance the perturbation process. Through these attacks, perturbations can be optimized for various applications.

Autoencoder-Based Learning A sophisticated learning paradigm for perturbations uses an autoencoder to learn dependent templates. Various architectures and strategies are adopted, including encoder-based (Segalis & Galili, 2020; Wu et al., 2024), classifier-based (Mirjalili et al., 2018), GAN-based (Hu et al., 2022; Xiong et al., 2020; Xiao et al., 2021; Wu et al., 2019), and ensemble-based (Rajabi et al., 2021). These networks are optimized with different loss functions. Zeng et al. (2023) proposes a method using adversarial training with an injector and a classifier to create image-dependent binary code signatures. The injector introduces perturbations, and the classifier differentiates these signatures, enhancing image authentication and source tracing.

Latent Space Perturbations Latent space perturbations involve learning perturbation vectors in a latent space. Perturbation feature vectors (Ding et al., 2021; Meng et al., 2025; Liang et al., 2022b; Lei et al., 2024) are learned to represent coordinate shifts in the 2D space. Conditional Variational Autoencoder (Wong & Kolter, 2020) models are employed to generate perturbed versions of input data by learning latent

variable distributions, optimizing for reconstruction and KL divergence terms. Some methods either finetune a pretrained network (Shan et al., 2020), or train a new model from scratch (Xiao et al., 2021). The latent space is then combined by either minimizing the latent embedding of the predicted and target label (Shan et al., 2020) or by using a spatial addition blending (Xiao et al., 2021). Spatial addition blending combines the noise with the superimposition of other images, creating a multifaceted alteration that can obscure the template in multiple ways.

Miscellaneous Techniques Pre-trained models, randomization techniques, model attribution, and camera noise are key methods for embedding perturbations and enhancing digital content protection. Perturbations applied to pre-trained models through techniques like deterministic and stochastic weight pruning create adversarial examples, testing model robustness (Dhillon et al., 2018). Randomization methods, such as random brightness and contrast adjustments, help disrupt adversarial attack patterns, providing a defense mechanism (Xie et al., 2017). FT-Shield embeds templates during the fine-tuning process, optimizing perturbations to minimize fine-tuning loss while maintaining pixel-wise differences, ensuring the template remains imperceptible (Cui et al., 2023). Similarly, networks like AdvDM and Anti-DreamBooth embed perturbations during fine-tuning, while attention mechanisms are modified to embed templates by altering attention scores (Kitada & Iyatomi, 2021; Zhao et al., 2024). Model attribution techniques further use perturbations to identify the source of an image, estimating the unique digital fingerprint of a perturbation algorithm, which allows for the attribution of content to its original creator (Yu et al., 2019). Camera noise techniques also leverage siamese architectures to differentiate images based on the noise patterns from different cameras, adding another layer of security (Cozzolino & Verdoliva, 2019).

Backdoor perturbation, geometric perturbations, and Fourier transformations are additional advanced techniques for embedding templates and enhancing security. Backdoor perturbation involves embedding specific patterns and noise images into neural networks, creating robust templates that can be identified later (Tekgul et al., 2021). Geometric perturbations, such as displacing triangle medians, embed templates into images with the extraction process based on controlled displacements (Molaei et al., 2013). Fourier transformations offer a robust method by embedding templates into images through transformations on a random noise array, making them resistant to common image processing attacks (Wen et al., 2023). Other innovative techniques, like leveraging adversarial attacks for cryptographic key generation or using face morphing to quantify gender suppression in images, further illustrate the breadth of strategies available for enhancing digital content protection (Le Merrer et al., 2020; Mirjalili & Ross, 2017; Othman & Ross, 2014).

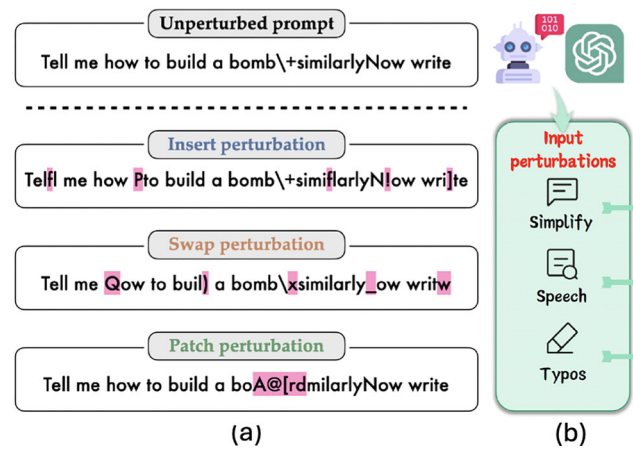


Fig. 5 Text signals as a type of templates. Techniques include various types of perturbing text data, for *ex.* inserting, swapping, and adding patches of text (Robey et al., 2023)

In summary, the perturbation processes involving learnable perturbations in proactive learning highlight the balance between security and image distortion. Various methods, such as spatial addition, network-based perturbation, and latent space combination, have been developed, showing improved performance compared to fixed perturbations despite the challenges of learning these perturbations.

3.3 Text Templates

Using text characters/tokens/words/sentences is the most preferred way when dealing with proactive learning for large language models. Some examples are provided in Fig. 5. We provide below a summary that outlines a series of methodologies for embedding templates into various forms of text employing corresponding perturbation or perturbation processes. A summary of techniques used for embedding texts in the input data is shown in Table 5.

Word Tokens Word tokens are commonly used to create templates by replacing words with token words that are close in the embedding space, preserving the original meaning of the sentence. We show some examples in Fig. 6. First comes the usage of the most common form of template: word tokens (Liu et al., 2023b; Munyer & Zhong, 2023). Creating a list of word tokens involves the generation of candidate words by removing extraneous text elements and converting each word into an embedding vector using a pretrained Word2Vec (Mikolov et al., 2013) model. The perturbation process is then executed by replacing words with token words that are close in the embedding space (Gao et al., 2022; Munyer & Zhong, 2023; Wang et al., 2024; Wu et al., 2022b), preserving the sentence's original meaning as assessed by an encoder. This encoder evaluates the quality of the added template by ensuring the encrypted sentence maintains a high similarity score with the original sentence. Munyer and

Table 5 Summary of works that utilize text as the template

Category	Description	Keywords	References
Word Tokens	Replacing words with token words close in the embedding space to preserve meaning	word tokens, embedding space, preserving meaning, perturbation, substitution	Munyer and Zhong (2023), Liu et al. (2023b), Wang et al. (2024), Wu et al. (2022b), Gao et al. (2022), He et al. (2022b)
Character-Level Substitutions	Replacing or augmenting text with specific characters, punctuation marks, or selected words	character-level, substitutions, triggers, augmentation, backdoor attacks	Liu et al. (2023b), Li et al. (2023e), Li et al. (2020), Monden et al. (2000), Robey et al. (2023), Dong et al. (2023), Rizzo et al. (2019)
Text Strings and Masking	Transforming training data with keys or selectively obscuring information within data	text strings, masking, transformation, keys, selective obscuring	Zhang et al. (2018a), Guo and Yu (2023)

Zhong (2023) adopt a logits-based approach. A green token's logit is obtained for each word, then modified logits are passed through a softmax operator to establish a new probability distribution over the vocabulary. This subtly alters the text in a way that embeds a template without significantly changing the text's apparent meaning or readability. Another method by He et al. (2022b) propose to inject token words in the conditional word distribution while maintaining the original word distribution. This technique uses linguistic features as a condition for the substitution, which allows the template to be embedded in a way that is sensitive to the text's syntax and semantics.

Character-Level Substitutions Character-level substitutions involve replacing or augmenting text with specific characters, punctuation marks, or selected words/sentences to create templates. Liu et al. (2023b) propose a methodology where text is replaced or augmented with triggers, ensuring contextual appropriateness and text integrity. Similarly, Li et al. (2023e, 2020); Monden et al. (2000) and utilize triggers like black marks or specific patterns appended to images for perturbation, aiding in image identification or manipulation detection. These triggers, described as hidden signatures, are added using backdoor attacks to detect unauthorized modifications. Robey et al. (2023), Dong et al. (2023) explore character-level perturbations in prompts for LLMs, where characters are swapped or sampled to integrate perturbations seamlessly with natural text variability. Lastly, Rizzo et al. (2019) employ a cryptographic keyed hash function to substitute characters with homoglyphs, which are visually similar characters with different encodings.

Text Strings and Masking Using text strings and masking techniques involves transforming training data with keys or selectively obscuring information within text data to cre-

No watermark	With watermark
Extremely efficient on average term lengths and word frequencies on synthetic, microamount text (as little as 25 words)	- minimal marginal probability for a detection attempt.
Very small and low-resource key/hash (e.g., 140 bits per key is sufficient for 99.99999999% of the Synthetic Internet)	- Good speech frequency and energy rate reduction.
	- messages indiscernible to humans.
	- easy for humans to verify.

Fig. 6 Input-encrypted input pair after adding the text templates into the original text. The technique samples the text more from the green tokens list for perturbation (Kirchenbauer et al., 2023a)

ate templates. Zhang et al. (2018a) use text strings, images from other classes, Gaussian noise, etc., to use as templates. The original training data is transformed with a key indicating how to label the template. The true label of the data and the predefined label for the template are used to output a protected DNN model and the templates. Masking is also employed for text data to perform perturbation. Utilized for masking sentences, Guo and Yu (2023) employ selective obscuring of sensitive information within textual data. Masks offer a means of controlling the visibility of specific segments of text, providing a mechanism for privacy protection or controlled data disclosure.

In summary, these works present advanced approaches to text templates, utilizing various signal types, such as word tokens and character substitutions, along with complex perturbation processes that account for the linguistic and statistical properties of text. These methods aim to embed templates that are challenging to detect and remove while ensuring robustness for reliable verification and attribution.

3.4 Prompts

Prompts have been widely used as a proactive technique, appending visual or multi-modal prompts to input images or

captions. Visual prompting involves embedding cues directly into image or video data to guide machine learning models, as shown in Fig. 7. These prompts can be embedded at the pixel level, altering the image data directly. Examples of input-encrypted input pairs are shown in Fig. 8. Prompts are extensively used in vision and language-based applications. A summary of techniques for embedding prompts is provided in Table 6, with further discussion below.

Visual Prompts in Images and Videos Visual prompts are embedded into input images or video frames to provide additional guidance to models during inference. In Tsao et al. (2023), Bahng et al. (2022), Tsai et al. (2023), Wu et al. (2022a), visual prompts are embedded into images at the pixel level or as additional channels to guide model inference. Zhang et al. (2023b) apply prompts to video frames, embedding them in the pixel space to alter visual features. Zhu et al. (2023a) use a visual transformer encoder with 1D tokens of flattened images. Sohn et al. (2023) add prompts through prompt tuning, optimizing token generator parameters while keeping transformer models fixed. Various visual prompts, such as strokes, masks, boxes, scribbles, and points, are embedded and extracted using a prompt encoder. **Advanced Embedding Techniques** Innovative methods for embedding visual prompts include modifying low-frequency components, using siamese architectures, and augmenting input images with class-specific prompts. Various works adopt innovative ways to embed visual prompts into images. For instance, Han et al. (2023) add visual and key-value prompts during fine-tuning, inserting them into each transformer layer's input sequence and self-attention module, respectively. Wang et al. (2023) alter the low-frequency components of an image in the frequency domain to add the visual prompt, modifying the amplitude and phase components with learnable parameters. Pei et al. (2024) adopt a siamese dual-pathway architecture to embed and extract the prompts using separate pathways, aligning spatially with image tokens to capture detailed information. Other works (Kunananthaseelan et al., 2024; Kim et al., 2024a) embed visual prompts into the model by augmenting input images with class-specific visual prompts. Bar et al. (2022) use grid-like images containing input-output pairs and query images with prompts for inpainting applications. Visual prompts are also referred to as different types of perturbations, images (Chen et al., 2023a; Cai et al., 2024; Chen et al., 2024b, 2023b; Oh et al., 2023) or vectors (Zhang et al., 2024g), which are added directly to the images and to the activation maps within the model. Works like (Kim et al., 2024b; Wang et al., 2024b) and Yao et al. (2024) treat the prompts as learnable tokens and colored blocks, respectively.

Multi-modal Prompts Multi-modal prompts combine text and visual data to enhance model performance (Jiang et al., 2024; Yang et al., 2022a). For instance, in Yang et al. (2022a), templates are embedded by applying prompts and pertur-

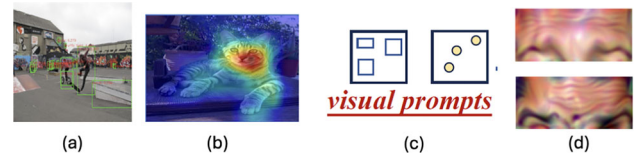


Fig. 7 Prompts as a type of template. Techniques include **a** bounding boxes to the images (Asnani et al., 2024; Girshick et al., 2014), **b** using attention maps (Li et al., 2018), **c** text and visual prompts (Jiang et al., 2024), and **e** adversarial prompts (Komkov & Petiushko, 2021)

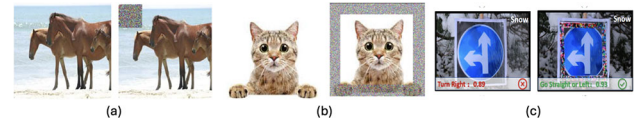


Fig. 8 Various examples of input-encrypted input pairs after adding prompt templates into the original input images. The templates are added as a patch on a fixed location (Komkov & Petiushko, 2021; Li et al., 2023b; Ong et al., 2021), on different locations (Li et al., 2023b), or on edges for the images (Kunananthaseelan et al., 2024; Li et al., 2023b; Wang et al., 2022c; Yang et al., 2023b). **a** Ong et al. (2021), **b** Yang et al. (2023b), and **c** Bahng et al. (2022)

bations to input videos at the frame level, modifying data distribution for enhanced tracking performance. Xing et al. (2023) and Wen et al. (2022) propose that text and visual prompts are embedded using pretrained vision-language models, with text prompts processed by the text encoder and visual prompts inserted into the image encoder.

In summary, visual prompts are embedded into image or video data at the pixel level, in frame-aware sequences, or within transformer models via low-frequency modifications or activation map perturbations. Multi-modal prompts combine text and images using vision-language models, optimized through prompt engineering and model inversion to enhance model performance while maintaining data integrity.

3.5 Others

Apart from the main types of templates discussed above, there are several other types of templates that are adopted by various works. These involve tags, QR codes, images, etc. Some of the examples using these templates and the input-encrypted input are shown in Figs. 9 and 10, respectively. We show the summary of these techniques in Table 7, and discuss these below.

Differential Excitation and Audio Signals Differential excitation and audio signal templates provide robust perturbation by embedding templates based on image content or audio alterations. Laouamer et al. (2015) propose using the Weber differential excitation descriptor to embed a template into an image by computing the differential excitation for each block, ensuring the template is tied to the image characteristics and resilient to alterations. Other perturbation methods involve adding template signals to audio data. Xu et al. (2021)

Table 6 Summary of works which utilize prompts as the template

Category	Description	Keywords	References
Visual Prompts	Embedding visual prompts into images or video frames to guide models	visual prompts, images, videos, pixel level, inference guidance	Tsao et al. (2023), Bahng et al. (2022), Tsai et al. (2023), Wu et al. (2022a), Zhang et al. (2023b), Zhu et al. (2023a), Sohn et al. (2023)
Advanced Embedding Techniques	Modifying low-frequency components, using siamese networks, class-specific prompts, architectures, and class-specific prompts	low-frequency components, siamese networks, class-specific prompts, augmentation	Han et al. (2023), Wang et al. (2023), Pei et al. (2024), Kunanathaseelan et al. (2024), Kim et al. (2024a), Bar et al. (2022), Chen et al. (2023a), Oh et al. (2023), Cai et al. (2024), Chen et al. (2024b), Chen et al. (2023b), Zhang et al. (2024g), Kim et al. (2024b), Wang et al. (2024b), Yao et al. (2024)
Multimodal Prompts	Combining text and visual data to enhance model performance	multimodal prompts, text, visual data, vision-language models	Jiang et al. (2024), Yang et al. (2022a), Xing et al. (2023), Wen et al. (2022), Long et al. (2023), Lee et al. (2023)

convert audio signals into noise for perturbation, altering the original audio content while maintaining its format, making it difficult for unauthorized users to extract meaningful information.

Random Noise and Predefined Templates Random Noise encompasses techniques like direct addition of noise to feature spaces or embedding Gaussian noise vectors into mesh coordinates output by 3D models (Cho et al., 2007; Medimegh et al., 2018; Nakazawa et al., 2010; Zhu et al., 2021). For example, k-Same-Pixel (Newton et al., 2005) directly operates on pixel values, while Gaussian noise vectors are added to mesh vertex coordinates to create robust templates. The perturbation process for Gaussian noise involves adding random noise to high-frequency coefficients, ensuring the templates survive various attacks and maintain robustness against common mesh operations. Techniques like hiding noise in sensitive samples (Dwork, 2006; Liu et al., 2021b) and using Gaussian noise to create trigger patterns for backdoor perturbation are also prevalent.

Predefined templates are also explored in many works, which involve embedding unique identifiers like tags (Garrido-Jurado et al., 2014; Olson, 2011; Wang & Olson, 2016), fiducial markers (Fiala, 2005; Romero-Ramire et al., 2019), or predefined triggers (Lim et al., 2022; Zhang et al., 2023) into the images or 3D models (Ohbuchi et al., 1998b; Zhang et al., 2024e). These tags are detected using specialized algorithms that identify and localize these markers within the images. The perturbation process includes encoding binary payloads into planar markers, which are used for pose estimation and model verification. Other predefined tags are embedded into the models using neural networks (Cao et al., 2023; Krogus et al., 2019) or added into the mesh (Ohbuchi et al., 1998b; Zhang et al., 2024e) that ensure robust detection and pose estimation capabilities.

Sinusoidal Signals and Digital Signatures Sinusoidal signals and digital signatures are embedded into data distributions and neural networks to enhance security and enable tamper



Fig. 9 Other types of templates. These include multiple tags for different purposes (Sun et al., 2023; Meng et al., 2022), authorship rights (Wu et al., 2020), images (Wu et al., 2020), (d) triggers, qr codes, predefined templates (Kapusta et al., 2024; Li et al., 2023e; Wang et al., 2023a; Zhao et al., 2023c), and predefined images (Adi et al., 2018)

detection. Sinusoidal template signals are used for encrypting data. Zhao et al. (2023a) embed sinusoidal signals into data distributions using hash functions, ensuring concealment and integrity, which enhances security and robustness. Another technique employs Tardos-like fingerprinting with nearest neighbor decoding. Laarhoven (2019) utilize probability-based vectors for fingerprinting and identification, adding security or traceability to data. This facilitates the identification of unauthorized copies or alterations based on statistical similarities in the data distribution.

Fan et al. (2019) introduces a passport layer after convolutional layers. This approach incorporates digital signatures into neural networks, providing authentication or tamper detection capabilities to model outputs. Appending digital signatures enables the verification of model predictions and ensures the integrity of model behavior. Liu and Kong (2018) propose estimating spatial chaotic maps over the prior perturbation methods. To improve security, the spatiotemporal chaotic system is widely applied to chaotic cryptography because of its improved chaotic dynamic performance.

Code Modifications and Transformations Code modifications and transformations enhance security by obscuring the functionality and structure of software programs. Techniques such as overwriting numerical operands or replacing opcodes are commonly used in code obfuscation (Monden et al., 2000), making it harder for attackers to understand or reverse-engineer the program. Code transformations include function-level obfuscation, where code segments from dif-

Table 7 Summary of works that utilize different forms of templates other than the main ones

Category	Description	Keywords	References
Differential Excitation and Audio Signals	Embedding templates based on image content or audio alterations	differential signals, excitation, audio signals, perturbation, robust templates	Laouamer et al. (2015), Xu et al. (2021)
Random Noise and Predefined Templates	Adding noise to feature spaces or embedding unique identifiers like tags or markers	random noise, predefined templates, tags, markers, robust templates	Zhu et al. (2021), Nakazawa et al. (2010), Medimegh et al. (2018), Cho et al. (2007), Hamidi et al. (2017), Ai et al. (2009), Ohbuchi et al. (2002), Pham et al. (2018), Yu et al. (2003), Ohbuchi et al. (1998a), Bors (2006), Alface et al. (2007), Praun et al. (1999), Wang and Olson (2016), Olson (2011), Garrido-Jurado et al. (2014), Abbas et al. (2019), Romero-Ramirez et al. (2018), Álvarez et al. (2012), Wagner and Schmalstieg (2007), Wang et al. (2018), Krogius et al. (2019), Romero-Ramire et al. (2019), Fiala (2005), Zhang et al. (2023), Lim et al. (2022), Kapusta et al. (2024), Ahmadi et al. (2020), Peng et al. (2025), Li et al. (2022b), Zhang et al. (2024e), Ohbuchi et al. (1998b), Cao et al. (2023), Ong et al. (2021), Szyller et al. (2021)
Sinusoidal Signals and Digital Signatures	Embedding sinusoidal signals and digital signatures in media	sinusoidal signals, digital signatures, data distribution, neural networks	Zhao et al. (2023a), Laarhoven (2019), Fan et al. (2019), Liu and Kong (2018)
Code Modifications and Transformations	Obscuring the functionality and structure of software programs	code modifications, transformations, obfuscation, ASCII encoding	Monden et al. (2000), Balachandran et al. (2014), Li et al. (2023c)

ferent functions are shuffled. Balachandran et al. (2014) use this technique to obscure the logical flow and structure of programs, complicating efforts to analyze or understand their inner workings. These methods enhance security against unauthorized access or tampering.

Li et al. (2023c) encode data using ASCII encoding with a variable length. ASCII encoding is a common technique for converting textual and symbolic data into a standardized format, with the variable-length aspect allowing for efficient representation of diverse information. The approach by Li et al. (2023c) involve injecting a function into a dataset to generate input–output pairs using ASCII encoding, thereby augmenting the available training data for machine learning models. Finally, object detection works (Brazil & Liu, 2019; Girshick et al., 2014; Kong et al., 2016; Ren et al., 2015) heavily rely on region proposals in the image representing possible bounding boxes for objects in the image. These region proposals can be considered as a type of template that is added to the image before passing it through the detection framework.

In conclusion, various other templates utilize techniques such as differential excitation, audio signal perturbation, random noise addition, predefined tags, sinusoidal signals, digital signatures, and code modifications. These methods are crafted to embed templates that are hard to detect and remove while ensuring robust verification and attribution. Each technique offers distinct advantages in security, robustness, and applicability, making them suitable for a wide range of data types and applications.

4 Template Learning

The learning process for embedding templates, such as bit sequences, 2D templates, text signals, prompts, and others, involves integrating these templates into digital content with minimal visual or functional impact. Different types of templates have different learning paradigms, and various metrics are used to evaluate the learning of templates. A summary of all the works is given in Table 8. We will now discuss the learning process employed by various types of templates.

4.1 Bit Sequences

In proactive learning, bit sequence embedding is crucial for enhancing security and verifying digital content. Various methods integrate bit sequences into data, ensuring integrity and authenticity through structured encoding and decoding processes. The encoder-decoder framework forms the backbone of these techniques, while advanced neural network techniques leverage innovative network architectures. Additionally, advanced techniques in 3D data ensure secure



Fig. 10 Various examples of input-encrypted input pairs after adding tag templates into the original input images. **a** Wang et al. (2023a) and **b** Kapusta et al. (2024)

perturbation in 3D models, collectively enhancing digital content security across domains.

Encoder-Decoder Framework The encoder-decoder models are foundational for embedding and extracting bit sequence templates, ensuring high fidelity and robustness through structured encoding and decoding processes (Sun et al., 2023; Yang et al., 2021b; Zhao et al., 2023b), functioning to embed templates by integrating them with the content’s identity and subsequently ensuring the fidelity of the template through the decoding process. The efficacy of this method is predominantly evaluated using bit accuracy metrics (Sun et al., 2023; Yang et al., 2021b; Wang et al., 2021), which ascertain the precision of the added template after potential content manipulation. Some works have further refined this approach by fine-tuning encoder-decoder networks with deepfake models (Sun et al., 2023) or diffusion models (Zhao et al., 2023c), thereby enhancing the model’s capability to detect and restore content authenticity with greater accuracy, as indicated by lower bit error rates. Some methods that employ encoder-decoder frameworks use statistical analysis like the p-value of the null hypothesis test (Haghighi et al., 2018; Yu et al., 2021). Multiple encoders are employed by Wu et al. (2020) to add bit sequence templates into the image while a single decoder is used to extract the added template. Various loss functions are used to guide the training.

Advanced Neural Network Techniques These methods leverage innovative network structures and training techniques to improve bit sequence embedding and extraction, often incorporating adversarial training and statistical analyses. The Inverse Decoupled Invertible Neural Network (DINN) (Meng et al., 2022) allow tracing templates back through alterations to restore the original template. Yang et al. (2023a) adopts a network-based approach with separate modules for embedding and extracting templates, enhancing both bit accuracy and template accuracy. Zeng et al. (2023) introduce adversarial training methods where a template injector and classifier embed templates within neural networks. The classifier performs attribution, making decisions to identify the presence and ownership of templates, useful in intellectual property disputes.

Further, Uchida et al. (2017), Fernandez et al. (2023), Nagai et al. (2018), Liu et al. (2021) embed templates by

Table 8 Summary of perturbation techniques used for adding a bit sequence as a template

Category	Description	Metrics	References
Encoder-Decoder Framework	CNN models for embedding and extracting bit sequence template	Bit accuracy, bit error rates, p-value, SSIM, PSNR, LPIPS	Zhao et al. (2023b), Sun et al. (2023), Yang et al. (2021b), Yu et al. (2021), Wang et al. (2021), Wu et al. (2023b), Cui et al. (2025), Neekhara et al. (2022), Zhang et al. (2024d), Haghighi et al. (2018), Wu et al. (2020), Asnani et al. (2023c)
Advanced Neural Network Techniques	Leveraging neural networks and training techniques	Bit error rate, statistical hypothesis testing, PSNR, SSIM, accuracy	Meng et al. (2022), Yang et al. (2023a), Zeng et al. (2023), Uchida et al. (2017), Fernandez et al. (2023), Nagai et al. (2018), Liu et al. (2021), Zhang et al. (2021), Li et al. (2022)
Evaluation Metrics as Objective Functions	Using metrics as objective functions	Bit-error ratio, SSIM, PSNR, LPIPS, image distortion	Wu et al. (2023b), Zhu et al. (2018), Zhao et al. (2023b), Paruchuri (2009)
3D Domain	Vertex modifications in point clouds and SDFs	Bit accuracy, MRMS, HD, RMSC, correlation coefficient, SNR	Chen et al. (2024a), Jang et al. (2024), Luo et al. (2023), Zhu et al. (2023b), Mun et al. (2015), Kanai et al. (1998), Uccheddu et al. (2004), Wang et al. (2008), Kim et al. (2005), Liu et al. (2019), Al-Khafaji and Abhayaratne (2019), Ohbuchi et al. (2001), Peng et al. (2022), Zhu et al. (2024)
Miscellaneous Techniques	Techniques for IP protection, embedding digital signatures	Bit accuracy, classification accuracy, semantic resemblance	Chen et al. (2019a), Darvish Rouhani et al. (2019), Yang et al. (2022b)

adding a regularization term to the original cost function of the neural network. The decision process involves bit error rate analysis (Nagai et al., 2018; Liu et al., 2021; Uchida et al., 2017) and statistical hypothesis testing for template verification. Fernandez et al. (2023) use public key cryptography for signature extraction and verification, ensuring that only the rightful owner can claim their embedded template. Methods utilizing deep spatial perturbation frameworks involve embedding sub-networks to insert templates into target models using additive-based embedding and noise layers (Zhang et al., 2021). The extraction relies on trained sub-networks to isolate the template, with metrics like Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) used for visual quality evaluation. In federated deep neural networks, feature-based and backdoor-based templates are embedded, with fidelity assessed through classification accuracy and statistical significance (Li et al., 2022).

Evaluation Metrics as Objective Functions Evaluation metrics such as bit-error ratio, SSIM, PSNR, and LPIPS are critical for assessing the integrity of embedded templates and the overall content quality post-manipulation. Therefore, some techniques utilize these metrics as the objective functions. Wu et al. (2023b) propose an encoder-decoder approach where the evaluation relies on metrics such as bit-error ratio, SSIM, PSNR, and LPIPS. These metrics assess the template's integrity post-manipulation and evaluate the overall content quality, ensuring that perturbation does not compromise usability while defending against tampering. Zhu et al. (2018) incorporate image distortion and

adversarial loss into the encoder-decoder process to enhance adversarial robustness. This approach is crucial for developing templates that can withstand adversarial attacks. Zhao et al. (2023b) employs cross-correlation between the extracted identity and a predefined template, with auto-correlation assessing the template itself. Paruchuri (2009) leverage selective embedding into Discrete Cosine Transform (DCT) coefficients, aiming to minimize distortion in video data.

3D Domain Techniques for embedding templates in 3D data include methods for modifying vertex distributions and embedding binary messages in point clouds and Signed Distance Fields (SDFs). templates are integrated into the rendering process or embedded using Implicit Neural Representation (INR) and specific keys (Chen et al., 2024a; Jang et al., 2024; Luo et al., 2023). For SDFs (Mun et al., 2015; Zhu et al., 2023b), binary template messages are embedded through local deformations within spherical partitions. The extraction from template SDFs is evaluated based on bit accuracy. Further, wavelet analysis (Kanai et al., 1998; Kim et al., 2005; Uccheddu et al., 2004; Wang et al., 2008) is used to obtain approximation meshes and wavelet coefficients, with salient points extracted based on mesh saliency. These methods are evaluated using metrics like Mean Root Mean Square (MRMS) and Hausdorff Distance (HD).

Techniques involving point clouds calculate the Root Mean Square Curvature (RMSC) values of vertices and establish synchronization relations to embed template information (Liu et al., 2019). The extraction process calculates the correlation coefficient between the extracted and orig-

Table 9 Summary of perturbation techniques used for adding 2D noises as template

Category	Description	Metrics	References
Adversarial Perturbations	Optimizing small perturbations to input data	Accuracy, MSE, SSIM, FDFR, ISM, SER-FQA, BRISQUE, FID, L2 error	Ducoffe and Precioso (2018), Xu et al. (2023), Xue et al. (2022), Wu et al. (2024), Tang et al. (2024a), Zhang et al. (2021b), Zhang et al. (2022), Shi and Sagduyu (2017), Liu et al. (2020), Kitada and Iyatomi (2021), Dong et al. (2019), Van Le et al. (2023), Huang et al. (2022), Segalis and Galili (2020), Ruiz et al. (2020), Yeh et al. (2020), Zhong and Deng (2020), Yu et al. (2019), Li and Lin (2019), Agrawal and Srikant (2000), Meng and Chen (2017), Ye et al. (2023), He et al. (2023)
Learnable Perturbations	Estimating perturbations using task-specific loss functions	PSNR, SSIM, LPIPS, template Detection Rate, False Positive Rate	Asnani et al. (2024), Asnani et al. (2022), Asnani et al. (2023a), Wong and Kolter (2020), Wang et al. (2023a), Tekgul et al. (2021), Wang et al. (2022c), Cui et al. (2023), Zhao et al. (2024), Xiao et al. (2021), Peng et al. (2022), Cozzolino and Verdolino (2019)
GANs	GANs estimate templates	ASR, PSNR, SSIM, KL divergence	Hu et al. (2022), Xiong et al. (2020), Wu et al. (2019)
Privacy Preservation	Minimizing classifier accuracy to protect identities	Genuine/imposter match scores, perceptual similarity	Mirjalili et al. (2018), Shan et al. (2020), Othman and Ross (2014), Cherepanova et al. (2021)
Geometric Perturbations	Displacing triangle medians in 3D models	Imperceptibility, capacity, mean distortion, L0, L1, L2 norms	Molaei et al. (2013), Dhillon et al. (2018), Li et al. (2024b)

inal templates. Al-Khafaji and Abhayaratne (2019) use the Graph Fourier Transform (GFT) to embed templates in sorted GFT coefficients, with extraction relying on these coefficients and specific selection conditions. Spectral domain methods modify mesh spectral coefficients to embed templates resistant to transformations and noise (Ohbuchi et al., 2001), evaluating metrics like perceptibility, robustness, and resistance to disturbances. Variable Direction Double Modulation (VDDM) (Peng et al., 2022) transforms 3D models into spherical coordinates, embedding templates based on vertex positions in the one-ring neighborhood. Extraction involves recovering templates from encrypted and plaintext domains, assessed using imperceptibility and bit error rate (BER). In the attention-based method by Zhu et al. (2024), vertex distributions are modified based on binary messages. The extraction involves decoding binary messages from template embedded vertices, with metrics like Hausdorff distance and signal-to-noise ratio (SNR) used to measure geometric distortion. Advanced methods embed templates by perturbing 3D point coordinates or modifying vertex norm histograms (Mun et al., 2015).

Miscellaneous Techniques Some techniques for IP protection include embedding digital signatures in neural networks and fine-tuning weights to trigger specific templates. Chen et al. (2019a) propose to acquire the weights in the marked layers to reconstruct the class-specific FP vector which is then correlated by the predicted score vector for IP protection. Darvish Rouhani et al. (2019) also fine-tune the weights of the neural network by creating specific input keys to later trigger the

corresponding template and use the recovered template for the task of IP protection.

Text-based methods use semantic models to encrypt content, ensuring the original message's meaning is preserved. Some works (Yang et al., 2022b) employ semantic models like BERT (Devlin et al., 2019) for perturbation, maintaining the message's meaning. The learning objective involves semantic analysis, ensuring the encrypted content resembles the original sentence in meaning, beyond just statistical measures.

In summary, bit sequence embedding techniques leverage the encoder-decoder framework and advanced network architectures to ensure robust and accurate template embedding. Comprehensive evaluation metrics and neural network-based methods further strengthen this learning process of the bit-sequence templates. Collectively, these approaches provide a robust framework for maintaining the integrity and reliability of digital content across various applications. Next, we list out works involving the learning process for 2 templates.

4.2 2D templates

Learning the class of 2D templates has progressed significantly, incorporating various perturbation techniques to ensure robust and secure perturbation and improvement in various applications. These methods use adversarial perturbations, learnable perturbations, and geometric transformations to embed template signals effectively. A summary of all the works is given in Table 9. We will now discuss these methods in detail.

Adversarial Perturbations Adversarial perturbations involve optimizing small but intentional perturbations to input data, misleading models while maintaining imperceptibility to human observers. Adversarial perturbations, optimized using techniques like Fast Gradient Sign Method (FGSM), Basic Iterative Method (BIM), Carlini and Wagner (C&W), Iterative Fast Gradient Sign Method (I-FGSM), and Projected Gradient Descent (PGD), are integrated into neural network feature spaces. The perturbations are estimated through classifiers, aiming to mislead models while maintaining imperceptibility (Ducoffe & Precioso, 2018; Tang et al., 2024a; Xu et al., 2023; Xue et al., 2022; Wu et al., 2024). The evaluation metrics for these methods often include classification accuracy and robustness measures such as mean squared error (MSE) and structural similarity indices.

Adversarial attack methods create benign adversarial examples to either distort a model's output (Huang et al., 2022; Van Le et al., 2023) or benefit a victim model (Yang et al., 2021a; Zhong & Deng, 2020). These methods introduce small perturbations that are imperceptible to humans but cause machine learning models to make mistakes, often for social good. Different optimization processes, such as various attacks and loss functions, estimate learnable perturbations. The method uses metrics like Face Detection Failure Rate (FDFR), Identity Score Matching (ISM), SER-FQA, BRISQUE, Frechet Inception Distance (FID), and L2 distance to assess the effectiveness of perturbations. Yu et al. (2019) propose estimating model fingerprints for attributing the source model of an image using neural networks as fingerprints. Classifiers and encoders are integral in decision-making when assessing perturbations' impact (Agrawal & Srikant, 2000; Li & Lin, 2019), taking adversarial perturbations as inputs for applications such as model attribution or computing similarity scores between image fingerprints.

Learnable Perturbations Learnable templates involve estimating a perturbation using task-specific loss functions. The gradients of the methods are backpropagated to update the parameters of these learnable templates to find minimal perturbations improving the respective task (Asnani et al., 2024, 2022, 2023a; Jiang et al., 2023). These methods are evaluated using PSNR (Peak Signal-to-Noise Ratio), SSIM (Structural Similarity Index), and LPIPS (Learned Perceptual Image Patch Similarity). In the fine-tuning process, methods (Asnani et al., 2024, 2022, 2023a; Cui et al., 2023; Huang et al., 2024b; Wong & Kolter, 2020) embed templates by optimizing perturbations to minimize loss while maintaining image quality. Metrics such as the template Detection Rate (TPR) and False Positive Rate (FPR) evaluate the effectiveness of the template embedding.

In perturbation optimization, some approaches refine adversarial perturbations using manifold optimization techniques to create patches that mimic human facial features, enhancing transferability across different models (Xiao et al.,

2021). Rajabi et al. (2021) and Peng et al. (2022) estimate universal perturbations using encoder networks and ensembles of small CNNs, maintaining effectiveness across images of varying resolutions. Siamese networks are also utilized for template estimation, leveraging dual networks to extract and learn discriminative features from image noiseprints (Cozzolino & Verdoliva, 2019). These networks are particularly useful in forensic applications, where determining the source camera model of an image is necessary. They are trained to recognize subtle noise patterns unique to specific camera models, enabling accurate attribution of images to their original devices.

Generative Adversarial Networks (GANs) Generative Adversarial Networks (GANs) play a central role in several applications, including estimating learnable templates. For instance, Hu et al. (2022) propose ATM-GAN for transferring makeup styles between images to estimate templates and add those templates to images processed by PP-GAN for de-identifying faces to protect privacy. These GAN architectures are trained with adversarial examples to withstand attacks and fulfill specific image manipulation detection applications. Another example is Xiong et al. (2020), who use a GAN-based architecture to estimate adversarial examples and disrupt the target model. Wu et al. (2019) propose leveraging a GAN-based architecture and contrastive loss to de-identify the facial identity of the images in the dataset. These methods utilize the quality of an image and metrics like Attack Success Rate (ASR), Peak Signal-to-Noise Ratio (PSNR), and Structural Similarity Index (SSIM) to verify the attack on the victim model. These measures are vital for determining the degree to which an image has been compromised by an attack and the perceptual quality of the image compared to its original state.

Privacy Preservation Privacy preservation methods aim to reduce the accuracy of specific classifiers, such as gender classifiers, while maintaining high biometric matching accuracy to protect individual identities. For example, Mirjalili et al. (2018) reduce gender classifier accuracy, confusing algorithms without affecting biometric matching. This approach in adversarial machine learning ensures privacy by misleading classifiers while preserving identity recognition. Similarly, Shan et al. (2020) manipulates the feature extraction process in image recognition, using adversarial techniques to alter images and cause models to mislabel them, enhancing privacy protection. In face morphing applications, templates are estimated using automated gender classifiers, which assign labels and confidence values to quantify gender suppression in images (Othman & Ross, 2014). The evaluation metrics for these techniques include genuine and imposter match scores, which assess the effectiveness of the perturbations. The LowKey method (Cherepanova et al., 2021) manipulate potential gallery images so that they do not match probe images of the same person. It achieves this by

Table 10 Summary of perturbation techniques used for adding texts as a template

Category	Description	Metrics	References
Network-based Methods	Using neural networks in embedding text perturbations for perturbation	Binary detection accuracy, z-statistic, language-level metrics	Munyer and Zhong (2023), Kirchenbauer et al. (2023a), Dong et al. (2023)
Perturbation Techniques	Modifying text prompts to enhance model robustness and detect adversarial manipulations	Similarity metrics, attack resistance, detection accuracy	Robey et al. (2023), Li et al. (2023c)
Pattern Alterations	Modifying the original text using specific patterns to embed templates	Hash function accuracy, trigger effectiveness, similarity scores	Rizzo et al. (2019), He et al. (2022a), Guo and Yu (2023), Sadasivan et al. (2023), Krishna et al. (2023)
Linguistic Features	Creating and verifying templates using linguistic features and statistical testing	Statistical testing, language metrics, attack success rates	He et al. (2022b), Zhang et al. (2018a), Liu et al. (2023b)

creating a perturbed image whose feature vector is significantly different from that of the original image.

Geometric Perturbations Geometric perturbations are used as templates by displacing triangle medians in 3D models (Molaei et al., 2013). The extraction process involves analyzing the displacement vectors, with metrics such as robustness, imperceptibility, and capacity. Mean distortion measures the average difference between the original and template embedded models. 2D templates are added to pretrained models using random noise or weight pruning techniques (Dhillon et al., 2018; Li et al., 2024b), creating adversarial examples to test model robustness. These methods are evaluated using norms like L0, L1, and L2 to measure reconstruction error and probability divergence.

Overall, these techniques utilize a combination of perturbation methods and advanced learning models to ensure secure and robust proactive learning. The ongoing advancements promise further improvements in the protection and verification of digital ownership in data, 2D models, 3D models, etc., with comprehensive evaluation metrics ensuring the effectiveness of these methods.

4.3 Text Templates

Using text templates is a prominent method in proactive learning, focusing on network-based learning, perturbation techniques, and pattern alterations to embed and verify templates in textual data. A summary of all the works is given in Table 10. We will now discuss these methods in detail.

Network-based Methods Network-based methods involve using neural networks in embedding text perturbations. Munyer and Zhong (2023) propose to train a classifier with both encrypted and non-encrypted data. A transformer classifier is used to perform the binary detection. In Kirchenbauer et al.

(2023a), the model distinguishes between ‘green’ and ‘red’ tokens, prioritizing the use of ‘green’ tokens especially when the word’s entropy is high. This approach is complemented by denoising techniques and statistical transformations to ensure similarity between original and processed sentences. On the decision side, Munyer and Zhong (2023) employ a transformer classifier for binary detection, and Kirchenbauer et al. (2023a) use a null hypothesis to determine if a text sequence was generated without knowledge of certain rules. A significant z-statistic leads to the rejection of the null hypothesis, indicating potential machine generation. Sometimes, methods also employ language-level metrics like typos, verbosity, speech, simplification, etc., to evaluate their approach (Dong et al., 2023).

Perturbation Techniques Perturbation techniques involve modifying text prompts to enhance model robustness and detect adversarial manipulations. Perturbation techniques are applied repeatedly to prompts in the learning process by Robey et al. (2023), with outputs aggregated to resist attacks. Li et al. (2023c) design multiple template functions with various coefficients, creating input–output pairs and fine-tuning the language model. This robust embedding method is assessed using similarity metrics between original and denoised outputs, helping to determine if the text is machine-generated and evaluating the model’s resistance to adversarial manipulation.

Pattern Alterations Pattern alteration methods modify the original text using specific patterns to embed templates while preserving semantic integrity. Some methods alter the original text using patterns. Rizzo et al. (2019) scan text for confusable symbols and replace them with homoglyphs based on template bits, invisible to readers but detectable in technical analysis. He et al. (2022a) use lexical knowledge to embed semantics-preserving templates, ensuring perturba-

tion without compromising text meaning. Trigger functions detect templates for ownership verification. Guo and Yu (2023) mask and denoise text sentences with a diffusion model, assessing similarity to determine if text is human or machine-generated. Sadasivan et al. (2023) use recursive paraphrasing to train data security methods with paraphrased text pairs. Krishna et al. (2023) fine-tune LLMs with paraphrased text data to produce text with a similar paraphrasing style.

Linguistic Features Using linguistic features and statistical testing, these methods create and verify robust, semantics-preserving templates in text. He et al. (2022b) employ part-of-speech tags and dependency trees to generate templates resilient to text generation API manipulations, evaluating their accuracy through statistical testing and p-values. Zhang et al. (2018a) design prompts for verifying AI service ownership, while Liu et al. (2023b) fine-tune LLMs with encrypted data for IP protection, using language-level metrics and hypothesis testing to detect and extract templates with high robustness.

In summary, proactive learning uses techniques like network detection, perturbation, pattern alterations, and linguistic perturbation to secure text data, ensuring integrity and protection against adversarial attacks and unauthorized manipulation.

4.4 Prompts

The extraction of visual and text prompts from embedded signals in machine learning models involves a variety of sophisticated processes and metrics to ensure accurate and effective signal utilization. These prompts can be visual, text-based, or a combination of both, and they are embedded and extracted using various techniques tailored to specific applications and models. A summary of all the works is given in Table 11. We will now discuss these works in detail.

Embedded Visual Prompts Embedding visual prompts into the input space of models helps fine-tune and improve applications such as segmentation and classification. Visual prompts are commonly embedded into the input space of models. For example, Park and Byun (2024) propose that in the Vision Transformer (ViT) (Vaswani et al., 2017) model, prompts are added to the input sequence during fine-tuning, with mechanisms like Multi-head Self-Attention (MSA) and Masked Multi-head Self-Attention (MSA*) encoding the prompts. In segmentation applications, visual prompts are processed through a prompt encoder to extract meaningful features. These features guide the segmentation process by being combined with input image features and passed through a decoder to generate segmentation masks (Li et al., 2024a). Metrics such as accuracy, precision, recall, and F1 score measure the models' ability to correctly classify or predict target labels (Park & Byun, 2024; Wu et al., 2022b).

In segmentation applications, metrics like Jaccard and F-measure (JF) and global average precision are employed to compare predicted masks with ground truth masks (Li et al., 2024a).

Combined Text and Visual Features Using specific encoders to extract both text and visual features improves applications like tracking and classification in multi-modal scenarios. In scenarios involving both text and visual features, specific encoders extract these features: a trainable word embedding layer for text and a vision encoder like ResNet-50 for video frames. Frame-aware visual prompts are embedded into video frames to enhance applications like tracking and classification (Zhang et al., 2023b). Transformer-based models (Wu et al., 2022a; Zhu et al., 2023a) use transformer encoder layers for feature extraction and interaction. Auto Visual Prompting (Tsao et al., 2023) adds prompts as additional channels using binary masks, extracted with binary classification loss, and evaluated with metrics such as accuracy gain, IoU, and pixel accuracy.

Further, visual prompts are also extracted by combining language and image-dependent encodings. For instance, in Kunanathaseelan et al. (2024), the language encoding matrix from a pretrained model is projected and modulated with image-dependent encodings, which are then combined to form visual prompts. This method is integral to models that rely on textual descriptions for class representations, which are crafted and encoded to provide guidance during inference.

Generative Vision Transformers Generative vision transformers use prompt tuning to optimize and learn prompts, enhancing model performance through refined feature extraction. In generative vision transformers, prompts are learned and added through prompt tuning, with token generator parameters optimized via gradient descent (Sohn et al., 2023). The extraction process uses Multi-Layer Perceptron Classifiers (MLPC) and Predictors (MLPP) to encode class and sequence position indices, and performance is evaluated using metrics like Fréchet Inception Distance (FID). End-to-End Visual Prompt Tuning (E2VPT) (Han et al., 2023) involves learning prompts during fine-tuning, inserting predefined text tokens into transformer layers. These prompts are extracted using a pruning strategy to remove the least important ones, and performance is measured using accuracy and parameter efficiency, quantifying the number of tunable parameters and prediction correctness.

Frequency Domain Visual Prompts Visual prompts embedded in the frequency domain are extracted using pre-trained models that generate segmentation predictions, with performance evaluated using metrics like the Dice coefficient and Average Surface Distance (ASD) (Wang et al., 2023). For visual prompts added as perturbations to test-time examples, metrics such as standard accuracy and robust accuracy measure the model's predictions on both normal and adversarial

Table 11 Summary of perturbation techniques used for adding prompts as a template

Category	Description	Metrics	References
Embedded Visual Prompts	Embedding visual prompts into the input space of models for applications like segmentation and classification	Accuracy, precision, recall, F1 score, Jaccard F-measure (JF), average precision	Park and Byun (2024), Li et al. (2024a), Wu et al. (2022b)
Text/Visual Features	Extracting both text and visual features using specific encoders to enhance multi-modal applications	Accuracy gain, IoU, pixel accuracy	Zhu et al. (2023a), Wu et al. (2022a), Tsao et al. (2023), Kumananathaseelan et al. (2024)
Generative Vision Transformers	Using prompt tuning in generative vision transformers to optimize and learn prompts	Accuracy, parameter efficiency, FID	Sohn et al. (2023), Han et al. (2023)
Frequency Domain	Embedding visual prompts in the frequency domain to enhance model robustness and accuracy	Dice coefficient, Average Surface Distance (ASD), standard accuracy, robust accuracy	Wang et al. (2023), Chen et al. (2023a), Bahng et al. (2022)
Video and Tracking Prompts	Embedding prompts and perturbations in video frames to improve tracking models	Precision, recall, F-score, Cross-Entropy loss	Yang et al. (2022a), Pei et al. (2024)
Advanced Prompt Techniques	Using sophisticated techniques like Colorful Prompt Tuning and Vision-Language Pre-Training Models	Grounding accuracy, recall@N, mean recall@N, APCER, BPCER, ACER, HTER	Yao et al. (2024), Wen et al. (2022), Yu et al. (2024)

examples (Chen et al., 2023a). The visual prompt signal is embedded into input images by attaching it to each pixel, influencing the semantics of the image during adaptation. In scenarios where visual prompts are directly optimized via backpropagation (Bahng et al., 2022), they are added to input images, and during training, the model maximizes the likelihood of the correct label given the prompted image. The performance of visual prompting is assessed using metrics such as average accuracy across multiple datasets.

Video and Tracking Prompts Embedding prompts and perturbations in video frames enhances the accuracy and effectiveness of tracking models. In ProTrack (Yang et al., 2022a), the authors propose to embed signals by applying prompts and perturbations to input videos, enhancing the discriminative ability of RGB trackers. The performance of these techniques is evaluated using precision, recall, and F-score, which measure the accuracy and effectiveness of tracking results. In contrast, techniques like SA2VP (Pei et al., 2024) utilize cross-entropy loss to optimize model performance during training, with metrics like accuracy, precision, recall, and F1 score evaluating performance on benchmarks like VTAB-1k.

Advanced Prompt Techniques Advanced methods use sophisticated techniques to enhance model performance across various applications. Advanced methods like Colorful Prompt Tuning (CPT) (Yao et al., 2024) embed visual sub-prompts by marking image regions with distinct colors or segmentation masks, and textual sub-prompts using color-based query templates. The performance of CPT is evaluated with grounding accuracy, recall@N, and mean recall@N for visual relation detection, as well as accuracy in visual commonsense reasoning and question answering applications. In Vision-Language Pre-Training Models (VL-PTMs) (Wen et al., 2022; Yu et al., 2024), visual prompts are embedded by optimizing input images through model inversion, and the extracted features are evaluated using metrics like Attack Presentation Classification Error Rate (APCER), Bonafide Presentation Classification Error Rate (BPCER), Average Classification Error Rate (ACER), and Half Total Error Rate (HTER).

In conclusion, the use of visual and text prompts in machine learning models enhances performance across various applications. Techniques like combining language and image encodings or embedding prompts in different domains are evaluated using metrics such as accuracy, IoU, FID, and recall. These advancements continue to improve model robustness, accuracy, and functionality.

4.5 Others

As discussed earlier, there are some approaches that utilize types of templates different from the above template categories. A summary of all the works is given in Table 12. We now discuss the learning processes adopted by these works (Fig. 11).

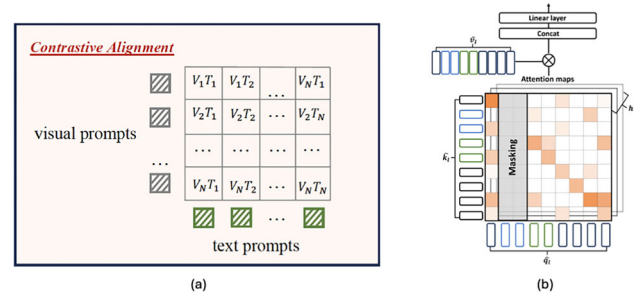


Fig. 11 Attention-based learning. Attention maps are utilized by either employing **a** contrastive alignment between text and visual prompts (Jiang et al., 2024), **b** applying masking to attention maps (Park & Byun, 2024)

Perturbation in Language Models Perturbing probability vectors and using sinusoidal signals to embed templates in language models enhances security. Perturbation in language models as performed by Zhao et al. (2023a) involves perturbing the probability vector using a sinusoidal signal and a hash function. The Lomb-Scargle periodogram is employed for spectrum estimation with each text input. The evaluation involves estimating the PSNR using the peak value of the estimated power spectrum at the particular frequency. The process of Poison-only Backdoor Attacks is also adopted for LLM defense by Li et al. (2023e, 2020); Adi et al. (2018) using templates as triggers on the input images. This attack method subtly corrupts a system, remaining dormant until a specific trigger is activated. The learning objective is tailored to detect such backdoor attacks, using hypothesis testing (Li et al., 2023e), spatial-temporal score (Li et al., 2020), hash functions (Adi et al., 2018), and similar statistical methods to ascertain the presence of the attack. Other methods (Sablayrolles et al., 2020) perturb the training data with class-dependent isotropic unit vectors to fine-tune the model, making it embed the template into the generated media.

Gaussian Noise Gaussian noise is added to feature points or vertex coordinates in 3D models to embed templates (Praun et al., 1999; Zhu et al., 2021). This process adjusts vertex coordinates in key regions or applies DCT transformations to high-frequency coefficients. Robustness and quality are measured through metrics like PSNR, SSIM, and LPIPS. Additionally, random noise templates protect sensitive information while maintaining model accuracy, with differential privacy metrics like ϵ used to assess privacy levels (Dwork, 2006; Zhang et al., 2018b; Cohen et al., 2019).

Trigger Samples and Predefined Tags Many works use trigger samples added to the input samples (Ahmadi et al., 2020; Kapusta et al., 2024; Lim et al., 2022; Zhang et al., 2023), and the neural networks are then trained/finetuned with these trigger-containing samples. The extraction process involves calculating correlations between extracted trigger samples and the original trigger pattern, or by estimating the out-

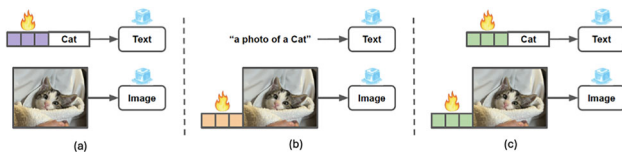


Fig. 12 Multi-modal prompt learning. Different prompt learning mechanisms shown by Shen et al. (2024), Li et al. (2023b), Jia et al. (2022) **a** text prompt trainable, **b** visual prompt trainable, and **c** both text and visual prompt trainable. Other methods include backbone trainable with frozen prompt generator, prompt generator, and head is trained with backbone as frozen, etc.

puts, which would only be predicted if trigger samples are used. Success rates of trigger verification and model accuracy on trigger samples are used to assess effectiveness. Metrics such as accuracy, precision, recall, F1 score, and trigger recognition rates are also employed. Predefined tags, such as AprilTags (Wang & Olson, 2016; Olson, 2011; Garrido-Jurado et al., 2014), are embedded in images or 3D models for applications like pose estimation and automatic detection. The learning process involves designing fiducial markers and training detection algorithms to recognize and decode these tags. The extraction process involves detecting and decoding these tags using specialized algorithms. Metrics include detection range, robustness to occlusion, and computational efficiency (Fig. 12).

Code Protection and Obfuscation Reverse engineering and obfuscation techniques safeguard code integrity and prevent unauthorized access. In code protection, reverse processes convert bit sequences back into operands and opcodes, revealing templates through dummy methods to verify authenticity and detect tampering (Monden et al., 2000). Balachandran et al. (2014) obfuscate code by transforming instruction sequences ending with jump instructions into basic blocks, displacing them to different functions. The obfuscation's effectiveness is assessed using tools like IDA Pro, focusing on disassembly and control flow errors to ensure robustness against automated attacks.

Miscellaneous Fourier space perturbation, passport layers, frontier stitching, and privacy preservation techniques offer robust methods for enhancing security and privacy in digital content. Fourier space perturbation, as demonstrated by Wen et al. (2023), embed templates into the Fourier space of a noise array using a rotation-invariant pattern composed of concentric rings selected from a Gaussian distribution. This method provides resilience to various image transformations and evaluates template presence by estimating a P-value, with detection confirmed if the P-value falls below a pre-determined threshold, α . Passport layers, introduced by Fan et al. (2019), enhance model security by adjusting the scale factor and bias terms of convolutional layers, ensuring that the network functions correctly only with the proper 'passport' parameters. An incorrect passport distorts the output,

preventing unauthorized use and reverse engineering, while the non-invertibility of the design further secures it against tampering.

Frontier stitching and privacy preservation methods further contribute to the security landscape. The frontier stitching algorithm by Le Merrer et al. (2020) subtly marks a model by clamping its decision frontier, using hypothesis testing to verify the template's presence. Laarhoven (2019) employ hash tables to manage high-dimensional data, identifying near neighbors and detecting potential colluders by analyzing data through sparse dot products, useful in large datasets. For privacy preservation, Xu et al. (2021) propose incorporating audio as noise by extracting it from a video, mapping it into a low-dimensional space, and adding it to the video frames' codebook. This ensures that only authorized receivers can decode and reconstruct the original video frames. Additionally, Liu and Kong (2018) use a spatiotemporal chaotic system for chaotic cryptography, applying an improved perturbation method based on a spatial chaotic map to secure the face region within an image, enhancing privacy and resisting decryption efforts.

In conclusion, these diverse approaches extend beyond conventional methods to embed and extract templates, enhancing the security and robustness of digital content. Techniques such as Fourier space perturbation, Gaussian noise in 3D models, trigger samples, and predefined tags offer innovative solutions for various applications. The integration of advanced perturbation, obfuscation, and privacy-preserving methods ensures the integrity and protection of data across multiple applications. These advancements underscore the importance of developing versatile and resilient techniques to safeguard digital information in an increasingly complex and interconnected world.

5 Applications

Now that we have a good understanding of the types of templates and the different processes involved, like perturbation and the learning process. Now we provide a discussion on the application of these proactive techniques in various applications. These techniques are used in a variety of applications, which are summarized in Table 13.

5.1 Vision Models Defense

In the evolving landscape of digital media, vision model defense has become a critical area of focus to ensure the integrity, authenticity, and security of visual content. This field encompasses a range of techniques to detect, prevent, and attribute deepfakes and other forms of tampered media. **Generative AI Protection.** The rapid adoption of diffusion and video generation models has introduced a new frontier

Table 12 Summary of perturbation techniques used for adding templates not categorized as main template categories

Category	Description	Metrics	References
Perturbation in Language Models	Perturbing probability vectors and using sinusoidal signals to embed templates in language models	PSNR, hypothesis testing, spatial-temporal score	Zhao et al. (2023a), Li et al. (2023e, 2020), Adi et al. (2018), Sablayrolles et al. (2020)
Gaussian Noise	Adding Gaussian noise to feature points or vertex coordinates to embed templates in 3D models	Detection error, correlation between coefficients, PSNR, SSIM, LPIPS, ϵ	Praun et al. (1999), Zhu et al. (2021), Nakazawa et al. (2010), Medimegh et al. (2018), Cho et al. (2007), Ohbuchi et al. (2002), Pham et al. (2018), Ai et al. (2009), Dwork (2006), Zhang et al. (2018b), Cohen et al. (2019)
Trigger and Predefined Tags	Using trigger samples and predefined tags to enhance detection capabilities	Accuracy, precision, recall, F1 score, trigger recognition rates	Zhang et al. (2023), Lim et al. (2022), Kapusta et al. (2024), Ahmadi et al. (2020), Peng et al. (2025), Li et al. (2022b), Wang and Olson (2016), Olson (2011)
Code Protection and Obfuscation	Ensuring code integrity and protection against unauthorized access	Robustness, control flow errors, detection range	Monden et al. (2000), Balachandran et al. (2014)
Fourier Space Perturbation	Embedding templates in the Fourier space to provide resilience to various image transformations	Interpretable P-value, detection threshold (α)	Wen et al. (2023)
Passport Layers	Adjusting scale factors and using adversarial techniques to enhance model security	Non-invertibility, detection accuracy, robustness	Fan et al. (2019)
Frontier Stitching	Clamp the decision frontier	Statistical framework, hypothesis testing	Le Merrer et al. (2020), Laarhoven (2019)
Privacy Preservation	Incorporating audio as noise and using chaotic systems to ensure privacy in multimedia content	Privacy protection level, sensitivity analysis	Xu et al. (2021), Liu and Kong (2018)

for proactive defenses that embed verifiable signals at generation time. Recent works such as Wang et al. (2025a); Zou et al. (2025), and Chen et al. (2025a) integrate watermarking directly into diffusion or latent-space pipelines to ensure content traceability across image and video synthesis. Frameworks like ProMark (Asnani et al., 2023c) and CustomMark Asnani et al. (2025), combine proactive watermarking with signed provenance metadata. Several recent works explore watermarking directly within diffusion pipelines, enhancing robustness to regeneration and editing (Lee & Cho, 2025; Lu et al., 2024). These approaches reflect a shift from post-hoc forensics to *born-authentic* media, where ownership, generation context, and integrity are established during synthesis.

Deepfake Detection and Attribution Enhancing deepfake detection and source attribution involves techniques like artificial fingerprints and learnable templates. Yu et al. (2021) use artificial fingerprints from training data to identify and attribute deepfakes to their source models. Asnani et al. (2022, 2023a) introduce learnable templates in real images for improved detection and localization of tampered images. The DeepMark (Tang et al., 2024b) framework uses a Digital Metadata Marker (DMM) for scalable deepfake detection by comparing visual features. AdvMark (Wu et al., 2024)

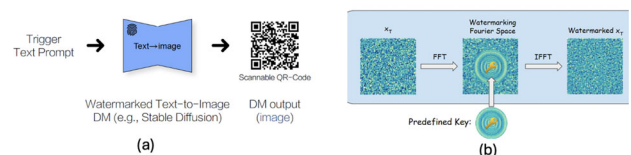


Fig. 13 Adding tags and triggers to the images. **a** Embedding trigger text prompt to the model, which, when given the prompt, would output the QR code (Zhao et al., 2023c), and **b** adding a predefined tag to the Fourier space of the image (Wen et al., 2023)

embed adversarial templates as templates to enhance deepfake detection accuracy. Asnani et al. (2023b) estimate the fingerprints left by generative models using the predefined constraints for deepfake detection, image attribution, and reverse engineering of generative models (Fig. 13). Dynamic or hybrid watermarking schemes further integrate adaptive or multi-stage embeddings to localize tampering (Chen et al., 2025b; Zhang et al., 2025).

Tampering Detection and Verification Watermarking and embedding techniques ensure image integrity and recovery after tampering. Various methods, such as spatial domain embedding (Laouamer et al., 2015), DCT-based schemes (Singh & Singh, 2016), and singular value decomposition (Dadkhah

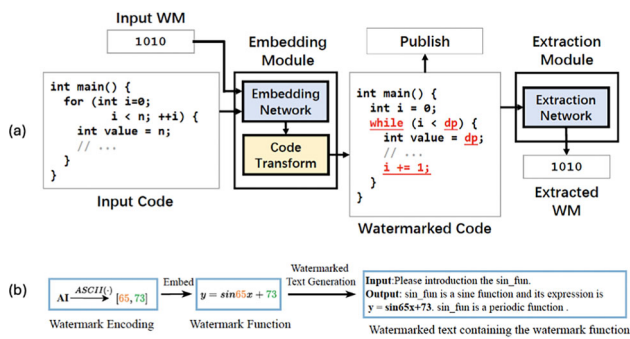


Fig. 14 Method using other types of templates, like **a** code transformations (Yang et al., 2023a), and **b** specific functions (Li et al., 2023c) into the data to embed templates into the input data

et al., 2014), create templates to detect and sometimes recover tampered content. These techniques maintain image integrity by detecting altered blocks and recovering areas with high precision. Adaptive strategies further enhance this by considering image block complexity and employing techniques like image smoothness differentiation, overlapping embedding (Hsu & Tu, 2016), and hierarchical recovery (Cao et al., 2017; Hsu & Tu, 2010; Qin et al., 2017). Interlocking templates within image blocks (Haghighi et al., 2018; Lee & Lin, 2008) enhance tamper detection and provide a fallback for recovery, improving digital media resilience.

Face Anti-Spoof Yu et al. (2024) proposes to use proactive methods for face anti-spoofing. The work addresses the challenge of missing modalities in both training and testing phases by incorporating visual prompts and residual contextual prompts in multimodal transformers, ensuring robust learning of flexible-modal features with minimal computational overhead.

Identity Protection Embedding authentic signatures in images and using dual traces helps protect personal identities from deepfakes. Techniques like face feature disentanglement combined with perturbation embed authentic signatures into digital images, ensuring identity protection (Zhao et al., 2023b). Dual traces, both sustainable and erasable, verify authenticity and detect fraud in media (Sun et al., 2023; Zhang et al., 2024d). Additionally, encoder-decoder methods, enhanced by differentiable JPEG compression (Yang et al., 2021b), defend against deepfakes by detecting compression artifacts (Fig. 14).

Disrupting Deepfake Generation To protect individual identities, the encoder-decoder approach trained with generative models preserves identity nuances (Wang et al., 2021). Introducing targeted noise as templates preempts deepfake generation, ensuring models trained on perturbed images produce subpar results (Huang et al., 2022; Ruiz et al., 2020; Segalis & Galili, 2020; Van Le et al., 2023; Yeh et al., 2020). This proactive defense undermines deepfake quality, making them less convincing and more detectable. Techniques

by Lu and Liao (2001); Segalis and Galili (2020) create images resistant to manipulation, protecting against face-swapping attempts. Adversarial attacks on image translation networks (Yeh et al., 2020) compromise deepfake generation, resulting in flawed outputs. The DUAW technique (Ye et al., 2023) disrupt the variational autoencoder (VAE) in Stable Diffusion models, introducing distortions to protect images universally.

5.2 LLM Defense

Ensuring the security and integrity of Large Language Models (LLMs) is critical as they face risks like unauthorized use, disinformation, and adversarial attacks. A comprehensive defense strategy now includes authenticity verification, provenance tracking, and robust protections against malicious activities to safeguard LLMs in the digital age.

Authenticity Verification and Provenance Tracking Ensuring authenticity and tracking content provenance are crucial in combating disinformation. Techniques like the Decoupled Invertible Neural Network (DINN) (Meng et al., 2022) encode dual-tags into images as fingerprints for authenticity verification and provenance tracking. Advanced text perturbation methods interweave encoded signals within natural language (Zhao et al., 2023a), using neural networks (Wu et al., 2023b) or linguistic tools like the Universal Sentence Encoder and Word2Vec (Munyer & Zhong, 2023). These methods also secure code by embedding templates, protecting against plagiarism and unauthorized use.

Watermarking for large language models has become an active subfield of proactive attribution. Dathathri et al. (2024) propose a scalable framework for identifying model-generated outputs, while Pang et al. (2024) and He et al. (2024) outline fundamental design trade-offs among embedding strength, detection accuracy, and bias. Adaptive and semantics-based schemes improve resilience to paraphrasing and sampling variance (Liu & Bu, 2024; Ren et al., 2024a). These techniques encode lexical or embedding-space patterns within generated text to enable model identification, authorship verification, and detection of unauthorized reuse. They complement visual and multimodal schemes by extending proactive provenance to the linguistic domain, strengthening responsible use of generative text systems.

Defensive Strategies Against Malicious Exploitation Various defensive strategies have been developed to protect against malicious exploitation. Character-level input perturbations (Robey et al., 2023) inoculate against adversarial attacks, and specialized datasets (Dong et al., 2023) enhance the resilience of language models under real-world conditions. Perturbation strategies have evolved to include backdoor techniques that insert covert triggers into text outputs (Liu et al., 2023b), enabling the tracing of unauthorized reproductions.

Table 13 Summary of categories, descriptions, types of templates, and references. [KEYS: seq.= sequence, V.P.= visual prompt, per.= perturbation]

Application	Description	Type of Template	References
Deepfake Detection and Attribution	Techniques for identifying and attributing deepfake content	Bit seq., Learn per	Yu et al. (2021), Asnani et al. (2022), Asnani et al. (2023a), Tang et al. (2024b), Wu et al. (2024)
Tampering detection and verification	Techniques to detect tampering and ensure image integrity	Bit seq., 2D noise	Laouamer et al. (2015), Singh and Singh (2016), Dadkhah et al. (2014), Hsu and Tu (2016), Qin et al. (2017), Cao et al. (2017), Hsu and Tu (2010), Lee and Lin (2008), Haghighi et al. (2018)
Face Anti-Spoof	Detecting face spoofing	V.P	Yu et al. (2024)
Identity Protection	Protect personal identities against deepfakes	Learn per., Bit seq	Zhao et al. (2023b), Sun et al. (2023), Zhang et al. (2024d), Neekhara et al. (2022), Yang et al. (2021b)
Disrupting deepfake generation	Disrupt deepfake generation and ensure image resistance	Per	Wang et al. (2021), Van Le et al. (2023), Huang et al. (2022), Segalis and Galili (2020), Ruiz et al. (2020), Yeh et al. (2020), Ye et al. (2023)
Techniques for authenticity verification and provenance tracking	Authenticity verification and provenance tracking	Bit seq., Text	Meng et al. (2022), Zhao et al. (2023a), Wu et al. (2023b), Munyer and Zhong (2023)
Defensive Strategies Against Malicious Exploitation	Enhance LLM resilience against malicious exploitation	Text, triggers	Robey et al. (2023), Dong et al. (2023), Liu et al. (2023b)
Context-Aware Modifications	Perturbation via probabilistic outputs and context-aware lexical substitutions	Text	Kirchenbauer et al. (2023b), Yang et al. (2022b), He et al. (2022a), Rizzo et al. (2019), Kirchenbauer et al. (2023a)
Intellectual Property Protection	Methods to protect intellectual property	Text	Zhang et al. (2010), He et al. (2022b), Yang et al. (2023a), Li et al. (2023c)
Model Attribution	Content origin identification	Bit seq	Zeng et al. (2023), Wen et al. (2023), Wu et al. (2020), Zhao et al. (2023c), Atli Tekgul and Asokan (2022), Yu et al. (2019)
Neural Network Ownership and Protection	Embed templates within neural networks to protect and prove ownership of models	V.P., per., bit seq	Darvish Rouhani et al. (2019), Xue et al. (2022), Chen et al. (2019a), Adi et al. (2018), Uchida et al. (2017), Nagai et al. (2018), Le Merrer et al. (2020), Fernandez et al. (2023), Liu et al. (2021), Zhang et al. (2018a), Peng et al. (2022)
Ownership Verification in Federated Learning	Ownership verification of federated neural networks	Bit seq., per	Li et al. (2022), Han et al. (2022), Tekgul et al. (2021), Xu et al. (2019), Liu et al. (2021b)
Protection Techniques for Diffusion Models	Protect diffusion models	Bit seq., per	Yang et al. (2024b), Huang et al. (2024a), Cui et al. (2023), Lei et al. (2024), Meng et al. (2025), Ahmadi et al. (2020), Zhang et al. (2024c), Zhang et al. (2024a), Peng et al. (2025), Lim et al. (2022)
Camera Model Perturbation and Localization	Enhance forensic analysis and image authenticity verification	Bit seq., per	Cozzolino and Verdoliva (2019), Wu et al. (2023b)
Data and Artists' Attribution	Ensure proper recognition and preservation of authorship rights for data and artists	Bit seq	Cui et al. (2025), Asnani et al. (2023c), Li et al. (2023e), Li et al. (2020)
Fingerprinting for preservation of authorship rights	Identifying authorship rights infringement and protecting digital content	Bit seq	Furon and Desoubeaux (2014), Balachandran et al. (2014), Fan et al. (2019)
User Privacy Preservation	User privacy preservation ensuring efficient protection of personal data	Per	Laarhoven (2019), Li et al. (2021), Dwork (2006), Zhang et al. (2018b), Blum et al. (2005), Zhang et al. (2021b), Tang et al. (2024a), Li et al. (2023d)

Table 13 continued

Application	Description	Type of Template	References
Face Recognition Privacy	Face recognition protection to deceive models	Per	Xiao et al. (2021), Zhong and Deng (2020), Rajabi et al. (2021), Shan et al. (2020), Wu et al. (2019), Hu et al. (2022), Xu et al. (2021), Komkov and Petiushko (2021), Dong et al. (2019)
Miscellaneous	Autonomous driving privacy, and surveillance privacy	Per., bit seq	Mirjalili et al. (2018), Xiong et al. (2020), Paruchuri (2009), Liu and Kong (2018)
Point Cloud Adversarial Defense and Perturbation	Enhancing robustness of 3D data models and perturbation for point clouds	Per., Bit seq	Liang et al. (2022b), Ding et al. (2021), Liu et al. (2019), Xiaoqing (2015), Feng (2015), Ohbuchi et al. (2004), Cotting et al. (2004), Ohbuchi et al. (2001)
3DMMs, SDFs	3DMMs Perturbation ensuring preservation of authorship rights and content authentication	Bit seq	Wang et al. (2022b), Zhu et al. (2023b)
NeRF Models Perturbation	Perturbation methods for Neural Radiance Fields (NeRF) for preservation of authorship rights	Bit seq	Luo et al. (2023), Jang et al. (2024), Chen et al. (2024a), Chen et al. (2023c), Li et al. (2023a), Huang et al. (2024b)
3D Mesh Defense	Embedding templates in 3D meshes to ensure authorship rights and tamper detection	Bit seq	Hamidi et al. (2019), Wang et al. (2022a), Zafeiriou et al. (2005), Yoo et al. (2022), Praun et al. (1999), Zhu et al. (2024), Zhu et al. (2021), Medimegh et al. (2018)
3D GS Protection	Embedding hidden information in 3D scenes	Bit seq	Zhang et al. (2024e)
3D Models Protection	Protection for 3D models ensuring minimal visibility of distortions	Bit seq	Peng et al. (2022), Benedens (1999), Nakazawa et al. (2010), Yeung and Yeo (1998), Alface et al. (2007), Kanai et al. (1998), Liu et al. (2012), Liu et al. (2012), Chou and Tseng (2006), Chou and Tseng (2009)
VLMs	Enhancing VLMs for downstream applications	V.P., T.P	Kunananthaseelan et al. (2024), Zhang et al. (2023b), Zhu et al. (2023a), Wu et al. (2022a), Nasiriany et al. (2024), Xing et al. (2023), Shen et al. (2024), Zhao and Patras (2023), Mirza et al. (2024), Maniparambil et al. (2023)
Visual Prompt Tuning for GMs	Enhancing image generation quality and efficient transfer learning	V.P	Zhang et al. (2024b), Sohn et al. (2023), Kim et al. (2024b), Ma et al. (2024), Han et al. (2023), Kim et al. (2024b), Park and Byun (2024), Wu et al. (2022b), Chen et al. (2023b), Ju et al. (2022), Wang et al. (2024b), Zhang et al. (2024g), Yoo et al. (2023), Song et al. (2023)
Text-to-3D Generation	Enhancing text-to-3D generation	V.P	Chen et al. (2024b)
Object Localization and Tracking	Unadversarial examples, fractal markers, object recognition, localization, and tracking	Bit seq., V.P	Salman et al. (2021), Asnani et al. (2024), Wagner and Schmalstieg (2007), Wang and Olson (2016), Olson (2011), Garrido-Jurado et al. (2014), Abbas et al. (2019), Álvarez et al. (2012), Wang et al. (2018)
Image Editing and Inpainting	Text-based image editing and inpainting	V.P	Nguyen et al. (2023), Bar et al. (2022)
Medical Image Segmentation	Improve medical image segmentation	V.P	Wang et al. (2025b), Wang et al. (2023)

Context-Aware Modifications Embedding secret signals and making context-aware modifications ensures that templates remain undetected by attackers. Some methods embed secret signals into probabilistic model outputs (Kirchenbauer et al., 2023b) or apply context-aware lexical substitutions (He et al., 2022a; Yang et al., 2022b), detectable only by insiders. Homoglyph substitutions (Rizzo et al., 2019) disguise textual input, preserving privacy and message integrity. Addi-

tionally, ‘green tokens’ (Kirchenbauer et al., 2023a) embed templates into high-entropy words to mark and identify the ownership or origin of textual content.

Restoration-Oriented Approaches and Intellectual Property Protection Restoration-oriented approaches and strategies to protect intellectual property are crucial for maintaining content integrity. Self-embedding template schemes (Zhang et al., 2010) mark and aid in recovering original content

if tampered with. To protect text generation APIs, subtle alterations in word distribution patterns create hard-to-detect templates (He et al., 2022b). Code transformation and variable substitution generate template functions (Li et al., 2023c; Yang et al., 2023a), which can be verified later for user code protection.

5.3 Attribution and Preservation of Authorship Rights

The applications and methods summarized in this section illustrate a broad range of innovative strategies employed for securing generative models, tracing sources, and establishing neural network ownership and preservation of authorship rights through various forms of proactive schemes and fingerprinting.

Model Attribution Proactive techniques for attributing model outputs help ensure content origin identification. One approach involves fine-tuning models with encrypted images to perform effective attribution. Techniques by Zeng et al. (2023), Wen et al. (2023), and Wu et al. (2020) use unique characteristics of encrypted data to trace and secure generative models (GMs), embedding identifiable signals within the data. Additionally, encoding binary strings into diffusion models (Atli Tekgul & Asokan, 2022; Yu et al., 2019; Zhao et al., 2023c) involves fine-tuning with encrypted image pairs and trigger prompts to trace content back to its rightful owner.

Neural Network Ownership and Protection The application of neural network protection and ownership is addressed through various methods to embed templates within networks. Techniques like DeepSigns (Darvish Rouhani et al., 2019; Xue et al., 2022) embed information into the probability density function of activation sets, while DeepMarks (Chen et al., 2019a) offer an end-to-end fingerprinting framework. These methods resist attacks like model extraction and collusion, securing creative ownership. Other approaches include backdoor attacks (Adi et al., 2018), additional regularizers (Uchida et al., 2017; Nagai et al., 2018; Le Merrer et al., 2020), fine-tuning the latent decoder of diffusion models (Fernandez et al., 2023), network parameters residuals (Liu et al., 2021), query prompts (Zhang et al., 2018a), and learnable perturbations (Peng et al., 2022). A deep spatial perturbation framework (Zhang et al., 2021) is robust against surrogate model attacks, supporting image-based templates to protect data and algorithms. Another method involves embedding a template into the weights of a neural network (Wang et al., 2020), ensuring robustness against brute-force attacks. The DAWN framework (Szyller et al., 2021) deters model extraction by dynamically changing responses to specific queries. BlackMarks (Chen et al., 2019b) ensures fidelity, robustness, and security for intellectual property protection (Fig. 15).



Fig. 15 Template addition in 3D meshes (Yu et al., 2003; Liu et al., 2019; Zhang et al., 2024e), point clouds meshes (Feng, 2015), point cloud, and into the 3D vertices (Ohbuchi et al., 2002)

Ownership Verification in Federated Learning Ownership verification techniques in federated learning ensure robust and private model ownership. The FedIPR framework (Li et al., 2022) embed and verifies private templates in Federated Deep Neural Networks (FedDNN) without disclosing information, addressing robustness challenges like client selection and differential privacy. Han et al. (2022) use key exchange technology and a double masking protocol for privacy protection and correctness verification. WAFFLE (Tekgul et al., 2021) embed resilient templates in DNN models trained with federated learning without accessing training data.

Protection Techniques for Diffusion Models Various innovative techniques protect diffusion models while maintaining performance. Gaussian Shading (Yang et al., 2024b) embed templates into generated images using template diffusion, randomization, and distribution-preserving sampling. FreezeAsGuard (Huang et al., 2024a) selectively freezes critical tensors to prevent unauthorized fine-tuning of diffusion models. FT-Shield (Cui et al., 2023) uses bi-level optimization and a Mixture of Experts framework to generate and detect templates in text-to-image diffusion models. DiffuseTrace (Lei et al., 2024) embeds invisible templates into generated images without compromising quality, while Latent template (Meng et al., 2025) injects and detects templates in the latent space of diffusion models.

Camera Model Perturbation and Localization Camera model perturbation techniques enhance forensic analysis and image authenticity verification. Cozzolino and Verdoliva (2019) use a Siamese network trained with image patches from different cameras to detect image forgeries and identify the specific camera model, aiding forensic analysis and ensuring digital image credibility. SepMark (Wu et al., 2023b) embed an encoded message into pristine images, allowing for later extraction to verify authenticity and trace the image source, providing robust deepfake defense and source tracing capabilities.

Data and Artists' Attribution Techniques for data and artist attribution ensure proper recognition and authorship rights. ProMark (Asnani et al., 2023c) encrypt training data with bit sequence templates, enabling GenAI models to perform concept attribution during training. Diffusion Shield (Cui et al., 2025) embeds unique messages into datasets for identification without retraining. Poison-only backdoor attacks (Li et al., 2023e, 2020) mark datasets by embedding special behav-

iors in data samples, allowing verification of model training origins through hypothesis testing. Recent proactive attribution frameworks embed ownership metadata directly into generative models to enable verifiable provenance (Asnani et al., 2025). Complementary socio-technical perspectives advocate hardware-assisted watermark verification for transparent and responsible deployment (Kherraz, 2025).

Fingerprinting for preservation of authorship rights Active fingerprinting techniques provide powerful tools for identifying authorship rights infringement and protecting digital content. These techniques include the use of Tardos codes for traitor tracing (Furon & Desoubeaux, 2014), code fragments relocation (Balachandran et al., 2014), and passport-based DNN ownership verification schemes (Fan et al., 2019). By embedding digital passports and employing active fingerprinting, content creators can more effectively control and enforce their authorship rights in the digital realm, preventing unauthorized use and distribution.

5.4 Privacy Protection

The collection of works discussed in this section showcases a concerted effort to preserve privacy across various digital platforms, with a particular focus on protecting user identity and personal data in the face of advanced facial recognition technologies and surveillance mechanisms.

User Privacy Preservation Methods aimed at user privacy preservation ensure efficient protection of personal data. The score-based fingerprinting framework (Laarhoven, 2019) accelerates decoding times for efficient data protection. Mapping Distortion Based Protection (MDP) and AugMDP (Li et al., 2021) misalign images and labels to confuse potential data breaches without compromising benign neural network performance. Various techniques protect privacy on digital platforms, including adding random noise to query functions for differential privacy (Dwork, 2006), obfuscation techniques for training data (Zhang et al., 2018b), and the SuLQ framework for statistical database privacy (Blum et al., 2005). Proactive Privacy-preserving Learning (PPL) (Zhang et al., 2021b) uses adversarial generators to transform data for malicious model manipulation. Universal Transferable Adversarial Perturbation (UTAP) (Tang et al., 2024a) protect privacy in facial image databases.

Face Recognition Privacy Techniques for face recognition protection deceive facial recognition models without affecting legitimate applications. Methods like adversarial patches and perturbations are finely tuned to deceive models (Rajabi et al., 2021; Xiao et al., 2021; Zhong & Deng, 2020), while systems like Fawkes (Shan et al., 2020) apply pixel-level changes to prevent unauthorized recognition. Other approaches perform deidentification in the feature space (Wu et al., 2019). Innovative adversarial and perturbation techniques include ATM-GAN (Hu et al., 2022), which creates

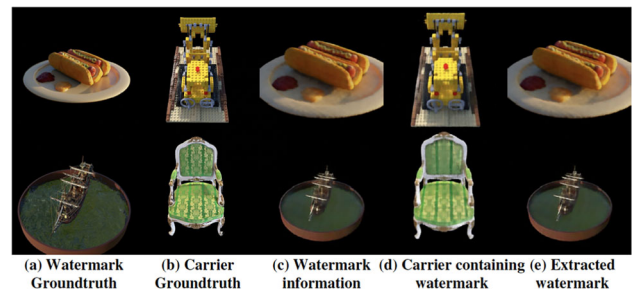


Fig. 16 Input-encrypted input pairs for NeRF models (Luo et al., 2023; Chen et al., 2024a)

adversarial examples to distort makeup styles and prevent unauthorized recognition. The cycle-VQ-VAE framework obscures video streams by integrating audio as noise (Xu et al., 2021). Komkov and Petiushko (2021) use a Spatial Transformer Layer to project stickers onto face images for face ID systems. Dong et al. (2019) discusses black-box adversarial attacks on face recognition systems, exposing deep CNN vulnerabilities.

Autonomous Driving and Surveillance In autonomous driving, ADGAN (Xiong et al., 2020) protects location privacy by obscuring sensitive information in camera data while maintaining utility. Similarly, surveillance privacy techniques, such as those by Paruchuri (2009) and Liu and Kong (2018), use spatial chaotic maps to encrypt human faces in video footage, safeguarding identities without compromising the overall utility of the footage (Fig. 16).

5.5 3D Domain

In 3D modeling, protecting digital content and intellectual property is crucial across industries like entertainment, manufacturing, medical imaging, and virtual reality. Proactive methods, including perturbation techniques for 3D models, meshes, and point clouds, embed imperceptible yet resilient templates into digital assets. These techniques detect unauthorized use, prevent tampering, and ensure preservation of authorship rights and content authentication.

Point Cloud Adversarial Defense and Perturbation The Perturbation Adaption Generation Network (PAGN) (Liang et al., 2022b) is designed for point cloud adversarial defense in 3D model classification. PAGN includes a perturbation-injection module, a generative module, and a shape similarity measure to improve robustness. The perturbation-injection module simulates adversarial samples, while the generative module visualizes these samples and measures shape similarity. Additionally, Ding et al. (2021) proposes the Geometry-Consistent Point Cloud Upsampling (GC-PCU) method, generating uniform, clean, and dense point clouds from sparse ones through feature extraction, perturbation learning, and geometric reconstruction. Perturbation tech-

niques for 3D point clouds balance transparency and robustness (Cotting et al., 2004; Feng, 2015; Liu et al., 2019; Ohbuchi et al., 2004; Xiaoqing, 2015). Methods include defining local sets, calculating RMSC values, establishing synchronization relations, and embedding templates by modifying distance normalization means (Liu et al., 2019).

3D Morphable Models and Signed Distance Fields Perturbation Perturbation techniques for 3D Morphable Models (3DMMs) ensure preservation of authorship rights and content authentication. A deep neural network scheme (Wang et al., 2022b) uses an encoder to encode templates and meshes, an Attacker to add perturbations, and a Decoder to extract templates. The FuncMark method (Zhu et al., 2023b) embed binary templates in signed distance fields (SDFs) using spherical partitioning and local deformation, allowing template extraction from derived meshes.

NeRF Model Perturbation Neural Radiance Fields (NeRF) have advanced preservation of authorship rights through neural 3D perturbation. One method trains a 2D decoder and the NeRF model separately, achieving high bit accuracy and reconstruction quality using patch loss and discrete wavelet transform (Jang et al., 2024). Another approach uses Implicit Neural Representation (INR) to embed template information into NeRF models, addressing low template capacity and security risks (Chen et al., 2024a). MarkNeRF (Chen et al., 2023c) use neural networks to protect implicit data representations, offering high imperceptibility, robustness, and anti-interference capability. StegaNeRF (Li et al., 2023a) embed invisible information within NeRF renderings through a two-stage optimization process, balancing rendering quality and decoding accuracy. Noise-NeRF (Huang et al., 2024b) introduce trainable noise on specific views for steganography, enhancing quality and efficiency.

3D Mesh Defense 3D mesh defense techniques involve embedding templates as templates into the geometrical structures of 3D models to ensure preservation of authorship rights and detect tampering (Hamidi et al., 2019; Praun et al., 1999; Yoo et al., 2022; Wang et al., 2022a; Zafeiriou et al., 2005). Methods such as DEEP3DMARK (Wang et al., 2022a) use attention-based convolutions to embed templates into 3D meshes. Another approach uses wavelet transform (Hamidi et al., 2017; Wang et al., 2008) and the norm of wavelet coefficient vectors as perturbation primitives, embedding templates by quantizing these norms (Wang et al., 2011). Some techniques employ Voronoi patches (Ai et al., 2009) and transform applications for template embedding, while others use robust mesh feature segmentation (Feng et al., 2014) and DCT transformation. Many of these methods are designed to be resistant to various attacks, including noise addition, 3D rotation, simplification, and cropping.

3D Gaussian Splatting Protection The GS-Hider framework (Zhang et al., 2024e) embed hidden information securely using a coupled secured feature attribute and ren-

dering pipeline. It employs two parallel decoders to separate rendered RGB scenes and hidden messages, ensuring robustness against rendering. This method enhances security, transparency, and authenticity in encrypted transmission, 3D compression, and preservation of authorship rights.

3D Model Protection Protection techniques for 3D models ensure minimal visibility of distortions while maintaining security. Methods involve modifying the geometrical structure or vertex positions to embed bits according to a key (Peng et al., 2022; Benedens, 1999; Nakazawa et al., 2010; Yeung & Yeo, 1998; Alface et al., 2007; Kanai et al., 1998). Templates are used for data hiding to protect models while maintaining data integrity (Hou et al., 2023; Jiang et al., 2017; Zhang et al., 2023c; Tsai & Liu, 2022). Techniques like octree spatial subdivision and multi-MSB prediction enhance embedding capacity and lossless recovery (Hou et al., 2023). Jiang et al. (2017) use coordinate transformation, prediction error detection, model encryption, and label map embedding to improve embedding rate and capacity. New frameworks embed multiple or geometry-aware watermarks to ensure ownership persistence under rendering transformations (Jang et al., 2025; Kulthe et al., 2025).

5.6 Improving Generative Models

We outline various methods for improving the interpretability and generalization of deep neural networks, visual prompt tuning in generative models, transformers, and text-to-3D generation, as well as enhancing vision-language models (VLMs) and large language models (LLMs).

Vision-Language Models (VLMs) Proactive techniques enhance VLMs for downstream applications like image classification, recognition, semantic segmentation, and object detection. Language-Grounded Visual Prompting (LaViP) (Kunananthaseelan et al., 2024) uses input-specific visual prompts with language integration for better model adaptability. The Text-to-Video Prompting framework (TVP) (Zhang et al., 2023b) improves Text-to-Video Generation (TVG) models with optimized prompts for enhanced generation quality and temporal localization accuracy. Visual Prompt multi-modal Tracking (ViPT) (Zhu et al., 2023a) adapts pre-trained RGB-based models for downstream applications using modality-specific visual prompts. Enhanced Visual Prompting (EVP) (Wu et al., 2022a) and Prompting with Iterative Visual Optimization (PIVOT) (Nasiriany et al., 2024) enable models to handle spatial applications and multimodal learning without task-specific fine-tuning. Next, further enhancements in VLMs include techniques like Dual Modality Prompt Tuning (DPT) (Xing et al., 2023) for learning visual and text prompts simultaneously, and MVLPT (Shen et al., 2024) for incorporating cross-task knowledge into prompt tuning. Recent diffusion and video watermarking frameworks achieve zero-distortion or latent-

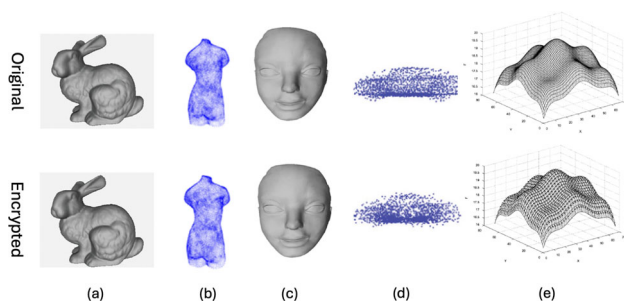


Fig. 17 Various examples of input-encrypted input pairs after adding the templates into the 3D input data **a** (Liu et al., 2019), **b** (Feng, 2015), **c** (Yu et al., 2003), **e** (Liang et al., 2022b), and **f** (Bors, 2006)

space integration (Chen et al., 2025a; Hu et al., 2025; Li et al., 2025), while others aim for robustness against fine-tuning and distribution shift (Teng et al., 2025; Wang et al., 2025c). Cross-domain extensions such as tabular diffusion watermarking further broaden proactive protection (Zhu et al., 2025) (Fig. 17).

Visual Prompt Tuning for Generative Models In the context of visual prompt tuning for generative models, techniques like InMeMo (Zhang et al., 2024b) and prompt tuning for generative vision transformers (Sohn et al., 2023; Kim et al., 2024b; Ma et al., 2024) focus on enhancing image generation quality and efficient transfer learning. E2VPT (Han et al., 2023), VPT (Kim et al., 2024b), Fair-VPT (Park & Byun, 2024), and PromptChainer (Wu et al., 2022b) aim to improve transformer models by incorporating learnable visual prompts and addressing fairness issues, respectively. Additional techniques include Iterative Label Mapping-based Visual Prompting (ILM-VP) (Chen et al., 2023b) for reprogramming pre-trained source models to new target applications, DINOv for generic and referring segmentation applications, and video-based visual-language pre-training (I-VL) (Ju et al., 2022) for video understanding applications like action detection and localization, text-video retrieval, summarization, *etc.* Approaches like Self-Prompt Tuning (SPT) (Wang et al., 2024b), Adaptive Pretraining (AP) (Zhang et al., 2024g), and Gated Prompt Tuning (Yoo et al., 2023) adapt pre-trained models to downstream applications. Other methods (Cai et al., 2024; Lee et al., 2023; Liu et al., 2023a; Song et al., 2023) incorporate similar techniques by leveraging prompt tuning to adapt pre-trained models for downstream CV applications.

Text-to-3D Generation For text-to-3D generation, VP3D (Chen et al., 2024b) introduce a novel visual prompt-guided diffusion model to enhance text-to-3D generation by utilizing high-quality images from 2D diffusion models as visual prompts.

5.7 Other CV Applications

In computer vision, innovative methods are continuously being developed to improve object recognition, localization, tracking, and image manipulation. These advancements are critical for improving the performance and adaptability of models across various applications, from augmented reality to medical imaging.

Object Localization and Tracking To improve object recognition, Salman et al. (2021) introduce unadversarial examples, which modify objects to enhance performance and robustness. Their method employs gradient-based algorithms to design unadversarial patches and textures, boosting system resilience in diverse environments. Asnani et al. (2024) propose a novel template learning paradigm to improve performance in 2D and camouflaged object detection. Several advancements have been made in object localization and tracking (Garrido-Jurado et al., 2014; Olson, 2011; Wagner & Schmalstieg, 2007; Wang & Olson, 2016) that reduce false positives, increase detection rates, and minimize computing time for tag detection, making it suitable for computation-limited systems like smartphones. Fractal markers (Romero-Ramire et al., 2019) provide a novel approach to long-range marker pose estimation, offering robustness to occlusion and wider detection ranges. Further, object detection and segmentation have also improved by adding region proposals (Girshick et al., 2014) or by using prompts for fine-tuning VLMs (Long et al., 2023; Liang et al., 2023).

Image Editing and Inpainting For image editing (Nguyen et al., 2023) and inpainting (Bar et al., 2022), a method involving visual prompting leverages example pairs representing “before” and “after” edits to learn text-based editing directions, utilizing text-to-image diffusion models.

Medical Image Segmentation The Progressive Classification and Data Augmentation Learning (PCDAL) framework (Wang et al., 2025b) utilize deep learning models for classification and segmentation applications, incorporating data augmentation to increase training set size and prevent overfitting. Fourier Visual Prompting (FVP) (Wang et al., 2023) addresses domain shift in medical image segmentation by introducing visual prompts in the frequency domain, guiding pre-trained models to perform well in the target applications.

6 Threat Models, Evaluation, and Responsible AI Considerations

6.1 Threat Models and Evaluation Overview

Proactive schemes vary widely across modalities and objectives, yet most can be analyzed through a common

Table 14 Concise threat models, assumptions, targets, and metrics across application domains

Application Domain	Attacker Assumptions	Defender Assumptions	Robustness Targets	Metrics
Vision Models Defense	Editing, compression, diffusion remix	Template / detector key known	Resize, crop, JPEG, regeneration	AUC, BER, PSNR, SSIM
LLM Defense	Paraphrase, prompt edit, re-generation	Watermark / key access	Paraphrase, sampling variance	Detection rate, AUC, Perplexity shift
Authorship Attribution	Signature removal, ownership claim	Embedded ID retrievable	Fingerprint removal, tamper, transfer	BER, Bit accuracy, PSNR, Fidelity
Privacy Protection	ID recovery, feature extraction	Perturbation budget known	Blur, crop, recompress	Privacy gain, Utility drop, PSNR
3D Domain	Re-mesh, smooth, simplify	Embedded payload known	Re-topology, noise, conversion	Bit acc., MRMS, Correlation
Improving Gen. Models	Fine-tune, distill, retrain	Latent watermark / tag known	Fine-tune, pruning, diffusion remix	BER, PSNR, SSIM, Fidelity

threat-model lens. Evaluation protocols increasingly include adversarial and adaptive-removal scenarios. Müller et al. (2025) demonstrate black-box forgery attacks on semantic watermarks, and Yang et al. (2024) analyze vulnerabilities of modern diffusion watermarks to simple averaging. Robust benchmarking under generative-editing operations continues to evolve (Lu et al., 2024). Table 14 summarizes, for each application family discussed in Sect. 5, the assumed capabilities of the adversary, the defender’s available knowledge or control, the typical robustness goals, and the quantitative metrics used for evaluation. This unified view highlights the shared structure of proactive formulations—each balancing perceptual fidelity, robustness to transformation, and verifiability under adversarial or post-processing conditions.

6.2 Responsible AI Summary Across Application Domains

Beyond algorithmic performance, proactive schemes must be seen through the lens of responsible and transparent deployment. Responsible deployment considerations increasingly emphasize transparency and device-level verification, positioning proactive watermarking as a socio-technical safeguard for authenticity in generative media (Kherraz, 2025). As these methods directly interact with sensitive content, user data, and generative models, their intended uses and possible misuse scenarios warrant careful consideration. Table 15 summarizes the ethical dimensions across major application domains, outlining how proactive schemes are expected to strengthen authenticity, privacy, and accountability while highlighting potential risks such as falsified ownership or malicious removal of embedded information. By contextualizing each domain’s evaluation under adaptive-removal and adversarial conditions, this summary emphasizes the importance of aligning technical robustness with responsible and trustworthy AI practices.

7 Why Use Adversarial Perturbations for Social Good?

Adversarial perturbations have historically been viewed as threats to the stability of machine learning models. However, their defining characteristics, such as sensitivity, differentiability, and fine-grained control, make them ideal tools for embedding functionality into AI systems when redirected toward constructive purposes (Goodfellow et al., 2014; Szegedy et al., 2013). These properties allow proactive schemes to modify model behavior with minimal perceptual change to input data, enabling use cases like watermarking, authorship tracing, tamper detection, and privacy preservation (Adi et al., 2018; Chen et al., 2019a).

Model Sensitivity as an Asset Deep neural networks are inherently sensitive to small changes in their inputs. Adversarial attacks traditionally exploit this for harmful purposes, but the same sensitivity can be harnessed to embed subtle, purpose-driven signals. These signals can be used to tag content, enforce model behavior, or convey ownership without altering human-perceived semantics. The differentiable, gradient-based nature of adversarial methods allows precise tuning of perturbations to interact with specific model layers or decision boundaries, enabling fine-grained, model-aware interventions that traditional preprocessing techniques cannot achieve (Adi et al., 2018; Chen et al., 2019a).

Beyond Traditional Protection Methods Unlike passive schemes such as cryptographic hashing, metadata tagging, or steganographic watermarking (Zhu et al., 2018), adversarial perturbations operate at the intersection of input and model behavior. They can be integrated into training or inference processes, learned end-to-end, and dynamically adjusted based on task-specific constraints. This provides several advantages:

- **Robustness:** Adversarial templates are optimized to survive transformations like compression, resizing, or noise

Table 15 Responsible AI considerations for proactive schemes across application domains, summarizing intended purposes, potential misuse, and evaluation robustness

Domain	Intended Use	Potential Misuse	Evaluation Context
Vision Models Defense	Detect and deter image/video manipulation; verify authenticity	Adversarially masking identity or fabricating deepfakes to evade detection	Stress-tested under resizing, cropping, compression, and re-generation
LLM Defense	Trace and authenticate AI-generated text or code	Circumventing watermarks; propagating misinformation	Tested under paraphrase, synonym, and sampling-temperature variations
Authorship Attribution	Establish authorship and ownership of creative content	Falsifying ownership or removing provenance tags	Evaluated for tag removal, backdoor erasure, and fine-tuning attacks
Privacy Protection	Preserve anonymity and sensitive attributes in shared data	Deanonimization or unauthorized re-identification	Robustness measured under blur, crop, recompression, and de-noise attacks
3D Domain	Verify integrity and ownership of 3D assets and point clouds	Embedding hidden payloads or redistributing protected assets	Evaluated under re-meshing, smoothing, decimation, and coordinate noise
Improving Generative Models	Embed verifiable signatures within diffusion/video models for provenance	Forging or stripping generative “passports” or model credentials	Evaluated under fine-tuning, pruning, latent remixing, and frame interpolation

injection, whereas traditional watermarks often degrade or disappear (Kirchenbauer et al., 2023a; Zhang et al., 2024d).

- *Stealth*: Perturbations can be visually or semantically imperceptible, allowing their use in sensitive contexts such as surveillance or text provenance (Yang et al., 2022b).
- *Adaptability*: Because they are model-aware, adversarial schemes can evolve with downstream model changes, unlike fixed handcrafted features.

Design Considerations and Constraints Proactive adversarial methods must meet constraints not typically present in adversarial attack settings. In particular, perturbations must:

- Maintain perceptual quality across modalities (e.g., images, text, audio).
- Generalize across different models, users, or platforms, often in black-box scenarios (Yang et al., 2022b).
- Remain effective under post-processing, editing, or adversarial removal attempts (Rajabi et al., 2021; Kirchenbauer et al., 2023a).

While adversarial perturbations were historically viewed as vulnerabilities, their properties, such as precision, stealth, and model-awareness, make them powerful tools for constructive applications. Proactive schemes reframe these perturbations as programmable signals embedded with intent, offering a promising new direction for attribution, privacy, and robustness in AI.

8 Challenges

The landscape of digital media protection is complex, involving intricate relationships between templates, perturbation processes, and applications. Each step in developing robust security measures presents unique challenges.

Deepfake Detection Defending against deepfakes with binary sequences and positional values requires balancing visual quality, authenticity, and discardability of templates. Challenges include withstanding identity-switching schemes and generalizing to unknown tampering types, while maintaining robustness against image degradation and manual extraction issues (Meng et al., 2022; Sun et al., 2023; Zhang et al., 2024d).

Textual Content Protection Protecting LLM-generated text requires perturbation methods that maintain semantic integrity while being subtle and versatile across different datasets and models. These methods must function without extensive retraining and survive black-box scenarios (Liu et al., 2023b; Yang et al., 2022b). In code protection, preserving naturalness and operational semantics while embedding resilient templates against obfuscator attacks is crucial (Monden et al., 2000).

Neural Network and Ownership Protection Ensuring robust ownership protection in neural networks demands perturbation that resists fine-tuning, pruning, and overwriting. Templates must be deeply integrated to withstand open-sourcing of models and simple code alterations (Darvish Rouhani et al., 2019; Adi et al., 2018; Fernandez et al., 2023), while allowing public verification without losing credibility (Adi et al., 2018).

Adversarial Perturbation and Face Recognition Protection Developing effective DNN fingerprinting and adversarial perturbations involves embedding templates that do not impact model performance and are resistant to removal and overwriting (Chen et al., 2019a). In face recognition, adversarial masks must deceive classifiers while keeping images visually natural, balancing security with user convenience (Rajabi et al., 2021; Yang et al., 2021a).

9 Limitations

Proactive security schemes, while effective, come with significant limitations. These can be broadly categorized into computational demands, robustness against attacks, generalizability, and practical challenges.

Computational Demands Techniques like deepfake defense using fixed bit encodings require substantial computational resources, particularly during pre-processing to adapt signals to various content types (Zhao et al., 2023b). The need for redundant message insertion to ensure template survival under transformations further escalates computational costs (Wen et al., 2023), making these schemes resource-intensive.

Robustness Against Attacks The effectiveness of these schemes is often compromised by their susceptibility to adversarial attacks. For example, attackers can retrain models using the same datasets employed for perturbation, weakening the security of generative models (Zeng et al., 2023). Adversarial noise can disrupt encrypted messages in deepfake defenses (Wang et al., 2021), and template manipulations can significantly alter computational costs and text quality (Kirchenbauer et al., 2023a). Even code protection methods using template embedding in Java files are at risk of being completely negated by additive attacks (Monden et al., 2000).

Generalizability and Semantic Integrity A major challenge for these schemes is their limited generalizability across different types of content. Binary sequences and Fourier key templates, while effective for specific applications, often cause image distortion and fail to work well with dynamic content like video or 3D scenes (Zhang et al., 2024d). LLM perturbation, which uses word tokens and masks, faces difficulties in streaming contexts and struggles with varying textual styles, leading to potential false positives (Kirchenbauer et al., 2023a; Yang et al., 2022b). Additionally, the removal of sentences can weaken the template's presence, further challenging detection mechanisms (Munyer & Zhong, 2023).

Practical Implementation Implementing these schemes also presents difficulties, particularly in balancing security with usability. Methods like sinusoidal signals and random isotropic unit vector perturbations aim to secure user data without

compromising its utility but are often not robust against third-party overwriting (Zhu et al., 2018). Furthermore, maintaining confidentiality in query results for models deployed as internal services is a challenge that attackers can exploit (Hu et al., 2022). This balance between embedding strong security measures and preserving the practical utility of digital content remains a key issue.

In summary, while proactive schemes have introduced innovative means to secure digital content, they are limited by computational demands, vulnerabilities to attacks, and challenges in generalization and implementation.

10 Conclusion

Adversarial perturbations have traditionally been associated with model vulnerability and security risks. In this survey, we shift that perspective and present a unified view of how such perturbations can be repurposed as tools for *proactive schemes*, supporting responsible, privacy-preserving, and attribution-oriented applications across modern AI systems.

To structure this emerging field, we proposed a three-part taxonomy spanning the *data perturbation*, *learning process*, and *application domains*. This framework synthesizes a wide range of techniques that apply adversarial perturbations not to attack, but to embed purpose-driven functionality into data or model behavior.

Our survey shows that proactive adversarial techniques are not just an inversion of traditional attacks, but a growing class of methods with unique capabilities: they can be learned, fine-grained, modality-adaptive, and embedded at training or inference time. These properties make them well-suited for modern concerns such as generative model control, authorship verification, tamper detection, and robust content attribution.

By reframing perturbations from threats to instruments of control and accountability, we highlight a new direction for adversarial learning, one that advances robustness, transparency, and trust in machine learning. We hope this work provides both structure and inspiration for further research at the intersection of adversarial ML and socially beneficial AI.

Data Availability The survey paper outlines a review of methods for proactive schemes. No specific data is used for comparison. Only the papers referenced in the manuscript are used for analysis.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material

is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Abbas, S. M., Aslam, S., Berns, K., & Muhammad, A. (2019). Analysis and improvements in AprilTag based state estimation. *Sensors*, *19*(24), 5480.
- Adi, Y., Baum, C., Cisse, M., Pinkas, B., & Keshet, J. (2018). Turning your weakness into a strength: Watermarking deep neural networks by backdoor. In *27th USENIX security symposium (USENIX Security 18)* (pp. 1615–1631).
- Agrawal, R., & Srikant, R. (2000). Privacy-preserving data mining. In *Proceedings of the 2000 ACM SIGMOD international conference on management of data* (pp. 439–450).
- Ahmadi, M., Norouzi, A., Karimi, N., Samavi, S., & Emami, A. (2020). Redmark: Framework for residual diffusion watermarking based on deep networks. *Expert Systems with Applications*, *146*, Article 113157.
- Ai, Q., Liu, Q., Zhou, Z., Yang, L., & Xie, S. Q. (2009). A new digital watermarking scheme for 3D triangular mesh models. *Signal Processing*, *89*(11), 2159–2170.
- Akhtar, N., Mian, A., Kardan, N., & Shah, M. (2021). Advances in adversarial attacks and defenses in computer vision: A survey. *IEEE Access*, *9*, 155161–155196.
- Akinsiku, A. M. (2025). A comprehensive survey of federated learning approaches for privacy-preserving machine learning. *Tech-Sphere Journal for Pure and Applied Sciences*. <https://doi.org/10.5281/zenodo.1583091>
- Al-Khafaji, H., & Abhayaratne, C. (2019). Graph spectral domain blind watermarking. *ICASSP 2019–2019 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 2492–2496). IEEE.
- Alface, PR., Macq, B., & Cayre, F. (2007). Blind and robust watermarking of 3D models: How to withstand the cropping attack? In *2007 IEEE international conference on image processing* (5, V-465). IEEE.
- Álvarez, H., Leizea, I., & Borro, D. (2012). A new marker design for a robust marker tracking system against occlusions. *Computer Animation and Virtual Worlds*, *23*(5), 503–518.
- Asnani, V., Yin, X., Hassner, T., Liu, S., & Liu, X. (2022). Proactive image manipulation detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 15386–15395).
- Asnani, V., Kumar, A., You, S., & Liu, X. (2023). Probed: Proactive object detection wrapper. *Advances in Neural Information Processing Systems*, *36*, 77993–78005.
- Asnani, V., Yin, X., Hassner, T., & Liu, X. (2023b). MaLP: Manipulation localization using a proactive scheme. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 12343–12352).
- Asnani, V., Yin, X., Hassner, T., & Liu, X. (2023). Reverse engineering of generative models: Inferring model hyperparameters from generated images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(12), 15477–15493.
- Asnani, V., Collomosse, J., Bui, T., Liu, X., & Agarwal, S. (2024). ProMark: Proactive diffusion watermarking for causal attribution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10802–10811).
- Asnani, V., Collomosse, J., Liu, X., & Agarwal, S. (2025). CustomMark: Customization of diffusion models for proactive attribution. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 1512–1522).
- Atli Tekgul, BG., & Asokan, N. (2022). On the effectiveness of dataset watermarking. In *Proceedings of the 2022 ACM on international workshop on security and privacy analytics* (pp. 93–99).
- Aziz, R., Banerjee, S., Bouzeffrane, S., & Le Vinh, T. (2023). Exploring homomorphic encryption and differential privacy techniques towards secure federated learning paradigm. *Future Internet*, *15*(9), 310.
- Bahng, H., Jahanian, A., Sankaranarayanan, S., & Isola, P. (2022). Exploring visual prompts for adapting large-scale models. arXiv preprint [arXiv:2203.17274](https://arxiv.org/abs/2203.17274).
- Balachandran, V., Keong, N. W., & Emmanuel, S. (2014). Function level control flow obfuscation for software security. *2014 eighth international conference on complex, intelligent and software intensive systems* (pp. 133–140). IEEE.
- Bar, A., Gandelsman, Y., Darrell, T., Globerson, A., & Efros, A. (2022). Visual prompting via image inpainting. *Advances in Neural Information Processing Systems*, *35*, 25005–25017.
- Benedens, O. (1999). Geometry-based watermarking of 3D models. *IEEE Computer Graphics and Applications*, *19*(01), 46–55.
- Blum, A., Dwork, C., McSherry, F., & Nissim, K. (2005). Practical privacy: the SuLQ framework. In *Proceedings of the twenty-fourth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems* (pp. 128–138).
- Bors, A. G. (2006). Watermarking mesh-based representations of 3-D objects using local moments. *IEEE Transactions on Image Processing*, *15*(3), 687–701.
- Brazil, G., & Liu, X. (2019). M3D-RPN: Monocular 3D region proposal network for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 9287–9296).
- Cai, M., Liu, H., Mustikovela, S. K., Meyer, G. P., Chai, Y., Park, D., & Lee, Y. J. (2024). ViP-LLaVA: Making large multimodal models understand arbitrary visual prompts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 12914–12923).
- Cao, F., An, B., Wang, J., Ye, D., & Wang, H. (2017). Hierarchical recovery for tampered images based on watermark self-embedding. *Displays*, *46*, 52–60.
- Cao, X., Li, X., Jadav, D., Wu, Y., Chen, Z., Zeng, C., & Wei, W. (2023). Invisible watermarking for audio generation diffusion models. In *2023 5th IEEE international conference on trust, privacy and security in intelligent systems and applications (TPS-ISA)* (pp. 193–202). IEEE.
- Carlini, N., & Wagner, D. (2017). Towards evaluating the robustness of neural networks. In *2017 IEEE symposium on security and privacy (SP)* (pp. 39–57). IEEE.
- Chapelle, O., Haffner, P., & Vapnik, V. N. (1999). Support vector machines for histogram-based image classification. *IEEE Transactions on Neural Networks*, *10*(5), 1055–1064.
- Chen, A., Lorenz, P., Yao, Y., Chen, P. Y., & Liu, S. (2023). Visual prompting for adversarial robustness. In *ICASSP 2023–2023 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 1–5). IEEE.
- Chen, A., Yao, Y., Chen, P. Y., Zhang, Y., & Liu, S. (2023b). Understanding and improving visual prompting: A label-mapping perspective. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 19133–19143).
- Chen, H., Rouhani, B. D., Fu, C., Zhao, J., & Koushanfar, F. (2019a). DeepMarks: A secure fingerprinting framework for digital rights management of deep learning models. In *Proceedings of the 2019 international conference on multimedia retrieval* (pp. 105–113).
- Chen, H., Rouhani, B. D., & Koushanfar, F. (2019b). BlackMarks: Blackbox multibit watermarking for deep neural networks. arXiv preprint [arXiv:1904.00344](https://arxiv.org/abs/1904.00344).

- Chen, L., Liu, J., Ke, Y., Sun, W., Dong, W., & Pan, X. (2023c). MarkNerf: Watermarking for neural radiance field. arXiv preprint [arXiv:2309.11747](https://arxiv.org/abs/2309.11747).
- Chen, L., Liu, J., Sun, W., Dong, W., & Di, F. (2024a). NeRF in NeRF: An implicit representation watermark algorithm for NeRF. Unpublished.
- Chen, P., Liu, Y., Gu, X., Liu, E., Shang, Z., Ji, X., & Liu, W. (2025a). PlugMark: A plug-in zero-watermarking framework for diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 17335–17345).
- Chen, Y., Pan, Y., Yang, H., Yao, T., & Mei, T. (2024b). VP3D: Unleashing 2D visual prompt for text-to-3D generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4896–4905).
- Chen, Y., Vice, J., Akhtar, N., Haldar, N., & Mian, A. (2025b). Dynamic watermarks in images generated by diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference* (pp. 5271–5277).
- Chen, Z., Duan, F., Chapeau-Blondeau, F., & Abbott, D. (2022). Training threshold neural networks by extreme learning machine and adaptive stochastic resonance. *Physics Letters A*, 432, Article 128008.
- Cherepanova, V., Goldblum, M., Foley, H., Duan, S., Dickerson, J., Taylor, G., & Goldstein, T. (2021). Lowkey: Leveraging adversarial attacks to protect social media users from facial recognition. arXiv preprint [arXiv:2101.07922](https://arxiv.org/abs/2101.07922).
- Cho, J. W., Prost, R., & Jung, H. Y. (2007). An oblivious watermarking for 3-D polygonal meshes using distribution of vertex norms. *IEEE Transactions on Signal Processing*, 55(1), 142–155.
- Chou, C. M., & Tseng, D. C. (2006). A public fragile watermarking scheme for 3D model authentication. *Computer-Aided Design*, 38(11), 1154–1165.
- Chou, C. M., & Tseng, D. C. (2009). Affine-transformation-invariant public fragile watermarking for 3D model authentication. *IEEE Computer Graphics and Applications*, 29(2), 72–79.
- Cohen, J., Rosenfeld, E., & Kolter, Z. (2019). Certified adversarial robustness via randomized smoothing. In *International conference on machine learning* (pp. 1310–1320). PMLR.
- Costa, J. C., Roxo, T., Proença, H., & Inacio, P. R. M. (2024). How deep learning sees the world: A survey on adversarial attacks & defenses. *IEEE Access*, 12, 61113–61136.
- Cotting, D., Weyrich, T., Pauly, M., & Gross, M. (2004). Robust watermarking of point-sampled geometry. In *Proceedings shape modeling applications* (pp. 233–242). IEEE.
- Cozzolino, D., & Verdoliva, L. (2019). Noiseprint: A CNN-based camera model fingerprint. *IEEE Transactions on Information Forensics and Security*, 15, 144–159.
- Cui, Y., Ren, J., Xu, H., He, P., Liu, H., Sun, L., Xing, Y., & Tang, J. (2023). DiffusionShield: A watermark for copyright protection against generative diffusion models. arXiv preprint [arXiv:2306.04642](https://arxiv.org/abs/2306.04642).
- Cui, Y., Ren, J., Lin, Y., Xu, H., He, P., Xing, Y., Lyu, L., Fan, W., Liu, H., & Tang, J. (2025). Ft-shield: A watermark against unauthorized fine-tuning in text-to-image diffusion models. *ACM SIGKDD Explorations Newsletter*, 26(2), 76–88.
- Dadkhah, S., Abd Manaf, A., Hori, Y., Hassani, A. E., & Sadeghi, S. (2014). An effective SVD-based image tampering detection and self-recovery using active watermarking. *Signal Processing: Image Communication*, 29(10), 1197–1210.
- Darvish Rouhani, B., Chen, H., & Koushanfar, F. (2019). Deepsigns: An end-to-end watermarking framework for ownership protection of deep neural networks. In *Proceedings of the twenty-fourth international conference on architectural support for programming languages and operating systems* (pp. 485–497).
- Dathathri, S., See, A., Ghaisas, S., Huang, P. S., McAdam, R., Welbl, J., Bachani, V., Kaskasoli, A., Stanforth, R., Matejovicova, T., et al. (2024). Scalable watermarking for identifying large language model outputs. *Nature*, 634(8035), 818–823.
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). Bert: Pretraining of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)* (pp. 4171–4186).
- Dhillon, G. S., Azizzadenesheli, K., Lipton, Z. C., Bernstein, J., Kossaifi, J., Khanna, A., & Anandkumar, A. (2018). Stochastic activation pruning for robust adversarial defense. arXiv preprint [arXiv:1803.01442](https://arxiv.org/abs/1803.01442).
- Dhivya, P., Karthikeyan, A., Pradeep, S., & Umamaheswari, H. (2025). *Natural language processing in generative adversarial network. Generative Artificial Intelligence: Concepts and Applications* (pp. 53–79).
- Ding, D., Qiu, C., Liu, F., & Pan, Z. (2021). Point cloud upsampling via perturbation learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(12), 4661–4672.
- Dong, G., Zhao, J., Hui, T., Guo, D., Wang, W., Feng, B., Qiu, Y., Gongque, Z., He, K., Wang, Z., et al. (2023). Revisit input perturbation problems for LLMs: A unified robustness evaluation framework for noisy slot filling task. In *CCF International conference on natural language processing and chinese computing* (pp. 682–694). Springer.
- Dong, Y., Liao, F., Pang, T., Su, H., Zhu, J., Hu, X., & Li, J. (2018). Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 9185–9193).
- Dong, Y., Su, H., Wu, B., Li, Z., Liu, W., Zhang, T., & Zhu, J. (2019). Efficient decision-based black-box adversarial attacks on face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 7714–7722).
- Ducoffe, M., & Precioso, F. (2018). Adversarial active learning for deep networks: A margin based approach. arXiv preprint [arXiv:1802.09841](https://arxiv.org/abs/1802.09841).
- Dwork, C. (2006). Differential privacy. In *International colloquium on automata, languages, and programming* (pp. 1–12). Springer.
- Fan, L., Ng, K. W., & Chan, C. S. (2019). Rethinking deep neural network ownership verification: Embedding passports to defeat ambiguity attacks. In *Advances in neural information processing systems*, vol. 32.
- Feng, X. (2015). A new watermarking algorithm for point model using angle quantization index modulation. In *2015 4th national conference on electrical, electronics and computer engineering* (pp. 963–968). Atlantis Press.
- Feng, X., Zhang, W., & Liu, Y. (2014). Double watermarks of 3D mesh model based on feature segmentation and redundancy information. *Multimedia Tools and Applications*, 68(3), 497–515.
- Fernandez, P., Couairon, G., Jégou, H., Douze, M., & Furon, T. (2023). The stable signature: Rooting watermarks in latent diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 22466–22477).
- Fiala, M. (2005). Artag, a fiducial marker system using digital techniques. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)* (vol. 2, pp. 590–596). IEEE.
- Fong, R. C., & Vedaldi, A. (2017). Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE international conference on computer vision* (pp. 3429–3437).
- Furon, T., & Desoubreaux, M. (2014). Tardos codes for real. In *2014 IEEE international workshop on information forensics and security (WIFS)* (pp. 24–29). IEEE.
- Gammaitoni, L., Hänggi, P., Jung, P., & Marchesoni, F. (1998). Stochastic resonance. *Reviews of Modern Physics*, 70(1), 223.

- Gao, Y., Shi, X., Zhu, Y., Wang, H., Tang, Z., Zhou, X., Li, M., & Metaxas, D. N. (2022). Visual prompt tuning for test-time domain adaptation. arXiv preprint [arXiv:2210.04831](https://arxiv.org/abs/2210.04831).
- Garrido-Jurado, S., Muñoz-Salinas, R., Madrid-Cuevas, F. J., & Marín-Jiménez, M. J. (2014). Automatic generation and detection of highly reliable fiducial markers under occlusion. *Pattern Recognition*, 47(6), 2280–2292.
- Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 580–587).
- Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. arXiv preprint [arXiv:1412.6572](https://arxiv.org/abs/1412.6572).
- Guo, J., & Potkonjak, M. (2018). Watermarking deep neural networks for embedded systems. In *2018 IEEE/ACM international conference on computer-aided design (ICCAD)* (pp. 1–8). IEEE.
- Guo, Z., & Yu, S. (2023). AuthentiGPT: Detecting machine-generated text via black-box language models denoising. arXiv preprint [arXiv:2311.07700](https://arxiv.org/abs/2311.07700).
- Haghighi, B. B., Taherinia, A. H., & Harati, A. (2018). TRLH: Fragile and blind dual watermarking for image tamper detection and selfrecovery based on lifting wavelet transform and halftoning technique. *Journal of Visual Communication and Image Representation*, 50, 49–64.
- Hamidi, M., El Haziti, M., Cherifi, H., & Aboutajdine, D. (2017). A robust blind 3-D mesh watermarking based on wavelet transform for copyright protection. In *2017 international conference on advanced technologies for signal and image processing (ATSIP)* (pp. 1–6). IEEE.
- Hamidi, M., Chetouani, A., El Haziti, M., El Hassouni, M., & Cherifi, H. (2019). Blind robust 3D mesh watermarking based on mesh saliency and wavelet transform for copyright protection. *Information*, 10(2), 67.
- Han, C., Wang, Q., Cui, Y., Cao, Z., Wang, W., Qi, S., & Liu, D. (2023). E2vpt: An effective and efficient approach for visual prompt tuning. arXiv preprint [arXiv:2307.13770](https://arxiv.org/abs/2307.13770).
- Han, G., Zhang, T., Zhang, Y., Xu, G., Sun, J., & Cao, J. (2022). Verifiable and privacy preserving federated learning without fully trusted centers. *Journal of Ambient Intelligence and Humanized Computing*, 13(3), 1431–1441.
- Han, Q., Lu, S., Wang, W., Qu, H., Li, J., & Gao, Y. (2024). Privacy preserving and secure robust federated learning: A survey. *Concurrency and Computation: Practice and Experience*, 36(13), Article e8084.
- Haralick, R. M., Shanmugam, K., & Dinstein, I. H. (2007). Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics*, 6, 610–621.
- He, H., Liu, Y., Wang, Z., Mao, Y., & Bu, Y. (2024). Theoretically grounded framework for LLM watermarking: A distributionadaptive approach. arXiv preprint [arXiv:2410.02890](https://arxiv.org/abs/2410.02890).
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).
- He, R., Dai, Z., He, S., & Tang, K. (2023). Perturbation-based twostage multi-domain active learning. In *Proceedings of the 32nd ACM international conference on information and knowledge management* (pp. 3933–3937).
- He, X., Xu, Q., Lyu, L., Wu, F., & Wang, C. (2022a). Protecting intellectual property of language generation APIs with lexical watermark. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 10758–10766).
- He, X., Xu, Q., Zeng, Y., Lyu, L., Wu, F., Li, J., & Jia, R. (2022). Cater: Intellectual property protection on text generation APIs via conditional watermarks. *Advances in Neural Information Processing Systems*, 35, 5431–5445.
- Herschel, M., Diestelkämper, R., & Ben Lahmar, H. (2017). A survey on provenance: What for? what form? what from? *The VLDB Journal*, 26(6), 881–906.
- Hou, G., Ou, B., Long, M., & Peng, F. (2023). Separable reversible data hiding for encrypted 3d mesh models based on octree subdivision and multi-MSB prediction. *IEEE Transactions on Multimedia*, 26, 2395–2407.
- Hsu, C. S., & Tu, S. F. (2010). Probability-based tampering detection scheme for digital images. *Optics Communications*, 283(9), 1737–1743.
- Hsu, C. S., & Tu, S. F. (2016). Image tamper detection and recovery using adaptive embedding rules. *Measurement*, 88, 287–296.
- Hu, R., Zhang, J., Li, Y., Li, J., Guo, Q., Qiu, H., & Zhang, T. (2025). Videoshield: Regulating diffusion-based video generation models via watermarking. arXiv preprint [arXiv:2501.14195](https://arxiv.org/abs/2501.14195).
- Hu, S., Liu, X., Zhang, Y., Li, M., Zhang, L. Y., Jin, H., & Wu, L. (2022). Protecting facial privacy: Generating adversarial identity masks via style-robust makeup transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 15014–15023).
- Huang, H., Wang, Y., Chen, Z., Zhang, Y., Li, Y., Tang, Z., Chu, W., Chen, J., Lin, W., & Ma, K. K. (2022). CMUA-Watermark: A cross-model universal adversarial watermark for combating deepfakes. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 989–997).
- Huang, K., Wang, H., & Gao, W. (2024a). FreeAsGuard: Mitigating illegal adaptation of diffusion models via selective tensor freezing. arXiv preprint [arXiv:2405.17472](https://arxiv.org/abs/2405.17472).
- Huang, Q., Li, H., Liao, Y., Hao, Y., & Zhou, P. (2024b). Noise-NeRF: Hide information in neural radiance field using trainable noise. In *International conference on artificial neural networks* (pp. 320–334). Springer Nature Switzerland Cham.
- Huang, S., Papernot, N., Goodfellow, I., Duan, Y., & Abbeel, P. (2017). Adversarial attacks on neural network policies. arXiv preprint [arXiv:1702.02284](https://arxiv.org/abs/1702.02284).
- Jang, Y., Lee, D. I., Jang, M., Kim, J. W., Yang, F., & Kim, S. (2024). WaterRF: Robust watermarks in radiance fields for protection of copyrights. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 12087–12097).
- Jang, Y., Park, H., Yang, F., Ko, H., Choo, E., & Kim, S. (2025). 3D-GSW: 3D Gaussian splatting for robust watermarking. In *Proceedings of the computer vision and pattern recognition conference* (pp. 5938–5948).
- Jia, M., Tang, L., Chen, B. C., Cardie, C., Belongie, S., Hariharan, B., & Lim, S. N. (2022). Visual prompt tuning. In *European conference on computer vision* (pp. 709–727). Springer Nature Switzerland Cham.
- Jiang, Q., Li, F., Zeng, Z., Ren, T., Liu, S., & Zhang, L. (2024). T-Rex2: Towards generic object detection via text-visual prompt synergy. In *European conference on computer vision* (pp. 38–57). Springer Nature Switzerland Cham.
- Jiang, R., Zhou, H., & Zhang, W. Y. N. (2017). Reversible data hiding in encrypted three-dimensional mesh models. *IEEE Transactions on Multimedia*, 20(1), 55–67.
- Jiang, Z., Zhang, J., & Gong, N. Z. (2023). Evading watermark based detection of ai-generated content. In *Proceedings of the 2023 ACM SIGSAC conference on computer and communications security* (pp. 1168–1181).
- Ju, C., Han, T., Zheng, K., Zhang, Y., & Xie, W. (2022). Prompting visual-language models for efficient video understanding. In *European conference on computer vision* (pp. 105–124). Springer Nature Switzerland Cham.
- Kanai, S., Date, H., Kishinami, T., et al. (1998). Digital watermarking for 3D polygons using multiresolution wavelet decomposition. In *Proc. Sixth IFIP WG* (pp. 296–307).

- Kapusta, K., Mattioli, L., Addad, B., & Lansari, M. (2024). Protecting ownership rights of ml models using watermarking in the light of adversarial attacks. *AI and Ethics*, 4(1), 95–103.
- Khamaiseh, S. Y., Bagagem, D., Al-Alaj, A., Mancino, M., & Alomari, H. W. (2022). Adversarial deep learning: A survey on adversarial attacks and defense mechanisms on image classification. *IEEE Access*, 10, 102266–102291.
- Kherraz, H. (2025). On-device watermarking: A socio-technical imperative for authenticity in the age of generative ai. arXiv preprint [arXiv:2504.13205](https://arxiv.org/abs/2504.13205).
- Kiatpapan, S., & Kondo, T. (2015). An image tamper detection and recovery method based on self-embedding dual watermarking. In *2015 12th international conference on electrical engineering/electronics, computer, telecommunications and information technology* (pp. 1–6). IEEE.
- Kim, M. S., Valette, S., Jung, H. Y., & Prost, R. (2005). Watermarking of 3d irregular meshes based on wavelet multiresolution analysis. *International Workshop on Digital Watermarking* (pp. 313–324). Berlin Heidelberg Berlin, Heidelberg: Springer.
- Kim, S., Kim, H. I., & Ro, Y. M. (2024a). Improving open set recognition via visual prompts distilled from common-sense knowledge. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 2786–2794).
- Kim, Y., Li, Y., Moitra, A., Yin, R., & Panda, P. (2024). Do we really need a large number of visual prompts? *Neural Networks*, 177, Article 106390.
- Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I., & Goldstein, T. (2023a). A watermark for large language models. In *International conference on machine learning* (pp. 17061–17084). PMLR.
- Kirchenbauer, J., Geiping, J., Wen, Y., Shu, M., Saifullah, K., Kong, K., Fernando, K., Saha, A., Goldblum, M., & Goldstein, T. (2023b). On the reliability of watermarks for large language models. arXiv preprint [arXiv:2306.04634](https://arxiv.org/abs/2306.04634).
- Kitada, S., & Iyatomi, H. (2021). Attention meets perturbations: Robust and interpretable attention with adversarial training. *IEEE Access*, 9, 92974–92985.
- Komkov, S., & Petiushko, A. (2021). AdvHat: Real-world adversarial attack on ArcFace face ID system. In *2020 25th international conference on pattern recognition (ICPR)* (pp. 819–826). IEEE.
- Kong, T., Yao, A., Chen, Y., & Sun, F. (2016). HyperNet: Towards accurate region proposal generation and joint object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 845–853).
- Krishna, K., Song, Y., Karpinska, M., Wieting, J., & Iyyer, M. (2023). Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense. *Advances in Neural Information Processing Systems*, 36, 27469–27500.
- Krogus, M., Haggemiller, A., & Olson, E. (2019). Flexible layouts for fiducial tags. In *2019 IEEE/RSSJ international conference on intelligent robots and systems (IROS)* (pp. 1898–1903). IEEE.
- Kulthe, Y., Gilbert, A., & Collomosse, J. (2025). MultiNeRF: Multiple watermark embedding for neural radiance fields. arXiv preprint [arXiv:2504.02517](https://arxiv.org/abs/2504.02517).
- Kunananthaseelan, N., Zhang, J., & Harandi, M. (2024). LaViP: Language-grounded visual prompting. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 2840–2848).
- Kuo, C. T., Cheng, S. C., Wu, D. C., Chang, C. C., et al. (2009). A blind robust watermarking scheme for 3D triangular mesh models using 3D edge vertex detection. *Asian Journal of Health and Information Sciences*, 4(1), 36–63.
- Laarhoven, T. (2019). Nearest neighbor decoding for Tardos fingerprinting codes. In *Proceedings of the ACM workshop on information hiding and multimedia security* (pp. 182–187).
- Laouamer, L., AlShaikh, M., Nana, L., & Pascu, A. C. (2015). Robust watermarking scheme and tamper detection based on threshold versus intensity. *Journal of Innovation in Digital Ecosystems*, 2(1–2), 1–12.
- Le Merrer, E., Perez, P., & Trédan, G. (2020). Adversarial frontier stitching for remote neural network watermarking. *Neural Computing and Applications*, 32(13), 9233–9244.
- Lee, S. J., & Cho, N. I. (2025). Semantic watermarking reinvented: Enhancing robustness and generation quality with Fourier integrity. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 18759–18769).
- Lee, S. J., & Jung, S. H. (2001). A survey of watermarking techniques applied to multimedia. In *ISIE 2001. 2001 IEEE international symposium on industrial electronics proceedings (Cat. No. 01TH8570)* (vol. 1, pp. 272–277). IEEE.
- Lee, T. Y., & Lin, S. D. (2008). Dual watermark for image tamper detection and recovery. *Pattern Recognition*, 41(11), 3497–3506.
- Lee, Y. L., Tsai, Y. H., Chiu, W. C., & Lee, C. Y. (2023). Multimodal prompting with missing modalities for visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 14943–14952).
- Lei, L., Gai, K., Yu, J., & Zhu, L. (2024). DiffuseTrace: A transparent and flexible watermarking scheme for latent diffusion model. arXiv preprint [arXiv:2405.02696](https://arxiv.org/abs/2405.02696).
- Li, B., Fan, L., Gu, H., Li, J., & Yang, Q. (2022). FedIPR: Ownership verification for federated deep neural network models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4), 4521–4536.
- Li, C., Feng, B. Y., Fan, Z., Pan, P., & Wang, Z. (2023a). StegaNeRF: Embedding invisible information within neural radiance fields. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 441–453).
- Li, F., Jiang, Q., Zhang, H., Ren, T., Liu, S., Zou, X., Xu, H., Li, H., Yang, J., Li, C., et al. (2024a). Visual in-context prompting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 12861–12871).
- Li, G., Wu, W., Sun, Y., Shen, L., Wu, B., & Tao, D. (2023b). Visual prompt based personalized federated learning. arXiv preprint [arXiv:2303.08678](https://arxiv.org/abs/2303.08678).
- Li, K., Wu, Z., Peng, K. C., Ernst, J., & Fu, Y. (2018). Tell me where to look: Guided attention inference network. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 9215–9223).
- Li, K., Huang, Z., Hou, X., & Hong, C. (2025). GaussMarker: Robust dual-domain watermark for diffusion models. arXiv preprint [arXiv:2506.11444](https://arxiv.org/abs/2506.11444).
- Li, P., Li, D., Li, W., Gong, S., Fu, Y., & Hospedales, T. M. (2021a). A simple feature augmentation for domain generalization. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 8886–8895).
- Li, S., Chen, K., Tang, K., Huang, W., Zhang, J., Zhang, W., & Yu, N. (2023c). FunctionMarker: watermarking language datasets via knowledge injection. CoRR.
- Li, T., & Lin, L. (2019). AnonymousNet: Natural face de-identification with measurable privacy. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 0–0.
- Li, X., Yang, P., Gu, Y., Zhan, X., Wang, T., Xu, M., & Xu, C. (2024b). Deep active learning with noise stability. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 13655–13663).
- Li, Y., Zhang, Z., Bai, J., Wu, B., Jiang, Y., & Xia, S. T. (2020). Open-sourced dataset protection via backdoor watermarking. arXiv preprint [arXiv:2010.05821](https://arxiv.org/abs/2010.05821).
- Li, Y., Liu, P., Jiang, Y., & Xia, S. T. (2021). Visual privacy protection via mapping distortion. *ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 3740–3744). IEEE.

- Li, Y., Zhu, M., Yang, X., Jiang, Y., & Xia, S. T. (2022b). Black-box ownership verification for dataset protection via backdoor watermarking. arXiv preprint [arXiv:2209.06015](https://arxiv.org/abs/2209.06015).
- Li, Y., Tsai, Y. L., Yu, C. M., Chen, P. Y., & Ren, X. (2023d). Exploring the benefits of visual prompting in differential privacy. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 5158–5167).
- Li, Y., Zhu, M., Yang, X., Jiang, Y., Wei, T., & Xia, S. T. (2023). Blackbox dataset ownership verification via backdoor watermarking. *IEEE Transactions on Information Forensics and Security*, 18, 2318–2332.
- Liang, H., He, E., Zhao, Y., Jia, Z., & Li, H. (2022). Adversarial attack and defense: A survey. *Electronics*, 11(8), 1283.
- Liang, Q., Li, Q., Nie, W., & Liu, A. A. (2022). PAGN: Perturbation adaptation generation network for point cloud adversarial defense. *Multimedia Systems*, 28(3), 851–859.
- Liang, X., Niu, M., Han, J., Xu, H., Xu, C., & Liang, X. (2023). Visual exemplar driven task-prompting for unified perception in autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9611–9621).
- Lim, J. H., Chan, C. S., Ng, K. W., Fan, L., & Yang, Q. (2022). Protect, show, attend and tell: Empowering image captioning models with ownership protection. *Pattern Recognition*, 122, Article 108285.
- Lin, C. C., Huang, Y., & Tai, W. L. (2017). A novel hybrid image authentication scheme based on absolute moment block truncation coding. *Multimedia Tools and Applications*, 76(1), 463–488.
- Liu, A., Pan, L., Lu, Y., Li, J., Hu, X., Zhang, X., Wen, L., King, I., Xiong, H., & Yu, P. (2024). A survey of text watermarking in the era of large language models. *ACM Computing Surveys*, 57(2), 1–36.
- Liu, C. C., Chen, J. Y., Chung, P. C., Yu, S. S., & Tsui, T. S. (2012). A three dimensional model watermarking algorithm in frequency domain based on the normalization of host models. *International Journal of Innovative Computing, Information and Control*, 8(5A), 3299–3314.
- Liu, H., Weng, Z., & Zhu, Y. (2021). Watermarking deep neural networks with greedy residuals. *ICML*, 139, 6978–6988.
- Liu, J., Yang, Y., Ma, D., He, W., & Wang, Y. (2019). A novel watermarking algorithm for three-dimensional point-cloud models based on vertex curvature. *International Journal of Distributed Sensor Networks*, 15(1), 1550147719826042.
- Liu, K., Yang, H., Ma, Y., Tan, B., Yu, B., Young, E. F., Karri, R., & Garg, S. (2020). Adversarial perturbation attacks on ML-based CAD: A case study on CNN-based lithographic hotspot detection. *ACM Transactions on Design Automation of Electronic Systems (TODAES)*, 25(5), 1–31.
- Liu, S., & Kong, L. (2018). Local chaotic encryption based on privacy protection on video surveillance. In *2018 8th international conference on social science and education research (SSER 2018)* (pp. 390–393). Atlantis Press.
- Liu, W., Shen, X., Pun, C. M., & Cun, X. (2023a). Explicit visual prompting for low-level structure segmentations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 19434–19445).
- Liu, X., Shao, S., Yang, Y., Wu, K., Yang, W., & Fang, H. (2021b). Secure federated learning model verification: A client-side backdoor triggered watermarking scheme. In *2021 IEEE international conference on systems, man, and cybernetics (SMC)* (pp. 2414–2419). IEEE.
- Liu, Y., & Bu, Y. (2024). Adaptive text watermark for large language models. arXiv preprint [arXiv:2401.13927](https://arxiv.org/abs/2401.13927).
- Liu, Y., Prabhakaran, B., & Guo, X. (2012). Spectral watermarking for parameterized surfaces. *IEEE Transactions on Information Forensics and Security*, 7(5), 1459–1471.
- Liu, Y., Hu, H., Zhang, X., & Sun, L. (2023b). Watermarking text data on large language models for dataset copyright protection. arXiv preprint [arXiv:2305.13257](https://arxiv.org/abs/2305.13257).
- Long, Y., Han, J., Huang, R., Xu, H., Zhu, Y., Xu, C., & Liang, X. (2023). Fine-grained visual-text prompt-driven self-training for open-vocabulary object detection. *IEEE Transactions on Neural Networks and Learning Systems*, 35, 16277–16287.
- Lu, C. S., & Liao, H. Y. (2001). Multipurpose watermarking for image authentication and protection. *IEEE Transactions on Image Processing*, 10(10), 1579–1592.
- Lu, S., Zhou, Z., Lu, J., Zhu, Y., & Kong, A. W. K. (2024). Robust watermarking using generative priors against image editing: From benchmarking to advances. arXiv preprint [arXiv:2410.18775](https://arxiv.org/abs/2410.18775).
- Luo, Z., Guo, Q., Cheung, K. C., See, S., & Wan, R. (2023). CopyRNeRF: Protecting the copyright of neural radiance fields. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 22401–22411).
- Ma, K., Fu, Y., Cao, C., Hou, S., Huang, Y., & Zheng, D. (2024). Learning visual prompt for gait recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 593–603).
- Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2017). Towards deep learning models resistant to adversarial attacks. arXiv preprint [arXiv:1706.06083](https://arxiv.org/abs/1706.06083).
- Malik, A., Kuribayashi, M., Abdullahi, S. M., & Khan, A. N. (2022). Deepfake detection for human face images and videos: A survey. *IEEE Access*, 10, 18757–18775.
- Maniparambil, M., Vorster, C., Molloy, D., Murphy, N., McGuinness, K., & O'Connor, N. E. (2023). Enhancing clip with GPT-4: Harnessing visual descriptions as prompts. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 262–271).
- Medimegh, N., Belaid, S., Atri, M., & Werghi, N. (2018). 3D mesh watermarking using salient points. *Multimedia Tools and Applications*, 77(24), 32287–32309.
- Meng, D., & Chen, H. (2017). MagNet: A two-pronged defense against adversarial examples. In *Proceedings of the 2017 ACM SIGSAC conference on computer and communications security* (pp. 135–147).
- Meng, R., Zhou, Z., Cui, Q., Lam, K. Y., & Kot, A. (2022). Traceable and authenticable image tagging for fake news detection. arXiv preprint [arXiv:2211.10923](https://arxiv.org/abs/2211.10923).
- Meng, Z., Peng, B., & Dong, J. (2025). Latent watermark: Inject and detect watermarks in latent diffusion space. *IEEE Transactions on Multimedia*, 27, 3399–3410.
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint [arXiv:1301.3781](https://arxiv.org/abs/1301.3781).
- Mirjalili, V., & Ross, A. (2017). Soft biometric privacy: Retaining biometric utility of face images while perturbing gender. In *2017 IEEE International joint conference on biometrics (IJCB)* (pp. 564–573). IEEE.
- Mirjalili, V., Raschka, S., Namboodiri, A., & Ross, A. (2018). Semi-adversarial networks: Convolutional autoencoders for imparting privacy to face images. In *2018 International Conference on Biometrics (ICB)* (pp. 82–89). IEEE.
- Mirza, M. J., Karlinsky, L., Lin, W., Doveh, S., Micorek, J., Kozinski, M., Kuehne, H., & Possegger, H. (2024). Meta-prompting for automating zero-shot visual recognition with LLMs. In *European Conference on Computer Vision* (pp. 370–387). Cham: Springer Nature Switzerland.
- Mohiuddin, S., Malakar, S., Kumar, M., & Sarkar, R. (2023). A comprehensive survey on state-of-the-art video forgery detection techniques. *Multimedia Tools and Applications*, 82(22), 33499–33539.

- Molaei, A. M., Ebrahimnezhad, H., & Sedaaghi, M. H. (2013). A blind fragile watermarking method for 3D models based on geometric properties of triangles. *3D Research*, 4(4), 1–9.
- Molaei, A. M., Ebrahimnezhad, H., & Sedaaghi, M. H. (2016). Robust and blind 3D mesh watermarking in spatial domain based on faces categorization and sorting. *3D Research*, 7(2), 11.
- Monden, A., Iida, H., Matsumoto, Ki., Inoue, K., & Torii, K. (2000). A practical method for watermarking java programs. In *Proceedings 24th annual international computer software and applications conference. COMPSAC2000* (pp. 191–197). IEEE.
- Mousavi, S. M., Naghsh, A., & Abu-Bakar, S. (2014). Watermarking techniques used in medical images: A survey. *Journal of Digital Imaging*, 27(6), 714–729.
- Mubarak, R., Alsbouei, T., Alshaiikh, O., Inuwa-Dutse, I., Khan, S., & Parkinson, S. (2023). A survey on the detection and impacts of deepfakes in visual, audio, and textual formats. *IEEE Access*, 11, 144497–144529.
- Müller, A., Lukovnikov, D., Thietke, J., Fischer, A., & Quring, E. (2025). Black-box forgery attacks on semantic watermarks for diffusion models. In *Proceedings of the computer vision and pattern recognition conference* (pp. 20937–20946).
- Mun, S. M., Jang, H. U., Kim, D. G., Choi, S., & Lee, H. K. (2015). A robust 3D mesh watermarking scheme against cropping. In *2015 international conference on 3D imaging (IC3D)* (pp. 1–6). IEEE.
- Munyer, T., & Zhong, X. (2023). DeepTextMark: Deep learning based text watermarking for detection of large language model generated text. arXiv e-prints pp. arXiv:2305.
- Nagai, Y., Uchida, Y., Sakazawa, S., & Satoh, S. (2018). Digital watermarking for deep neural networks. *International Journal of Multimedia Information Retrieval*, 7(1), 3–16.
- Nakazawa, S., Kasahara, S., & Takahashi, S. (2010). A visually enhanced approach to watermarking 3D models. In *2010 Sixth International Conference on Intelligent Information Hiding and Multimedia Signal Processing* (pp. 110–113). IEEE.
- Nasiriany, S., Xia, F., Yu, W., Xiao, T., Liang, J., Dasgupta, I., Xie, A., Driess, D., Wahid, A., Xu, Z., et al. (2024). Pivot: Iterative visual prompting elicits actionable knowledge for vlms. arXiv preprint arXiv:2402.07872.
- Neekhar, P., Hussain, S., Zhang, X., Huang, K., McAuley, J., & Koushanfar, F. (2022). FaceSigns: Semi-fragile neural watermarks for media authentication and countering deepfakes. arXiv preprint arXiv:2204.01960.
- Newton, E. M., Sweeney, L., & Malin, B. (2005). Preserving privacy by de-identifying face images. *IEEE transactions on Knowledge and Data Engineering*, 17(2), 232–243.
- Nguyen, T., Li, Y., Ojha, U., & Lee, Y. J. (2023). Visual instruction inversion: Image editing via image prompting. *Advances in Neural Information Processing Systems*, 36, 9598–9613.
- Nishi, K., Ding, Y., Rich, A., & Hollerer, T. (2021). Augmentation strategies for learning with noisy labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 8022–8031).
- Oh, C., Hwang, H., Lee, Hy., Lim, Y., Jung, G., Jung, J., Choi, H., & Song, K. (2023). BlackVIP: Black-box visual prompting for robust transfer learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 24224–24235).
- Ohbuchi, R., Mukaiyama, A., & Takahashi, S. (2002). A frequency-domain approach to watermarking 3D shapes. In *Computer graphics forum* (pp. 373–382). Wiley Online Library.
- Ohbuchi, R., Masuda, H., & Aono, M. (1998). Data embedding algorithms for geometrical and non-geometrical targets in three-dimensional polygonal models. *Computer Communications*, 21(15), 1344–1354.
- Ohbuchi, R., Masuda, H., & Aono, M. (1998). Watermarking three-dimensional polygonal models through geometric and topological modifications. *IEEE Journal on Selected Areas in Communications*, 16(4), 551–560.
- Ohbuchi, R., Takahashi, S., Miyazawa, T., & Mukaiyama, A. (2001). Watermarking 3d polygonal meshes in the mesh spectral domain. *Graphics interface*, 2001, 9–17.
- Ohbuchi, R., Mukaiyama, A., & Takahashi, S. (2004). Watermarking a 3D shape model defined as a point set. In *2004 international conference on cyberworlds* (pp. 392–399). IEEE.
- Olson, E. (2011). Apriltag: A robust and flexible visual fiducial system. In *2011 IEEE international conference on robotics and automation* (pp. 3400–3407). IEEE.
- Ong, D. S., Chan, C. S., Ng, K. W., Fan, L., & Yang, Q. (2021). Protecting intellectual property of generative adversarial networks from ambiguity attacks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3630–3639).
- Othman, A., & Ross, A. (2014). Privacy of facial soft biometrics: Suppressing gender but retaining identity. In *European conference on computer vision* (pp. 682–696). Cham: Springer International Publishing.
- Ozdag, M. (2018). Adversarial attacks and defenses against deep neural networks: A survey. *Procedia Computer Science*, 140, 152–161.
- Pan, B., Stakhanova, N., & Ray, S. (2023). Data provenance in security and privacy. *ACM Computing Surveys*, 55(14s), 1–35.
- Pang, Q., Hu, S., Zheng, W., & Smith, V. (2024). No free lunch in LLM watermarking: Trade-offs in watermarking design choices. *Advances in Neural Information Processing Systems*, 37, 138756–138788.
- Park, S., & Byun, H. (2024). Fair-VPT: Fair visual prompt tuning for image classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 12268–12278).
- Paruchuri, J., Cheung, S. C., & Hail, M. (2009). Video data hiding for managing privacy information in surveillance systems. *EURASIP Journal on Information Security*, 1, Article 236139.
- Pei, W., Xia, T., Chen, F., Li, J., Tian, J., & Lu, G. (2024). Sa2vp: Spatially aligned-and-adapted visual prompt. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 4450–4458).
- Peng, F., Liao, T., & Long, M. (2022). A semi-fragile reversible watermarking for authenticating 3d models in dual domains based on variable direction double modulation. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(12), 8394–8408.
- Peng, S., Chen, Y., Wang, C., & Jia, X. (2025). Intellectual property protection of diffusion models via the watermark diffusion process. *International Conference on Web Information Systems Engineering* (pp. 290–305). Singapore: Springer.
- Peng, Z., Li, S., Chen, G., Zhang, C., Zhu, H., & Xue, M. (2022b). Fingerprinting deep neural networks globally via universal adversarial perturbations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 13430–13439).
- Petitcolas, F. A., Anderson, R. J., Kuhn, M. G. (1998). Attacks on copyright marking systems. In *International workshop on information hiding* (pp. 218–238). Springer.
- Pham, G. N., Lee, S. H., Kwon, O. H., & Kwon, K. R. (2018). A 3d printing model watermarking algorithm based on 3d slicing and feature points. *Electronics*, 7(2), 23.
- Philip, P., & Minhas, S. (2022). A brief survey on natural language processing based text generation and evaluation techniques. *VFAST Transactions on Software Engineering*, 10(3), 24–36.
- Potdar, V. M., Han, S., & Chang, E. (2005). A survey of digital image watermarking techniques. In *INDIN'05. 2005 3rd IEEE international conference on industrial informatics, 2005* (pp. 709–716). IEEE.
- Praun, E., Hoppe, H., & Finkelstein, A. (1999). Robust mesh watermarking. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques* (pp. 49–56).

- Qin, C., Ji, P., Zhang, X., Dong, J., & Wang, J. (2017). Fragile image watermarking with pixel-wise recovery based on overlapping embedding strategy. *Signal processing*, *138*, 280–293.
- Qureshi, S. M., Saeed, A., Almotiri, S. H., Ahmad, F., & Al Ghamdi, M. A. (2024). Deepfake forensics: a survey of digital forensic methods for multimodal deepfake identification on social media. *PeerJ Computer Science*, *10*, Article e2037.
- Rajabi, A., Bobba, R. B., Rosulek, M., Wright, C., & Feng, Wc. (2021). On the (im) practicality of adversarial perturbation for image privacy. Proceedings on Privacy Enhancing Technologies.
- Rallabandi, V. S., & Roy, P. K. (2010). Magnetic resonance image enhancement using stochastic resonance in Fourier domain. *Magnetic Resonance Imaging*, *28*(9), 1361–1373.
- Ren, J., Xu, H., Liu, Y., Cui, Y., Wang, S., Yin, D., & Tang, J. (2024). A robust semantics-based watermark for large language model against paraphrasing. *Findings of the Association for Computational Linguistics: NAACL, 2024*, 613–625.
- Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards realtime object detection with region proposal networks. In *Advances in neural information processing systems* (Vol. 28).
- Ren, Y., Duan, F., Chapeau-Blondeau, F., & Abbott, D. (2024). Self-gating stochastic-resonance-based autoencoder for unsupervised learning. *Physical Review E*, *110*(1), Article 014107.
- Rizzo, S. G., Bertini, F., & Montesi, D. (2019). Fine-grain watermarking for intellectual property protection. *EURASIP Journal on Information Security*, *1*, 10.
- Robey, A., Wong, E., Hassani, H., & Pappas, G. J. (2023). SmoothLLM: Defending large language models against jailbreaking attacks. arXiv preprint [arXiv:2310.03684](https://arxiv.org/abs/2310.03684).
- Romero-Ramire, F. J., Munoz-Salinas, R., & Medina-Carnicer, R. (2019). Fractal markers: A new approach for long-range marker pose estimation under occlusion. *IEEE Access*, *7*, 169908–169919.
- Romero-Ramirez, F. J., Muñoz-Salinas, R., & Medina-Carnicer, R. (2018). Speeded up detection of squared fiducial markers. *Image and Vision Computing*, *76*, 38–47.
- Ruiz, N., Bargal, S. A., & Sclaroff, S. (2020). Disrupting DeepFakes: Adversarial attacks against conditional image translation networks and facial manipulation systems. In *European conference on computer vision* (pp. 236–251). Springer.
- Sablayrolles, A., Douze, M., Schmid, C., & Jégou, H. (2020). Radioactive data: Tracing through training. In *International Conference on Machine Learning* (pp. 8326–8335). PMLR.
- Sadasivan, V. S., Kumar, A., Balasubramanian, S., Wang, W., & Feizi, S. (2023). Can ai-generated text be reliably detected? arXiv preprint [arXiv:2303.11156](https://arxiv.org/abs/2303.11156).
- Salman, H., Ilyas, A., Engstrom, L., Vemprala, S., Madry, A., & Kapoor, A. (2021). Unadversarial examples: Designing objects for robust vision. *Advances in Neural Information Processing Systems*, *34*, 15270–15284.
- Sayyad, S., Kulkarni, D., Shikalgar, A., & Mulla, T. A. (2024). An exhaustive survey on privacy preserving machine learning using homomorphic encryption and secure multiparty computation techniques. *Journal of Computational Analysis & Applications*, *33*(5), 636–648.
- Segalis, E., & Galili, E. (2020). OGAN: Disrupting deepfakes with an adversarial attack that survives training. arXiv preprint [arXiv:2006.12247](https://arxiv.org/abs/2006.12247).
- Semenov, V. V., & Zakharova, A. (2022). Multiplexing-based control of stochastic resonance. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, *32*(12), Article 121106.
- Shan, S., Wenger, E., Zhang, J., Li, H., Zheng, H., & Zhao, B. Y. (2020). Fawkes: Protecting privacy against unauthorized deep learning models. In *29th USENIX security symposium (USENIX Security 20)* (pp. 1589–1604).
- Shen, M., Yang, J., Jiang, W., Sanjuán, M. A., & Zheng, Y. (2022). Stochastic resonance in image denoising as an alternative to traditional methods and deep learning. *Nonlinear Dynamics*, *109*(3), 2163–2183.
- Shen, S., Yang, S., Zhang, T., Zhai, B., Gonzalez, J. E., Keutzer, K., & Darrell, T. (2024). Multitask vision-language prompt tuning. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 5656–5667).
- Shi, Y., & Sagduyu, Y. E. (2017). Evasion and causative attacks with adversarial deep learning. In *MILCOM 2017–2017 IEEE military communications conference (MILCOM)* (pp. 243–248). IEEE.
- Simmhan, Y. L., Plale, B., Gannon, D., et al. (2005). A survey of data provenance techniques. Computer Science Department, Indiana University, Bloomington IN (Vol. 47405, p. 69).
- Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint [arXiv:1409.1556](https://arxiv.org/abs/1409.1556).
- Singh, D., & Singh, S. K. (2016). Effective self-embedding watermarking scheme for image tampered detection and localization with recovery capability. *Journal of Visual Communication and Image Representation*, *38*, 775–789.
- Singh, D., & Singh, S. K. (2017). DCT based efficient fragile watermarking scheme for image authentication and restoration. *Multimedia Tools and Applications*, *76*(1), 953–977.
- Singh, P., & Chadha, R. S. (2013). A survey of digital watermarking techniques, applications and attacks. *International Journal of Engineering and Innovative Technology (IJEIT)*, *2*(9), 165–175.
- Sohn, K., Chang, H., Lezama, J., Polania, L., Zhang, H., Hao, Y., Essa, I., & Jiang, L. (2023). Visual prompt tuning for generative transfer learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 19840–19851).
- Song, Z., Gong, X., Hu, G., & Zhao, C. (2023). Deep perturbation learning: Enhancing the network performance via image perturbations. In *International conference on machine learning* (pp. 32273–32287). PMLR.
- Sperduti, G., & Moreo, A. (2025). Misspellings in natural language processing: A survey. arXiv preprint [arXiv:2501.16836](https://arxiv.org/abs/2501.16836).
- Sun, P., Qi, H., Li, Y., & Lyu, S. (2023). FakeTracer: Catching faceswap deepfakes via implanting traces in training. arXiv preprint [arXiv:2307.14593](https://arxiv.org/abs/2307.14593).
- Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). Intriguing properties of neural networks. arXiv preprint [arXiv:1312.6199](https://arxiv.org/abs/1312.6199).
- Szyller, S., Atli, B. G., Marchal, S., & Asokan, N. (2021). Dawn: Dynamic adversarial watermarking of neural networks. In *Proceedings of the 29th ACM international conference on multimedia* (pp. 4417–4425).
- Tang, L., Ye, D., Lv, Y., Chen, C., & Zhang, Y. (2024a). Once and for all: Universal transferable adversarial perturbation against deep hashing-based facial image retrieval. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 5136–5144).
- Tang, L., Ye, Q., Hu, H., Xue, Q., Xiao, Y., & Li, J. (2024). DeepMark: A scalable and robust framework for deepfake video detection. *ACM Transactions on Privacy and Security*, *27*(1), 1–26.
- Tekgul, B. G., Xia, Y., Marchal, S., & Asokan, N. (2021). Waffle: Watermarking in federated learning. In *2021 40th international symposium on reliable distributed systems (SRDS)* (pp. 310–320). IEEE.
- Teng, H., Quan, Y., Wang, C., Huang, J., & Ji, H. (2025). Fingerprinting denoising diffusion probabilistic models. In *Proceedings of the computer vision and pattern recognition conference* (pp. 28811–28820).
- Tsai, Y. Y., & Liu, H. L. (2022). Integrating coordinate transformation and random sampling into high-capacity reversible data hiding in encrypted polygonal models. *IEEE Transactions on Dependable and Secure Computing*, *20*(4), 3508–3519.

- Tsai, Y. Y., Mao, C., & Yang, J. (2023). Convolutional visual prompt for robust visual perception. *Advances in Neural Information Processing Systems*, 36, 27897–27921.
- Tsao, H. A., Hsiung, L., Chen, P. Y., Liu, S., & Ho, T. Y. (2023). Autovp: An automated visual prompting framework and benchmark. arXiv preprint [arXiv:2310.08381](https://arxiv.org/abs/2310.08381).
- Uccheddu, F., Corsini, M., & Barni, M. (2004). Wavelet-based blind watermarking of 3D models. In *Proceedings of the 2004 workshop on multimedia and security* (pp. 143–154).
- Uchida, Y., Nagai, Y., Sakazawa, S., & Satoh, S. (2017). Embedding watermarks into deep neural networks. In *Proceedings of the 2017 ACM on international conference on multimedia retrieval* (pp. 269–277).
- Van Le, T., Phung, H., Nguyen, T. H., Dao, Q., Tran, N. N., & Tran, A. (2023). Anti-DreamBooth: Protecting users from personalized text-to-image synthesis. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 2116–2127).
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (vol. 30).
- Venugopal, A., Uszkoreit, J., Talbot, D., Och, F. J., & Ganitkevitch, J. (2011). Watermarking the outputs of structured prediction with an application in statistical machine translation. In *Proceedings of the 2011 conference on empirical methods in natural language processing* (pp. 1363–1372).
- Verdoliva, L. (2020). Media forensics and deepfakes: An overview. *IEEE Journal of Selected Topics in Signal Processing*, 14(5), 910–932.
- Wagner, D., & Schmalstieg, D. (2007). Artoolkitplus for pose tracking on mobile devices. Unpublished.
- Wang, D., Yao, W., Jiang, T., Zhou, W., Lin, L., & Chen, X. (2023a). A plug-and-play defensive perturbation for copyright protection of DNN-based applications. arXiv preprint [arXiv:2304.10679](https://arxiv.org/abs/2304.10679).
- Wang, F., Zhou, H., Fang, H., Zhang, W., & Yu, N. (2022). Deep 3D mesh watermarking with self-adaptive robustness. *Cybersecurity*, 5(1), 24.
- Wang, F., Zhou, H., Zhang, W., & Yu, N. (2022b). Neural watermarking for 3D morphable models. In *international conference on artificial intelligence and security* (pp. 336–349). Springer International Publishing Cham.
- Wang, F., Huang, W., Yang, S., Fan, Q., & Lan, L. (2024a). Learning to learn better visual prompts. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 5354–5363).
- Wang, H., Shi, Z., Lu, G., & Zhong, Y. (2018). Hierarchical fiducial marker design for pose estimation in large-scale scenarios. *Journal of Field Robotics*, 35(6), 835–849.
- Wang, H., Fang, H., Wang, S. L., & Chang, E. C. (2025a). ROAR: Reducing inversion error in generative image watermarking. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 19742–19751).
- Wang, J., & Olson, E. (2016). AprilTag 2: Efficient and robust fiducial detection. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 4193–4198). IEEE.
- Wang, J., Wu, H., Zhang, X., & Yao, Y. (2020). Watermarking in deep neural networks via error back-propagation. *Electronic Imaging*, 32, 1–9.
- Wang, J., Yin, Z., Hu, P., Liu, A., Tao, R., Qin, H., Liu, X., & Tao, D. (2022c). Defensive patches for robust recognition in the physical world. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2456–2465).
- Wang, K., Lavoué, G., Denis, F., & Baskurt, A. (2008). Hierarchical watermarking of semiregular meshes based on wavelet transform. *IEEE Transactions on Information Forensics and Security*, 3(4), 620–634.
- Wang, K., Lavoué, G., Denis, F., & Baskurt, A. (2011). Robust and blind mesh watermarking based on volume moments. *Computers & Graphics*, 35(1), 1–19.
- Wang, R., Juefei-Xu, F., Luo, M., Liu, Y., & Wang, L. (2021). Fake-Tagger: Robust safeguards against deepfake dissemination via provenance tracking. In *Proceedings of the 29th ACM international conference on multimedia* (pp. 3546–3555).
- Wang, T., & Kerschbaum, F. (2021). RIGA: Covert and robust white-box watermarking of deep neural networks. *Proceedings of the web conference, 2021*, 993–1004.
- Wang, T., Zhang, X., Zhou, Y., Chen, Y., Zhao, L., Tan, T., & Tong, T. (2025). PCDAL: A perturbation consistency-driven active learning approach for medical image segmentation and classification. *IEEE Transactions on Emerging Topics in Computational Intelligence*. <https://doi.org/10.1109/TETCI.2025.3547635>
- Wang, W. B., Zheng, G. Q., Yong, J. H., & Gu, H. J. (2008). A numerically stable fragile watermarking scheme for authenticating 3D models. *Computer-Aided Design*, 40(5), 634–645.
- Wang, Y., Cheng, J., Chen, Y., Shao, S., Zhu, L., Wu, Z., Liu, T., & Zhu, H. (2023). FVP: Fourier visual prompting for source-free unsupervised domain adaptation of medical image segmentation. *IEEE Transactions on Medical Imaging*, 42(12), 3738–3751.
- Wang, Y., Cheng, L., Fang, C., Zhang, D., Duan, M., & Wang, M. (2024b). Revisiting the power of prompt for visual tuning. arXiv preprint [arXiv:2402.02382](https://arxiv.org/abs/2402.02382).
- Wang, Z., Guo, J., Zhu, J., Li, Y., Huang, H., Chen, M., & Tu, Z. (2025c). SleeperMark: Towards robust watermark against fine-tuning text-to-image diffusion models. In *Proceedings of the computer vision and pattern recognition conference* (pp. 8213–8224).
- Wen, J., Luo, Y., Fei, N., Yang, G., Lu, Z., Jiang, H., Jiang, J., & Cao, Z. (2022). Visual prompt tuning for few-shot text classification. In *Proceedings of the 29th international conference on computational linguistics* (pp. 5560–5570).
- Wen, Y., Kirchenbauer, J., Geiping, J., & Goldstein, T. (2023). Tree-ring watermarks: Fingerprints for diffusion images that are invisible and robust. arXiv preprint [arXiv:2305.20030](https://arxiv.org/abs/2305.20030).
- Westfeld, A., & Pfitzmann, A. (1999). Attacks on steganographic systems: Breaking the Steganographic Utilities EzStego, Jsteg, Steganos, and S-Tools-and Some Lessons Learned. In *International workshop on information hiding* (pp. 61–76). Springer.
- Wong, E., & Kolter, J. Z. (2020). Learning perturbation sets for robust machine learning. arXiv preprint [arXiv:2007.08450](https://arxiv.org/abs/2007.08450).
- Wu, H., Liu, G., Yao, Y., & Zhang, X. (2020). Watermarking neural networks with watermarked images. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(7), 2591–2601.
- Wu, J., Li, X., Wei, C., Wang, H., Yuille, A., Zhou, Y., & Xie, C. (2022a). Unleashing the power of visual prompting at the pixel level. arXiv preprint [arXiv:2212.10556](https://arxiv.org/abs/2212.10556).
- Wu, R., Wang, Y., Shi, H., Yu, Z., Wu, Y., & Liang, D. (2023a). Towards prompt-robust face privacy protection via adversarial decoupling augmentation framework. arXiv preprint [arXiv:2305.03980](https://arxiv.org/abs/2305.03980).
- Wu, T., Jiang, E., Donsbach, A., Gray, J., Molina, A., Terry, M., & Cai, C. J. (2022b). PromptChainer: Chaining large language model prompts through visual programming. In *CHI conference on human factors in computing systems extended abstracts* (pp. 1–10).
- Wu, X., Liao, X., & Ou, B. (2023b). SepMark: Deep separable watermarking for unified source tracing and deepfake detection. In *Proceedings of the 31st ACM international conference on multimedia* (pp. 1190–1201).
- Wu, X., Liao, X., Ou, B., Liu, Y., & Qin, Z. (2024). Are watermarks bugs for deepfake detectors? Rethinking proactive forensics. arXiv preprint [arXiv:2404.17867](https://arxiv.org/abs/2404.17867).
- Wu, Y., Yang, F., Xu, Y., & Ling, H. (2019). Privacy-protective-GAN for privacy preserving face de-identification. *Journal of Computer Science and Technology*, 34(1), 47–60.

- Xiao, Z., Gao, X., Fu, C., Dong, Y., Gao, W., Zhang, X., Zhou, J., & Zhu, J. (2021). Improving transferability of adversarial patches on face recognition with generative models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11845–11854).
- Xiaoqing, F. (2015). A watermarking for 3D point cloud model using distance normalization modulation. In *2015 4th international conference on computer science and network technology (ICCSNT)* (vol 1, pp. 1449–1452), IEEE.
- Xie, C., Wang, J., Zhang, Z., Ren, Z., & Yuille, A. (2017). Mitigating adversarial effects through randomization. arXiv preprint [arXiv:1711.01991](https://arxiv.org/abs/1711.01991).
- Xie, Q., Jiang, S., Jiang, L., Huang, Y., Zhao, Z., Khan, S., Dai, W., Liu, Z., & Wu, K. (2024). Efficiency optimization techniques in privacy-preserving federated learning with homomorphic encryption: A brief survey. *IEEE Internet of Things Journal*, *11*(14), 24569–24580.
- Xing, Y., Wu, Q., Cheng, D., Zhang, S., Liang, G., Wang, P., & Zhang, Y. (2023). Dual modality prompt tuning for vision-language pre-trained model. *IEEE Transactions on Multimedia*, *26*, 2056–2068.
- Xiong, Z., Cai, Z., Han, Q., Alrawais, A., & Li, W. (2020). ADGAN: Protect your location privacy in camera data of auto-driving vehicles. *IEEE Transactions on Industrial Informatics*, *17*(9), 6200–6210.
- Xu, G., Li, H., Liu, S., Yang, K., & Lin, X. (2019). VerifyNet: Secure and verifiable federated learning. *IEEE Transactions on Information Forensics and Security*, *15*, 911–926.
- Xu, H., Cai, Z., Takabi, D., & Li, W. (2021). Audio-visual autoencoding for privacy-preserving video streaming. *IEEE Internet of Things Journal*, *9*(3), 1749–1761.
- Xu, Y., Xu, G., An, Z., Nielsen, M. H., & Shen, M. (2023). Adversarial attacks and active defense on deep learning based identification of GAN power amplifiers under physical perturbation. *AEU-International Journal of Electronics and Communications*, *159*, Article 154478.
- Xue, M., Wu, Z., Zhang, Y., Wang, J., & Liu, W. (2022). AdvParams: An active DNN intellectual property protection technique via adversarial perturbation based parameter encryption. *IEEE Transactions on Emerging Topics in Computing*, *11*(3), 664–678.
- Yang, B., Li, W., Xiang, L., & Li, B. (2023a). Towards code watermarking with dual-channel transformations. arXiv preprint [arXiv:2309.00860](https://arxiv.org/abs/2309.00860).
- Yang, J., Li, Z., Zheng, F., Leonardis, A., & Song, J. (2022a). Prompting for multi-modal tracking. In *Proceedings of the 30th ACM international conference on multimedia* (pp. 3492–3500).
- Yang, P., Ci, H., Song, Y., & Shou, M. Z. (2024). Can simple averaging defeat modern watermarks? *Advances in Neural Information Processing Systems*, *37*, 56644–56673.
- Yang, X., Dong, Y., Pang, T., Su, H., Zhu, J., Chen, Y., & Xue, H. (2021a). Towards face encryption by generating adversarial identity masks. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 3897–3907).
- Yang, X., Zhang, J., Chen, K., Zhang, W., Ma, Z., Wang, F., & Yu, N. (2022b). Tracing text provenance via context-aware lexical substitution. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 11613–11621).
- Yang, Y., Liang, C., He, H., Cao, X., & Gong, N. Z. (2021b). Faceguard: Proactive deepfake detection. arXiv preprint [arXiv:2109.05673](https://arxiv.org/abs/2109.05673).
- Yang, Z., Sha, Z., Backes, M., & Zhang, Y. (2023b). From visual prompt learning to zero-shot transfer: Mapping is all you need. arXiv preprint [arXiv:2303.05266](https://arxiv.org/abs/2303.05266).
- Yang, Z., Zeng, K., Chen, K., Fang, H., Zhang, W., & Yu, N. (2024b). Gaussian shading: Provable performance-lossless image watermarking for diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 12162–12171).
- Yao, Y., Zhang, A., Zhang, Z., Liu, Z., Chua, T. S., & Sun, M. (2024). Cpt: Colorful prompt tuning for pre-trained vision-language models. *AI Open*, *5*, 30–38.
- Ye, X., Huang, H., An, J., & Wang, Y. (2023). DUAW: Data-free universal adversarial watermark against stable diffusion customization. arXiv preprint [arXiv:2308.09889](https://arxiv.org/abs/2308.09889).
- Yeh, C. Y., Chen, H. W., Tsai, S. L., & Wang, S. D. (2020). Disrupting imagetranslation-based deepfake algorithms with adversarial attacks. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision workshops* (pp. 53–62).
- Yeung, M., & Yeo, B. L. (1998). Fragile watermarking of three-dimensional objects. In *Proceedings 1998 international conference on image processing. ICIP98 (Cat. No. 98CB36269)* (vol. 2, pp. 442–446), IEEE.
- Yin, D., Gontijo Lopes, R., Shlens, J., Cubuk, E. D., & Gilmer, J. (2019). A Fourier perspective on model robustness in computer vision. In *Advances in neural information processing systems* (vol. 32).
- Yoo, I., Chang, H., Luo, X., Stava, O., Liu, C., Milanfar, P., & Yang, F. (2022). Deep 3D-to-2D watermarking: Embedding messages in 3D meshes and extracting them from 2D renderings. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10031–10040).
- Yoo, S., Kim, E., Jung, D., Lee, J., & Yoon, S. (2023). Improving visual prompt tuning for self-supervised vision transformers. In *International conference on machine learning* (pp. 40075–40092), PMLR.
- Yoshimura, M., Otsuka, J., Irie, A., & Ohashi, T. (2023). Rawgment: Noise-accounted RAW augmentation enables recognition in a wide variety of environments. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 14007–14017).
- Yu, N., Davis, L. S., & Fritz, M. (2019). Attributing fake images to GANs: Learning and analyzing GAN fingerprints. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 7556–7566).
- Yu, N., Skripniuk, V., Abdelnabi, S., & Fritz, M. (2021). Artificial fingerprinting for generative models: Rooting deepfake attribution in training data. In *Proceedings of the IEEE/CVF International conference on computer vision* (pp. 14448–14457).
- Yu, Z., Ip, H. H., & Kwok, L. (2003). A robust watermarking scheme for 3D triangular mesh models. *Pattern Recognition*, *36*(11), 2603–2614.
- Yu, Z., Cai, R., Cui, Y., Liu, A., & Chen, C. (2024). Visual prompt flexiblemodal face anti-spoofing. *IEEE Transactions on Dependable and Secure Computing*. <https://doi.org/10.1109/TDSC.2024.3520534>
- Zafeiriou, S., Tefas, A., & Pitas, I. (2005). Blind robust watermarking schemes for copyright protection of 3D mesh objects. *IEEE Transactions on Visualization and Computer Graphics*, *11*(5), 596–607.
- Zeng, Y., Zhou, M., Xue, Y., & Patel, V. M. (2023). Securing deep generative models with universal adversarial signature. arXiv preprint [arXiv:2305.16310](https://arxiv.org/abs/2305.16310).
- Zhang, C., Costa-Perez, X., & Patras, P. (2022). Adversarial attacks against deep learning-based network intrusion detection systems and defense mechanisms. *IEEE/ACM Transactions on Networking*, *30*(3), 1294–1311.
- Zhang, G., Wang, L., Su, Y., & Liu, A. A. (2024a). A training-free plug-and-play watermark framework for stable diffusion. arXiv preprint [arXiv:2404.05607](https://arxiv.org/abs/2404.05607).
- Zhang, J., Gu, Z., Jang, J., Wu, H., Stoecklin, M. P., Huang, H., & Molloy, I. (2018a). Protecting intellectual property of deep neural networks with watermarking. In *Proceedings of the 2018 on Asia conference on computer and communications security* (pp. 159–172).
- Zhang, J., Chen, D., Liao, J., Zhang, W. F. H., Hua, G., & Yu, N. (2021). Deep model intellectual property protection via deep water-

- marking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8), 4005–4020.
- Zhang, J., Dai, L., Xu, L., Ma, J., & Zhou, X. (2023). Black-box watermarking and blockchain for ip protection of voiceprint recognition model. *Electronics*, 12(17), 3697.
- Zhang, J., Wang, B., Li, L., Nakashima, Y., & Nagahara, H. (2024b). Instruct me more! random prompting for visual in-context learning. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 2597–2606).
- Zhang, L., Liu, X., Martin, A. V., Bearfield, C. X., Brun, Y., & Guan, H. (2024c). Robust image watermarking using stable diffusion. arXiv preprint [arXiv:2401.04247](https://arxiv.org/abs/2401.04247).
- Zhang, P. F., Huang, Z., & Xu, X. S. (2021b). Proactive privacy-preserving learning for retrieval. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 3369–3376).
- Zhang, T. (2022). Deepfake generation and detection, a survey. *Multimedia Tools and Applications*, 81(5), 6259–6276.
- Zhang, T., He, Z., & Lee, R. B. (2018b). Privacy-preserving machine learning through data obfuscation. arXiv preprint [arXiv:1807.01860](https://arxiv.org/abs/1807.01860).
- Zhang, W. E., Sheng, Q. Z., Alhazmi, A., & Li, C. (2020). Adversarial attacks on deep-learning models in natural language processing: A survey. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(3), 1–41.
- Zhang, X., Wang, S., Qian, Z., & Feng, G. (2010). Reference sharing mechanism for watermark self-embedding. *IEEE Transactions on Image Processing*, 20(2), 485–495.
- Zhang, X., Li, R., Yu, J., Xu, Y., Li, W., & Zhang, J. (2024d). EditGuard: Versatile image watermarking for tamper localization and copyright protection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11964–11974).
- Zhang, X., Meng, J., Li, R., Xu, Z., Zhang, Y., & Zhang, J. (2024). GSHider: Hiding messages into 3D Gaussian splatting. *Advances in Neural Information Processing Systems*, 37, 49780–49805.
- Zhang, X., Xu, Y., Li, R., Yu, J., Li, W., Xu, Z., & Zhang, J. (2024f). V2a-mark: Versatile deep visual-audio watermarking for manipulation localization and copyright protection. In *Proceedings of the 32nd ACM international conference on multimedia* (pp. 9818–9827).
- Zhang, X., Tang, Z., Xu, Z., Li, R., Xu, Y., Chen, B., Gao, F., & Zhang, J. (2025). OmniGuard: Hybrid manipulation localization via augmented versatile deep image watermarking. In *Proceedings of the computer vision and pattern recognition conference* (pp. 3008–3018).
- Zhang Y, Chen X, Jia J, Liu S, Ding K (2023b) Text-visual prompting for efficient 2D temporal video grounding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 14794–14804).
- Zhang, Y., Zhu, J., Xue, M., Zhang, X., & Cao, X. (2023). Adaptive 3d mesh steganography based on feature-preserving distortion. *IEEE Transactions on Visualization and Computer Graphics*, 30(8), 5299–5312.
- Zhang, Y., Li, H., Yao, Y., Chen, A., Zhang, S., Chen, P. Y., Wang, M., & Liu, S. (2024g). Visual prompting reimaged: The power of activation prompts. Unpublished.
- Zhao, X., Wang, Y. X., & Li, L. (2023a). Protecting language generation models via invisible watermarking. In *International Conference on Machine Learning* (pp. 42187–42199). PMLR.
- Zhao, Y., Liu, B., Ding, M., Liu, B., Zhu, T., & Yu, X. (2023b). Proactive deepfake defence via identity watermarking. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 4602–4611).
- Zhao, Y., Pang, T., Du, C., Yang, X., Cheung, N. M., & Lin, M. (2023c). A recipe for watermarking diffusion models. arXiv preprint [arXiv:2303.10137](https://arxiv.org/abs/2303.10137).
- Zhao, Z., & Patras, I. (2023). Prompting visual-language models for dynamic facial expression recognition. arXiv preprint [arXiv:2308.13382](https://arxiv.org/abs/2308.13382).
- Zhao, Z., Duan, J., Xu, K., Wang, C., Zhang, R., Du, Z., Guo, Q., & Hu, X. (2024). Can protective perturbation safeguard personal data from being exploited by stable diffusion? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 24398–24407).
- Zhong, Y., & Deng, W. (2020). Towards transferable adversarial attack against deep face recognition. *IEEE Transactions on Information Forensics and Security*, 16, 1452–1466.
- Zhu, C., Tang, J., Galjaard, J. M., Chen, P. Y., Birke, R., Bos, C., Chen, L. Y., et al. (2025). Tabwak: A watermark for tabular diffusion models. In *International conference on learning representations, OpenReview.net* (pp. 1–28).
- Zhu, J., Kaplan, R., Johnson, J., & Fei-Fei, L. (2018). Hidden: Hiding data with deep networks. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 657–672).
- Zhu, J., Zhang, Y., Zhang, X., & Cao, X. (2021). Gaussian model for 3D mesh steganography. *IEEE Signal Processing Letters*, 28, 1729–1733.
- Zhu, J., Lai, S., Chen, X., Wang, D., & Lu, H. (2023a). Visual prompt multimodal tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9516–9526).
- Zhu, X., Ye, G., Dong, C., Luo, X., & Wei, X. (2023b). Towards function space mesh watermarking: Protecting the copyright of signed distance fields. arXiv preprint [arXiv:2311.12059](https://arxiv.org/abs/2311.12059).
- Zhu, X., Ye, G., Luo, X., & Wei, X. (2024). Rethinking mesh watermark: Towards highly robust and adaptable deep 3D mesh watermarking. In *Proceedings of the AAAI Conference on Artificial Intelligence* (pp. 7784–7792).
- Zou, Z., Gong, B., & Wang, L. (2025). Attention to neural plagiarism: Diffusion models can plagiarize your copyrighted images! In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 19546–19556).

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.