NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Adversarial Attacks for Good: A Survey of Proactive Protection across the Visual Content Lifecycle

Jiaming Zhang[1], **Boyang Chen**[1], **Zherui Li**[1], **Fuyao Zhang**[1], **Xinyu Yan**[1], **Hong Xi Tae**[1], **Wenwen He**[1], **Xuan Wang**[1], **Siqi Guo**[1], **Junhao Dong**[1], **Kun Wang**[1], **Hanxun Huang**[2], **Yige Li**[3], **Xingjun Ma**[4], **Yang Cao**[5], **Lingjuan Lyu**[6], and **Wei Yang Bryan Lim**[1,‡]

[1]College of Computing and Data Science, Nanyang Technological University, Singapore
[2]School of Computing and Information Systems, University of Melbourne, Australia
[3]School of Computing and Information Systems, Singapore Management University, Singapore
[4]Institute of Trustworthy Embodied AI, Fudan University, China
[5]Department of Computer Science, School of Computing, Institute of Science Tokyo, Japan
[6]Sony AI, Japan

Abstract

Once visual content enters an AI pipeline, its owner often retains little technical control over how it is used. Legal and regulatory remedies can address misuse, but many technical interventions must be applied earlier, when content is released or accessed. This survey examines the protective paradigm that has grown around this intervention point, which we call *adversarial attacks for good*. Perturbations and structured signals long studied as attacks on learned models are instead applied by data owners, creators, platforms, or auditors to disrupt unauthorized automation or support later accountability. Five research communities have arrived at this inversion largely independently, each addressing a different stage of a visual asset's lifecycle: privacy filters against unwanted recognition at sharing time, unlearnable examples against unauthorized training, generative safeguards against malicious editing or imitation, adversarial CAPTCHAs for access control against automated agents, and provenance mechanisms for post-circulation attribution. Although developed in separate venues with incompatible success criteria, many of these methods exploit persistent gaps between human perception, semantic interpretation, and machine inference, suggesting that the paradigm remains relevant as visual pipelines evolve toward multimodal models and autonomous agents. To make their claims comparable, we evaluate all five families along shared axes of transferability, adaptability, and deployment readiness. We close by consolidating cross-stage countermeasures and open problems for robust, composable, and deployable owner-side protection.

Corresponding Author: Wei Yang Bryan Lim
Key Words: Adversarial machine learning; protective adversarial examples; visual content lifecycle

Adversarial Attacks for Good: A Survey of Proactive Protection across the Visual Content Lifecycle

Jiaming Zhang, Boyang Chen, Zherui Li, Fuyao Zhang, Xinyu Yan, Hong Xi Tae, Wenwen He, Xuan Wang, Siqi Guo, Junhao Dong, Kun Wang, Hanxun Huang, Yige Li, Xingjun Ma, Yang Cao, Lingjuan Lyu, and Wei Yang Bryan Lim

Abstract—Once visual content enters an AI pipeline, its owner often retains little technical control over how it is used. Legal and regulatory remedies can address misuse, but many technical interventions must be applied earlier, when content is released or accessed. This survey examines the protective paradigm that has grown around this intervention point, which we call *adversarial attacks for good*. Perturbations and structured signals long studied as attacks on learned models are instead applied by data owners, creators, platforms, or auditors to disrupt unauthorized automation or support later accountability. Five research communities have arrived at this inversion largely independently, each addressing a different stage of a visual asset's lifecycle: privacy filters against unwanted recognition at sharing time, unlearnable examples against unauthorized training, generative safeguards against malicious editing or imitation, adversarial CAPTCHAs for access control against automated agents, and provenance mechanisms for post-circulation attribution. Although developed in separate venues with incompatible success criteria, many of these methods exploit persistent gaps between human perception, semantic interpretation, and machine inference, suggesting that the paradigm remains relevant as visual pipelines evolve toward multimodal models and autonomous agents. To make their claims comparable, we evaluate all five families along shared axes of transferability, adaptability, and deployment readiness. Across the lifecycle, we find that most protections are still validated mainly against static or weakly adaptive adversaries, while evidence beyond controlled benchmarks remains scarce. We close by consolidating cross-stage countermeasures and open problems for robust, composable, and deployable owner-side protection.

Index Terms—Adversarial machine learning, protective adversarial examples

1 Introduction

The rapid proliferation of multimodal foundation models, generative pipelines, and autonomous AI agents has made the automated processing of visual content a basic feature of digital infrastructure. These systems operate at a scale and with capabilities that have produced a structural asymmetry: content creators and rights holders control their visual assets until those assets enter an AI pipeline, but have no practical authority over what the pipeline does with them afterward. The legal system has begun to register this condition as litigation. Creators and publishers have brought multiple cases, several still pending, alleging that their visual work was scraped, reproduced, or repurposed for model training without authorization, and major regulatory frameworks now require AI developers to disclose the provenance of training data.[1] These proceedings are symptoms rather than anomalies: litigation and regulation respond to misuse after it occurs, whereas the intervention points available to a content owner lie before the pipeline, not inside it.

One such intervention point takes its name from a foundational discovery in machine learning. Szegedy et al. [1] demonstrated that perturbations imperceptible to humans can cause deep neural networks to misclassify inputs with high confidence. These perturbations were named adversarial examples and subsequently studied for over a decade as a security threat, an attack surface to be measured and defended against. That framing, however, presupposes who is imposing the perturbation and why. If the party applying the signal is not an adversary seeking misclassification but a data owner seeking to prevent unauthorized automated use, the objective inverts: induced model failure becomes the protection goal rather than the attack objective. This inversion is what we call *adversarial attacks for good*, and a growing body of work has explored this direction [2], [3]. Adversarial examples expose a structural blind spot of gradient-trained models, one that is present in every

- *J. Zhang, B. Chen, X. Yan, Z. Li, F. Zhang, H. X. Tae, W. He, X. Wang, S. Guo, J. Dong, K. Wang and W. Y. B. Lim are with the College of Computing and Data Science, Nanyang Technological University, Singapore. W. Y. B. Lim is the corresponding author.*
- *H. Huang is with the School of Computing and Information Systems, University of Melbourne, Australia.*
- *Y. Li is with the School of Computing and Information Systems, Singapore Management University, Singapore.*
- *X. Ma is with the Institute of Trustworthy Embodied AI, Fudan University, China.*
- *Y. Cao is with the Department of Computer Science, School of Computing, Institute of Science Tokyo, Japan.*
- *L. Lyu is with Sony AI, Japan.*

1. Representative cases include *Andersen v. Stability AI* (N.D. Cal., filed 2023) and *Getty Images v. Stability AI* (D. Del., filed 2023); on the regulatory side, the EU AI Act (Regulation (EU) 2024/1689) requires providers of general-purpose AI models to publish training-content summaries and honor machine-readable copyright reservations.

TABLE 1
Comparison of this survey with related surveys on coverage and analytical approach (✓ = addressed; ○ = partial; × = not addressed).

Ref.	Focus	Lifecycle framing	Families (of 5)	Adaptive robustness	Deployment readiness
Broad surveys					
[2]	proactive schemes	×	3	×	×
[3]	broad adv. ML	×	3	×	○
Family-specific surveys					
[9]	facial privacy	×	1	✓	○
[5]	unlearnable examples	×	1	✓	○
[10]	deepfake defense	×	2	○	×
[6]	AIGC defense	×	2	×	×
[7]	CAPTCHA design	×	1	○	×
[8]	watermarking	×	1	✓	○
	This survey	✓	**5**	✓	✓

such model, from early image classifiers to the multimodal systems and autonomous agents now at the center of digital infrastructure, and that has not been eliminated by scale or architectural change. At its core, adversarial attacks for good is an argument about paradigm: as long as AI pipelines rely on gradient-trained models, this blind spot persists, and the protective approach extends to successive systems as readily as the underlying vulnerability does.

Protection is applied before the asset enters the pipeline, but it must take effect inside the pipeline, at whichever stage the anticipated misuse occurs. The specific form the inversion takes therefore depends on the stage being defended. We identify five lifecycle stages, each pairing a distinct misuse risk with a corresponding mechanism family: adversarial privacy filters against unwanted recognition at sharing time, unlearnable examples against unauthorized training at release time, generative safeguards against manipulation and imitation by generative models, adversarial CAPTCHAs against automated abuse of online services, and provenance mechanisms for attribution after circulation. The survey is organized accordingly (Fig. 1). Our scope is defined by two conditions: the protected asset is visual, and the protection mechanism is adversarial, operating through the structured-signal sensitivity of learned models. Differential privacy, federated learning, cryptographic access control, and research that treats adversarial perturbations as threats rather than protective tools fall outside this boundary.

Despite operating on the same mathematical property, the five communities that independently developed protective uses of adversarial mechanisms have evolved largely in isolation, publishing in separate venues, adopting incompatible vocabularies, and applying incommensurable threat models and success criteria [4], [5], [6], [7], [8]. No existing survey places these families side by side (Table 1). This incommensurability is precisely what demands a common evaluation vocabulary; we provide one through three shared axes, transferability (L_1), adaptability (L_2), and deployment readiness (L_3), defined formally in Section 2.

This survey makes three contributions.

1) **A durable protective paradigm.** We introduce adversarial attacks for good as a unified paradigm. Its operative principle, the perceptual gap between human observers and learned models, is a structural property of AI systems rather than an artifact of any particular architecture or training method.
2) **A lifecycle taxonomy.** We provide, to our knowledge, the first unified account of adversarial privacy filters, unlearnable examples, generative safeguards, adversarial CAPTCHAs, and provenance mechanisms as successive stages of a single protective lifecycle.
3) **A shared evaluation framework.** We compare all five families along three axes: transferability (L_1), adaptability (L_2), and deployment readiness (L_3), making threat models and robustness claims across the five communities directly commensurable.

The remainder of this survey is organized as follows. Section 2 formalizes protective adversarial transformations, and defines the comparison axes L_1–L_3. Sections 3 through 7 review the five lifecycle stages in order. Section 8 consolidates cross-stage countermeasures and open problems and concludes the survey.

2 A Unified Framework for Adversarial Protection

This section establishes the two devices used throughout the survey: the template behind all protective adversarial mechanisms (Section 2.1), and the axes L_1–L_3 along which Sections 3–7 compare them (Section 2.2). No familiarity with adversarial machine learning is assumed. Fig. 3 shows the distribution of surveyed papers by family and year.

2.1 Protective Adversarial Transformations

Szegedy et al. [1] discovered that deep networks can be steered by signals people cannot see: for almost any input x, there is a perturbation δ so faint that $x + \delta$ looks identical to x, yet the model's output flips, $f(x + \delta) \neq f(x)$. Two facts about these *adversarial examples* matter here. They are structural, not incidental: gradient-trained models rely on faint input statistics that human perception discards, so a decade of architectural change has not removed them [3], [2]. And they *transfer*: a perturbation computed on one model often deceives another trained independently [193], so the effect does not require access to its eventual target.

Attack research treats these properties as a threat; protection reverses the roles, and the reversal is easiest to see by following a single portrait photo after it is posted online (Fig. 1). A recognition service may index the face and link it to an identity. A scraper may fold the photo into a training corpus. A generative model may re-synthesize the person in scenes that never happened. The bots doing the scraping must first pass the hosting platform's human-verification challenges. And once copies circulate, the owner may need to prove where the image came from or how it was used. Sections 3–7 survey the protections built for these five moments, and they all share one move: embedding an adversarial signal in the asset *before* release.

Throughout, we call the asset owner the *protector* and the operator of the unwanted pipeline the *adversary*, inverting the classical usage. The protector controls the asset only until release; the adversary controls everything afterward, including manipulations $g \in \mathcal{G}$ (re-encoding,

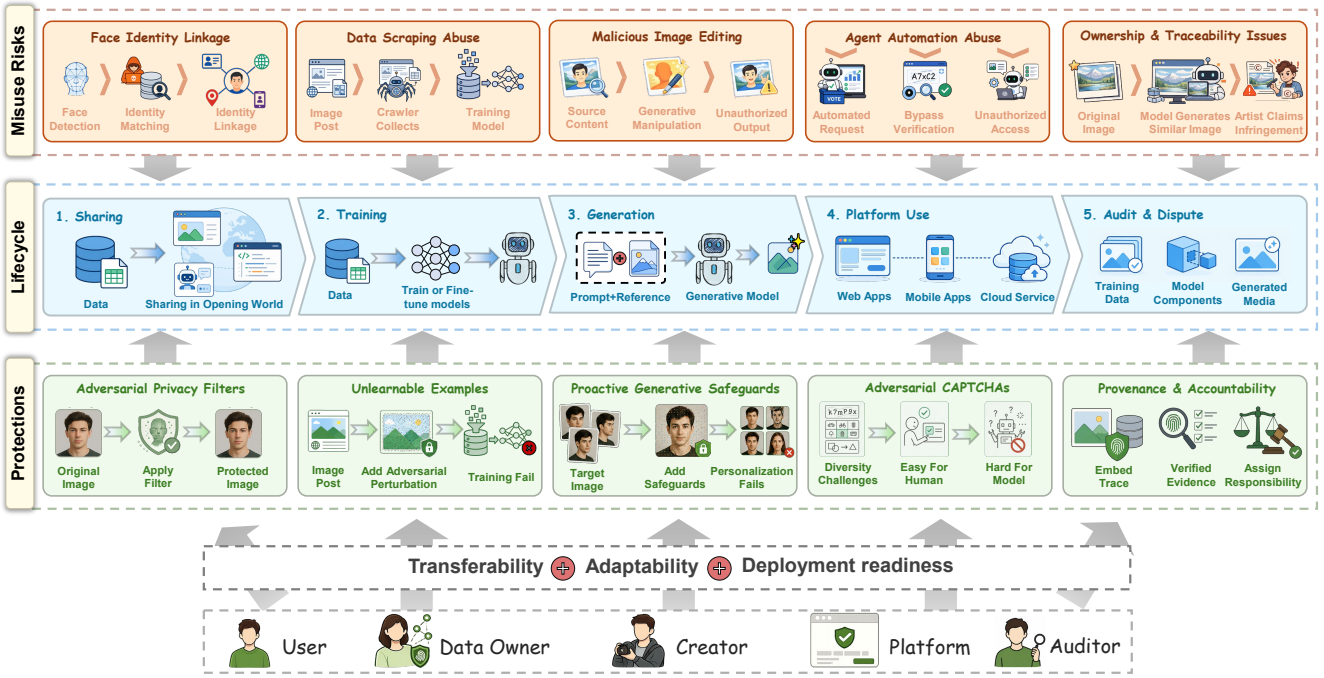

Fig. 1. Overview of misuse risks, lifecycle stages, and corresponding protection mechanisms, assessed along three axes: transferability (L_1), adaptability (L_2), and deployment readiness (L_3).

restoration, deliberate countermeasures) applied before the pipeline runs. The protector therefore applies a transformation T before release,

$$\tilde{x} = T(x), \qquad \underbrace{d(x, \tilde{x}) \leq \epsilon}_{\text{utility}}, \qquad \underbrace{F(g(\tilde{x})) \neq y}_{\text{protection}}, \tag{1}$$

where $F \in \mathcal{F}$ is the unwanted pipeline and y the output it is built to obtain: the utility condition keeps the asset valuable to its human audience, under a small budget ϵ, while the protection condition makes the pipeline fail regardless. Each family reads x, F, d, and failure in its own terms; each section's opening states its own reading. The one deliberate departure is provenance (Section 7): applied when misuse can no longer be prevented, it replaces failure with verification, $V(g(\tilde{x})) = 1$ for a designated verifier V, turning the signal into evidence rather than sabotage.

2.2 A Common Evaluation Lens: L_1–L_3

The five communities report robustness in mutually incompatible vocabularies; the axes make their claims commensurable, recording what a method assumes about the pipeline $\mathcal{F}$ (L_1), which manipulations $\mathcal{G}$ it has been validated under (L_2), and how mature its evidence is (L_3).

L_1**: transferability** grades the access under which protection is built and verified: *white-box* (the deployed target itself), *gray-box* (shared components such as a public backbone), or *black-box* (surrogates only, so that surrogate-to-target transfer carries the entire claim). The polarity is inverted relative to attacks: there, white-box is the informative worst case; here, black-box is the demanding level, because protectors can rarely inspect the service that will process their asset. A face cloak, for instance, must defeat the commercial API actually deployed, not the surrogate it was optimized against. In the accountability stage, L_1 instead records the verification access available to the party proving misuse.

L_2**: adaptability** grades the manipulations g a method has been shown to survive: *static* validation against a fixed pipeline, *routine* media operations applied without knowledge of the protection (compression, resizing, re-encoding), or informed *adaptive* countermeasures that target it (purification, adapted training, signal removal or forgery). Because the protector commits the signal at release and cannot re-optimize it afterward, robustness against uninformed adversaries says little about informed ones; this is where protection claims most often collapse.

L_3**: deployment readiness** grades the maturity of the evidence: *laboratory* demonstrations on in-house models, *external* validation against systems the authors do not control (commercial APIs, online services, human users), or sustained *operational* use. The axis separates what is feasible from what is genuinely available.

Sections 3 through 7 instantiate these axes under their own threat models in *domain-specific vocabulary*.

3 Adversarial Privacy Filters

Adversarial privacy filters are proactive mechanisms that protect facial identity before images enter recognition pipelines, injecting either imperceptible pixel-level noise or semantically meaningful appearance changes to mislead automated inference while preserving human-perceived quality. As facial recognition (FR) systems pervade social platforms, surveillance infrastructure, and increasingly multimodal large language models (MLLMs) and visual agents, we compare these filters along the survey's three axes: **transferability** (L_1, generalizing from surrogates to inaccessible

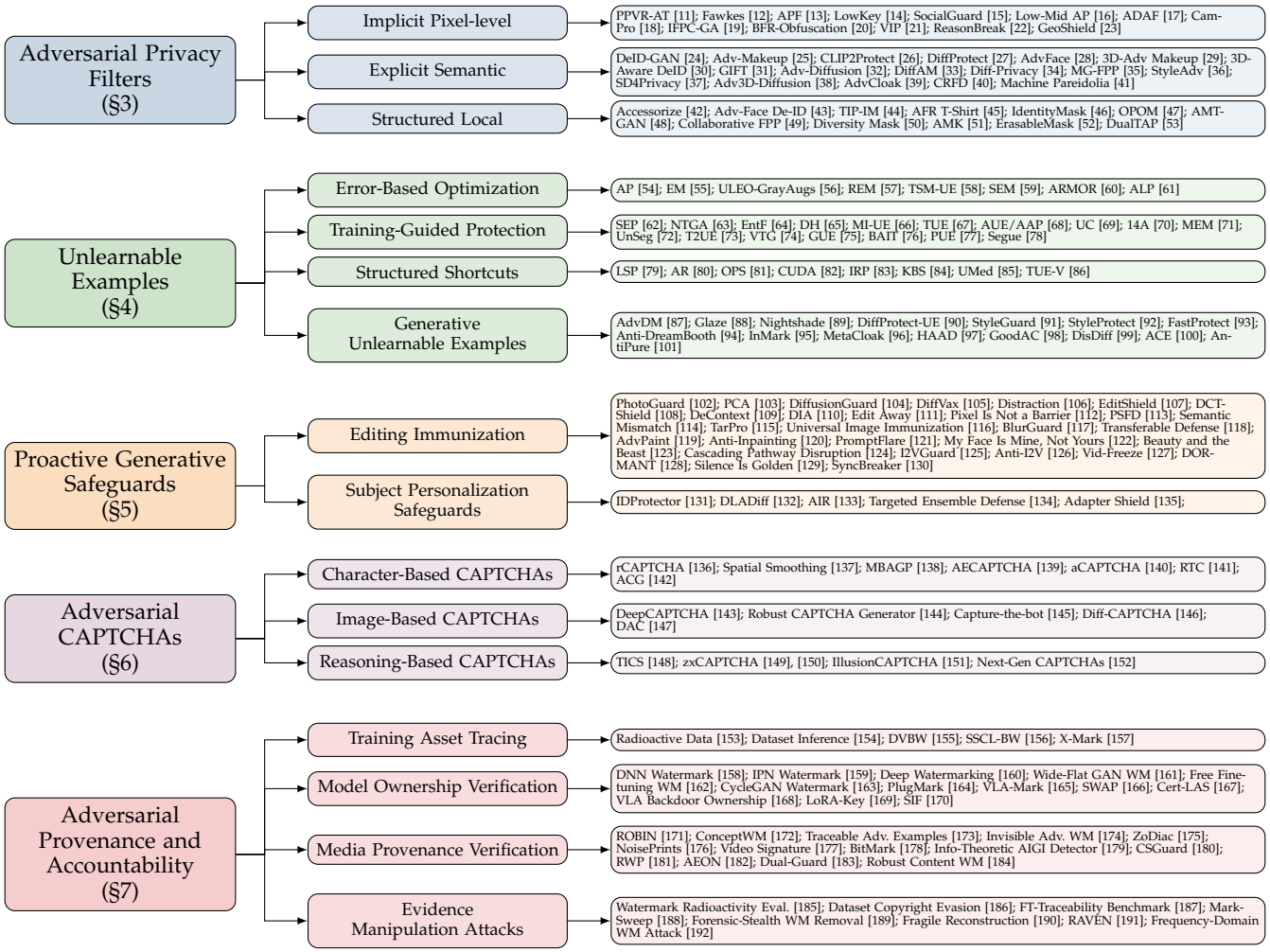

Fig. 2. Three-level taxonomy of protective adversarial mechanisms, organized by family, subcategory, and representative works.

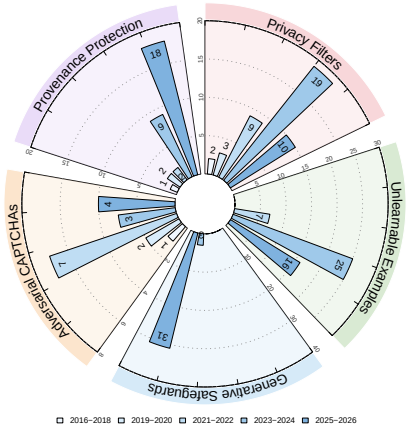

Fig. 3. Publication counts of surveyed papers by family and time period.

deployed models), **adaptability** (L_2, surviving content variation, platform processing, and supporting reversibility), and **deployment readiness** (L_3, maturing from client-side pre-upload toward source- and platform-side operation).

We organize existing methods by perturbation type, which most directly determines the threat model, visual fidelity, and deployment pathway. As summarized in Table 2, this section reviews three categories: (1) *Implicit Pixel-Level Perturbation*, hiding adversarial signals in imperceptible noise; (2) *Explicit Semantic Perturbation*, protecting identity via natural appearance transformations; (3) *Structured Local Perturbation*, confining modifications to facial or wearable regions.

3.1 Implicit Pixel-Level Perturbation

Implicit pixel-level protection develops as one trajectory: it first secures black-box transferability so a cloak works against recognizers the user cannot access, then hardens that cloak to survive the processing pipelines of real platforms, and finally moves protection earlier in the capture pipeline and outward to new inference targets.

Establishing black-box transferability. The defining scenario is a user who wants to share a photo but cannot query the commercial recognizer that will index it, so protection needF transfer from a surrogate to an unseen deployed model. Early work assumed white-box access—*PPVR-AT* [11] optimizes directly against the target. However, this assumption fails when the recognizer is proprietary. *Fawkes* [12] and *LowKey* [14] remove it by nudging fea-

TABLE 2
Taxonomy of adversarial privacy filter methods for privacy protection.

Paper	L_1-Transferability	L_2-Adaptability	L_3-Deployment Readiness	Target Model
Implicit Pixel-level				
PPVR-AT[ECCV'18] [11]	White-box	Non-Interactive	Client-side Pre-upload	Traditional FR
Fawkes[USENIX'20] [12]	Black-box	Non-Interactive	Client-side Pre-upload	Traditional FR
APF[ACM MM'20] [13]	Black-box	Non-Interactive	Client-side Pre-upload	Traditional FR
LowKey[ICLR'21] [14]	Black-box	Non-Interactive	Client-side Pre-upload	Traditional FR
SocialGuard[JISA'21] [15]	Black-box	Non-Interactive	Client-side Pre-upload	Traditional FR
Low-Mid AP[IS'23] [16]	Black-box	Non-Interactive	Client-side Pre-upload	Traditional FR
ADAF[Arxiv'23] [17]	Black-box	Content-Adaptive	Client-side Pre-upload	Traditional FR
CamPro[NDSS'24] [18]	Black-box	Non-Interactive	Physical Device-side	Traditional FR
IFPC-GA[ICASSP'24] [19]	Black-box	Non-Interactive	Client-side Pre-upload	Traditional FR
BFR-Obfuscation[ICML'24] [20]	Black-box	Non-Interactive	Client-side Pre-upload	Traditional FR
VIP[Arxiv'25] [21]	White-box	Content-Adaptive	Client-side Pre-upload	MLLM
ReasonBreak[ICLR'26] [22]	Black-box	Content-Adaptive	Client-side Pre-upload	MLLM
GeoShield[AAAI'26] [23]	Black-box	Content-Adaptive	Client-side Pre-upload	MLLM
Explicit Semantic				
DeID-GAN[ACM MM'21] [24]	Black-box	Non-Interactive	Client-side Pre-upload	Traditional FR
Adv-Makeup[IJCAI'21] [25]	Black-box	Non-Interactive	Client-side Pre-upload	Traditional FR
CLIP2Protect[CVPR'23] [26]	Black-box	Non-Interactive	Client-side Pre-upload	Traditional FR
DiffProtect[PR'26] [27]	Black-box	Non-Interactive	Client-side Pre-upload	Traditional FR
AdvFace[CVPR'23] [28]	White-box	Non-Interactive	Platform / Cloud-side	Traditional FR
3D-Adv Makeup[TPAMI'23] [29]	Black-box	Non-Interactive	Client-side Pre-upload	Traditional FR
3D-Aware DeID[MM Asia'23] [30]	Black-box	Non-Interactive	Client-side Pre-upload	Traditional FR
GIFT[ACM MM'24] [31]	Black-box	Non-Interactive	Client-side Pre-upload	Traditional FR
Adv-Diffusion[AAAI'24] [32]	Black-box	Non-Interactive	Client-side Pre-upload	Traditional FR
DiffAM[CVPR'24] [33]	Black-box	Non-Interactive	Client-side Pre-upload	Traditional FR
Diff-Privacy[TCSVT'24] [34]	Black-box	Reversible	Client-side Pre-upload	Traditional FR
MG-FPP[ECCV'24] [35]	Black-box	Non-Interactive	Client-side Pre-upload	Traditional FR
StyleAdv[PoPETs'24] [36]	Black-box	Reversible	Client-side Pre-upload	Traditional FR
SD4Privacy[ICME'24] [37]	Black-box	Non-Interactive	Client-side Pre-upload	Traditional FR
Adv3D-Diffusion[ICIA'25] [38]	Black-box	Non-Interactive	Client-side Pre-upload	Traditional FR
AdvCloak[PR'25] [39]	Black-box	Non-Interactive	Client-side Pre-upload	Traditional FR
CRFD[ESWA'25] [40]	Black-box	Reversible	Client-side Pre-upload	Traditional FR
Machine Pareidolia[AAAI'26] [41]	Black-box	Non-Interactive	Client-side Pre-upload	Traditional FR
Structured Local				
Accessorize[CCS'16] [42]	White-box	Non-Interactive	Physical Device-side	Traditional FR
Adv-Face De-ID[ICIP'19] [43]	White-box	Non-Interactive	Client-side Pre-upload	Traditional FR
TIP-IM[ICCV'21] [44]	Black-box	Non-Interactive	Client-side Pre-upload	Traditional FR
AFR T-Shirt[ICCCI'21] [45]	Black-box	Non-Interactive	Physical Device-side	Traditional FR
IdentityMask[TCSVT'22] [46]	Black-box	Reversible	Client-side Pre-upload	Traditional FR
OPOM[TPAMI'22] [47]	Black-box	Non-Interactive	Client-side Pre-upload	Traditional FR
AMT-GAN[CVPR'22] [48]	Black-box	Non-Interactive	Client-side Pre-upload	Traditional FR
Collaborative FPP[ICIC'23] [49]	Black-box	Non-Interactive	Platform / Cloud-side	Traditional FR
Diversity Mask[PETS'24] [50]	Black-box	Content-Adaptive	Client-side Pre-upload	Traditional FR
AMK[ACM MM'25] [51]	White-box	Reversible	Client-side Pre-upload	MLLM
ErasableMask[TMM'25] [52]	Black-box	Reversible	Client-side Pre-upload	Traditional FR
DualTAP[ECCV'26] [53]	Black-box	Content-Adaptive	Platform / Cloud-side	Visual Agent

ture embeddings toward decoy identities through surrogate ensembles, and this transferable formulation is what lets *SocialGuard* [15] succeed against live commercial recognition APIs rather than offline benchmarks.

Surviving real-world platform processing. Transferability alone is insufficient once the image is uploaded, because the platform itself re-encodes and restores it before any recognizer runs; the next question is therefore whether the perturbation survives this pipeline. Since JPEG recompression and blind face restoration erase raw pixel noise, *APF* [13], *Low-Mid AP* [16], and *BFR-Obfuscation* [20] relocate the signal into resilient frequency bands. *ADAF* [17] and *IFPC-GA* [19] complement this with content-adaptive noise budgets that concentrate perturbation where it survives while keeping the shared image visually clean.

Extending to source-level and multimodal deployment. With a cloak that both transfers and survives, the remaining frontier is where protection is applied and what it protects against. *CamPro* [18] moves protection off the client and into the camera sensor, obfuscating identity before an image is ever stored. In parallel, the target has shifted from traditional face recognition to multimodal models: *VIP* [21], *ReasonBreak* [22], and *GeoShield* [23] disrupt the visual encoders and reasoning chains of MLLMs, extending the pre-upload filter from blocking identity matching to blocking higher-level inference such as geolocation.

3.2 Explicit Semantic Perturbation

Explicit semantic protection replaces imperceptible noise with natural transformations like digital makeup [25], [24]. Its **transferability** relies on *black-box* assumptions, turning protection into a stylistic choice rather than a suspicious artifact. Its **adaptability** leverages diffusion models for user-friendly editing and reversibility for authorized recovery [33], [27], [36], [40]. For **deployment**, 3D-aware variants extend these tools to physical surveillance across varying viewpoints [29], [38].

Establishing protection through natural appearance. The motivating scenario is a user whose protection need not look like an attack, so that it survives the noise-stripping filters platforms apply and raises no suspicion. *Adv-Makeup* [25] realizes this with transferable cosmetic overlays, while *DeID-GAN* [24] and *3D-Aware DeID* [30] anonymize identity while preserving expression. Diffusion models then make these edits controllable rather than fixed: *DiffAM* [33] and *DiffProtect* [27] synthesize the protective appearance end-to-end, and *CLIP2Protect* [26], [35] lets users specify it by text prompt [32], while *GIFT* [31] and *AdvCloak* [39] further improve black-box semantic protection through transferable feature-level or cloak-style optimization. Control that in turn allows *AdvFace* [28] to explore platform-side protection, while *SD4Privacy* [37] applies generative editing in a pre-upload privacy-preserving setting.

Supporting reversibility for authorized recovery. Because a semantic edit alters the visible image rather than adding removable noise, the natural next question is whether a trusted party can recover the original—an adaptability property pixel-level cloaks rarely offer. *StyleAdv* [36] restores the protected face for authorized users holding the correct cryptographic key, and disentanglement-based *CRFD* [40] and *Diff-Privacy* [34] suppress identity while preserving expression and lighting, so recovered photos remain usable for downstream tasks such as video conferencing.

Extending to physical, cross-view surveillance. The final stage carries semantic protection out of the well-lit single image and into physical surveillance, where the protector controls neither viewpoint nor lighting. *3D-Adv Makeup* [29] binds adversarial cosmetics to the facial surface so evasion holds across camera angles, and *Adv3D-Diffusion* [38] adds generative refinement to remove residual video artifacts. *Machine Pareidolia* [41] pushes this furthest, letting users actively shape how a security camera perceives them rather than merely evading detection.

3.3 Structured Local Perturbation

Structured local protection confines adversarial signals to bounded semantic regions [42]. Its **transferability** relies on *black-box*, person-specific patterns in identity-critical areas [44], [47]. Its **adaptability** extends beyond plain cloaking to support reversibility, diversity, and multi-user coordination [52], [46], [50]. The key distinction is **deployment**, as spatial locality maps easily onto physical wearables, platform filters, and visual agent protection [45], [53].

Establishing protection through localized regions. The founding scenario is physically realizable evasion: *Accessorize* [42] shows that a printed pair of glasses alone can fool a recognizer, establishing that a signal confined to a small semantic region suffices to control identity. *Adv-Face De-ID* [43] and *AMT-GAN* [48] bring this localization into image space, and *TIP-IM* [44] and *OPOM* [47] make it transferable by learning person-specific patterns concentrated on identity-critical zones rather than a single model-tuned perturbation.

Supporting recovery and multi-user coordination. Once localized protection transfers, deploying it across a user community raises two adaptability demands the single-image view ignores: authorized recovery, and avoiding collapse when many users protect against the same service. *ErasableMask* [52] and *IdentityMask* [46] embed a recoverable signal inside the local patch, *Diversity Mask* [50] regularizes users away from converging on a shared decoy identity, and *Collaborative FPP* [49] coordinates perturbations server-side to reduce the community's aggregate identity leakage.

Extending to physical and agentic deployment. Finally, because the signal is spatially confined, it transfers naturally to deployment settings beyond the pre-upload image. *AFR T-Shirt* [45] scales the wearable from glasses to full-body clothing against video re-identification, while the target expands to automated visual agents: *AMK* [51] places reversible masks at patch-grid boundaries to exploit MLLM tokenization, and *DualTAP* [53] moves protection platform-side, intercepting image streams to suppress sensitive-attribute inference in tool-augmented pipelines.

3.4 Discussion

Countermeasures. Protection strength here is conditional on the adversary's post-release processing, and the manipulations that break these filters cluster around a single capability: restoring a near-clean image before the recognizer runs. JPEG recompression, blind face restoration, and diffusion-based purification attenuate pixel- and frequency-space cloaks, which is precisely why later pixel-level methods migrate the signal into resilient frequency bands and adaptive budgets [13], [20]. Explicit semantic and structured local perturbations resist naive noise removal because their signal is carried by natural appearance or a bounded region rather than additive noise, but they remain exposed to a complementary attack surface, switching the deployed recognizer backbone or detecting and filtering the protective pattern once it is characterized. Reversibility widens this surface further: a recoverable signal that a trusted party can invert is also a signal an adversary can target, so key management and trust become part of the threat model rather than an implementation detail. Across the three categories, then, no single design dominates; each trades protection, recoverability, fidelity, and cost against a different slice of the adversary's capabilities.

Open problems and future directions. The key challenge for adversarial privacy filters is moving beyond fixed face-recognition settings toward more adaptive and compositional visual inference pipelines. First, evaluation remains insufficiently standardized. Existing methods are often tested under different models, datasets, and preprocessing settings, making robustness claims hard to compare. Future benchmarks should explicitly include adaptive restoration and recognizer-switching attacks and measure residual privacy leakage rather than only recognition failure. Second, the protection objective should expand beyond identity matching. Modern VLMs can infer sensitive attributes, locations, activities, and social context even when face recognition is disrupted. Privacy filters therefore need finer leakage metrics and more selective protection mechanisms that suppress sensitive evidence without destroying benign visual utility. Third, visual agents introduce a stronger threat model. Unlike passive recognizers, agents can combine visual cues with instructions, memory, tools, and external search, turning weak residual evidence into private conclusions or actions. Future work should evaluate privacy filters by downstream agent behavior, asking whether they can serve as trust-aware privacy layers rather than single-image cloaks.

4 Unlearnable Examples

Unlearnable examples (UE), also studied under unlearnable data, availability protection, or data availability attacks, aim to restrict unauthorized use of released data for model training while preserving human-perceived or authorized utility. Whereas adversarial privacy filters protect facial identity before images enter recognition or identity-linkage pipelines, UE addresses the subsequent unauthorized-training stage: released images may still be scraped into downstream training pipelines, so protection shifts from blocking recognition to limiting what models can learn from the data. In discriminative settings, protected data should cause trained models to generalize poorly to clean test data; in generative settings, protection appears as degraded generation quality, failed personalization, or misdirected fine-tuning.

Table 3 summarizes the surveyed techniques by **transferability**, **adaptability**, **deployment readiness**, and target model families.

The development of UE can be organized into four stages. (1) *Error-based optimization* establishes the basic data-

TABLE 3
Taxonomy of unlearnable example methods for unauthorized training prevention.

Paper	L_1-Transferability	L_2-Adaptability	L_3-Deployment Readiness	Target Model
Error-Based Optimization				
AP[NeurIPS'21] [54]	Black-box	Transformation Resistant	Application Scenario	Discriminative
EM[ICLR'21] [55]	Black-box	Non-Adaptive	Application Scenario	Discriminative
ULEO-GrayAugs[arXiv'21] [56]	Black-box	Transformation Resistant	No Deployment Evidence	Discriminative
REM[ICLR'22] [57]	Black-box	Training-Pipeline Resistant	No Deployment Evidence	Discriminative
TSM-UE[CVPR'24] [58]	Black-box	Training-Pipeline Resistant	No Deployment Evidence	Discriminative
SEM[AAAI'24] [59]	Black-box	Training-Pipeline Resistant	Application Scenario	Discriminative
ARMOR[TPAMI'26] [60]	Black-box	Transformation Resistant	Application Scenario	Discriminative
ALP[arXiv'23] [61]	White-box	Non-Adaptive	Application Scenario	Discriminative
Training-Guided Protection				
SEP[ICLR'23] [62]	Black-box	Non-Adaptive	No Deployment Evidence	Discriminative
NTGA[ICML'21] [63]	Black-box	Non-Adaptive	No Deployment Evidence	Discriminative
EntF[ICLR'23] [64]	Black-box	Training-Pipeline Resistant	No Deployment Evidence	Discriminative
DH[TIFS'24] [65]	Black-box	Transformation Resistant	No Deployment Evidence	Discriminative
MI-UE[ICLR'26] [66]	Black-box	Non-Adaptive	No Deployment Evidence	Discriminative
TUE[ICLR'23] [67]	Black-box	Training-Pipeline Resistant	No Deployment Evidence	Discriminative
AUE/AAP[NeurIPS'24] [68]	Black-box	Training-Pipeline Resistant	No Deployment Evidence	Discriminative
UC[CVPR'23] [69]	Black-box	Non-Adaptive	External Evaluation	Discriminative
14A[ICML'24] [70]	Black-box	Non-Adaptive	Application Scenario	Discriminative
MEM[ACM MM'24] [71]	Black-box	Non-Adaptive	Application Scenario	Foundational
UnSeg[NeurIPS'24] [72]	Black-box	Non-Adaptive	Application Scenario	Discriminative
T2UE[ACM MM'25] [73]	Black-box	Non-Adaptive	Workflow Design	Foundational
VTG[NeurIPS'25] [74]	Black-box	Non-Adaptive	Application Scenario	Discriminative
GUE[AAAI'24] [75]	Black-box	Training-Pipeline Resistant	No Deployment Evidence	Discriminative
BAIT[ICLR'26] [76]	Black-box	Training-Pipeline Resistant	Application Scenario	Foundational
PUE[NDSS'25] [77]	White-box	Training-Pipeline Resistant	No Deployment Evidence	Discriminative
Segue[ICASSP'25] [78]	Black-box	Training-Pipeline Resistant	Application Scenario	Discriminative
Structured Shortcuts				
LSP[KDD'22] [79]	Black-box	Transformation Resistant	Application Scenario	Discriminative
AR[NeurIPS'22] [80]	Black-box	Transformation Resistant	No Deployment Evidence	Discriminative
OPS[ICLR'23] [81]	Black-box	Non-Adaptive	No Deployment Evidence	Discriminative
CUDA[CVPR'23] [82]	Black-box	Training-Pipeline Resistant	No Deployment Evidence	Discriminative
IRP[ECCV'24] [83]	Black-box	Purification Resistant	No Deployment Evidence	Discriminative
KBS[ACM MM'25] [84]	Black-box	Transformation Resistant	No Deployment Evidence	Discriminative
UMed[arXiv'24] [85]	Black-box	Non-Adaptive	Application Scenario	Discriminative
TUE-V[ICCV'25] [86]	Black-box	Non-Adaptive	Application Scenario	Discriminative
Generative Unlearnable Examples				
AdvDM[ICML'23] [87]	Black-box	Non-Adaptive	Application Scenario	Generative
Glaze[USENIX Security'23] [88]	Black-box	Non-Adaptive	External Evaluation	Generative
Nightshade[IEEE S&P'24] [89]	Black-box	Non-Adaptive	Application Scenario	Generative
DiffProtect-UE[ICLR'24] [90]	Black-box	Non-Adaptive	Application Scenario	Generative
StyleGuard[NeurIPS'25] [91]	Black-box	Purification Resistant	Application Scenario	Generative
StyleProtect[CVPR'26] [92]	Black-box	Non-Adaptive	Application Scenario	Generative
FastProtect[CVPR'25] [93]	Black-box	Non-Adaptive	Application Scenario	Generative
Anti-DreamBooth[ICCV'23] [94]	Black-box	Non-Adaptive	Application Scenario	Generative
InMark[CVPR'24] [95]	Black-box	Transformation Resistant	Application Scenario	Generative
MetaCloak[CVPR'24] [96]	Black-box	Transformation Resistant	External Evaluation	Generative
HAAD[ACM MM'25] [97]	Black-box	Non-Adaptive	Application Scenario	Generative
GoodAC[CVPR'25] [98]	Black-box	Non-Adaptive	Application Scenario	Generative
DisDiff[ACM MM'24] [99]	Black-box	Non-Adaptive	Application Scenario	Generative
ACE[ICLR'25] [100]	Black-box	Non-Adaptive	Application Scenario	Generative
AntiPure[ICCV'25] [101]	Black-box	Purification Resistant	Application Scenario	Generative

side perturbation paradigm. (2) *Training-guided protection* incorporates training dynamics, representations, objectives, and model priors into UE generation. (3) *Structured shortcuts* constructs non-semantic signals that models learn in place of task-relevant visual features. (4) *Generative UEs* extend protection to text-to-image fine-tuning, personalization, customization, and style mimicry.

4.1 Error-Based Optimization

Error-based optimization grows out of availability poisoning and reframes it as personal or sensitive data protection. Its application scenario is data release before downstream training: images are modified so that later unauthorized training becomes unreliable. Its **transferability** is usually evaluated through black-box transfer to unseen downstream models, its **adaptability** pressure mainly comes from transformation resistance against preprocessing or augmentation and training-pipeline resistance against adversarial training [56], [60], [57], [58], [59], and its **deployment readiness** develops through extensions from personal-image protection toward sensitive-domain application scenarios such as biomedical-image protection [55], [61].

From poisoning to personal-data protection. *AP* [54] establishes the availability-attack starting point by showing that adversarial examples can become strong training-time poisons. *EM* [55] turns this idea toward personal image protection, where imperceptible error-minimizing noise prevents released photos from supporting unauthorized model training. The key shift is from attacking a learner to protecting a data asset before it enters the training pipeline.

Robustness pressure. Once personal-data protection is established, the main question is whether the protection survives realistic training operations. *ULEO-GrayAugs* [56] shows that early UE can be weakened by grayscale filtering and augmentation, motivating more robust error-based protection. Several error-based methods instead focus on adversarial-training robustness [57], [58], [59], whereas *ARMOR* [60] emphasizes robustness to data augmentation. These works move the setting from clean benchmark training toward practical preprocessing and training pipelines.

Practical-domain extension. *ALP* [61] extends input-optimization UE to biomedical images, where unauthorized training is tied to data governance, institutional reuse, and sensitive-domain release. Rather than simply adding another dataset, this application evaluates error-based UE in a sensitive-domain release scenario where unauthorized training raises privacy, compliance, and data-governance concerns.

4.2 Training-Guided Protection

Training-guided protection shifts UE from directly optimizing input loss to modeling what unauthorized training will learn. Its **transferability** is mainly evaluated through black-box transfer across downstream models, while later work broadens the target settings beyond standard clas-

sification and architecture-level transfer, including label-agnostic use and zero-contact protection workflows [67], [69], [73]. Its **adaptability** is mainly captured by training-pipeline resistance, including learning-paradigm changes, adversarial training, pretrained backbones, and parameter recovery [67], [68], [75], [76], [77]. Its **deployment readiness** develops through label-agnostic evaluation, zero-contact protection workflows, and pipeline-aware face-image release scenarios [69], [73], [78].

Training guidance across representations, settings, and objectives. Early training-guided methods use the training process itself as a protection signal: *SEP* [62] uses checkpoint ensembles, and *NTGA* [63] uses generalization approximations to guide protected-data generation. Representation-oriented work then protects data by limiting the reusable features that unauthorized models can extract from protected examples [64], [65], [66]. The target settings also broaden from standard classification to cross-paradigm learning [67], [68], label-agnostic training [69], concept-level training [70], multimodal contrastive learning [71], segmentation [72], and cross-domain or cross-task transfer [74].

Adapting to stronger training pipelines. As downstream training pipelines become stronger, UE must address training choices that can weaken protection, such as switching learning paradigms, adversarial training, pretrained backbones, or parameter recovery. *GUE* [75] models the protector and learner as a game to improve robustness under adversarial training. *BAIT* [76] addresses pretrained-backbone bypass, where strong prior representations reduce the effect of traditional UE. *PUE* [77] studies learnability under parameter recovery or uncertainty, clarifying when protected data remain costly or unreliable to learn.

Deployment-aware workflows. Deployment-oriented work connects training guidance to more realistic data-release workflows. *UC* [69] provides label-agnostic external evaluation, addressing cases where the protector cannot assume the unauthorized learner's label taxonomy. *T2UE* [73] supports zero-contact protection workflows by generating UE from text descriptions rather than requiring access to private images. *Segue* [78] studies pipeline-aware face-image release under side information and adversarial training, with transmission distortion considered in the release pipeline.

4.3 Structured Shortcuts

In deep learning, shortcut learning describes the tendency of models to rely on simple but non-causal cues rather than task-relevant features [194]. Structured shortcuts adapt this idea to UE by explicitly constructing non-semantic signals that unauthorized models learn more readily than task-relevant visual features. The application goal is not to stop training from converging, but to make the learned rule depend on a shortcut that fails under clean deployment. Its **transferability** is mainly assessed through black-box transfer across downstream models, its **adaptability** centers on purification resistance against restoration or noise removal and transformation resistance against filtering or signal-domain processing [83], [84], and its **deployment readiness** develops as shortcut construction moves beyond image-level classification into task-structured settings such as medical segmentation and video tracking [85], [86].

Constructing explicit shortcut signals. *LSP* [79] shows that availability attacks can create learnable shortcuts without detailed victim-model knowledge, and *AR* [80] improves stability through autoregressive local structures. *OPS* [81] shows that shortcut signals can be extremely local, even at the one-pixel level, while *CUDA* [82] constructs class-wise convolutional transformations. Together, these works make shortcut learning an explicit protection mechanism rather than a byproduct of perturbation optimization.

Robust shortcuts under purification and transformations. After shortcuts become explicit, the main adaptability question is whether they remain learnable after countermeasures weaken the signal. *IRP* [83] exploits imperfections in restoration countermeasures, while *KBS* [84] designs frequency-domain shortcuts to resist filtering and visible detail loss. Together, these methods treat restoration and frequency processing as expected parts of the unauthorized training path.

Toward deployment in task-structured settings. *UMed* [85] adapts shortcut design to medical segmentation, where protection must preserve task-relevant spatial structure. *TUE-V* [86] extends shortcut protection to video object tracking, where temporal matching must be considered in protection design. These applications move structured shortcuts beyond image classification benchmarks toward downstream tasks where protection must account for spatial regions, contours, textures, or temporal correspondences.

4.4 Generative UEs

Generative UEs protect images that may be exploited as reference or training data in unauthorized generative pipelines. The central scenarios are style mimicry, concept learning, subject personalization, and customization services. Their **transferability** is usually black-box because the service, backbone, prompt, and customization method are outside the data owner's control. Their **adaptability** is mainly shaped by transformation resistance to compression or service-side processing and purification resistance against diffusion-based purification [96], [91], [101]. **Deployment readiness** is especially visible in artist-facing, privacy-facing, and service-facing studies [88], [96]. Generative UE methods then develop along two deployment directions: artist-facing style and concept protection and privacy-facing subject and identity protection.

Toward artist-facing style and concept protection. *AdvDM* [87] adapts adversarial perturbations to disrupt diffusion-based imitation, establishing an early path from perturbative UE to generative misuse. *Glaze* [88] turns this path into artist-facing style cloaking, while *Nightshade* [89] extends protection from style mimicry to prompt-specific concept-text misalignment. Later style-protection methods strengthen creative-asset protection against diffusion mimicry and style extraction [90], [92].

Toward privacy-facing subject and identity protection. A parallel deployment direction protects personal subjects and identities in personalization services. *Anti-DreamBooth* [94] disrupts subject-driven personalization, while *InMark* [95] targets personalized text-to-image workflows. Subsequent work broadens this privacy-facing direction to protecting personal subjects and identities across few-shot personalization, customization, and low-cost service settings [97],

[99], [100], [98], [93]. This branch is therefore privacy-facing, with personalization and customization services serving as the main deployment setting rather than artist-facing style mimicry.

Countermeasure-aware adaptation across deployment pipelines. Generative UE must also survive the transformations that platforms or downstream users apply before customization. *StyleGuard* [91] incorporates purification-aware style protection, *MetaCloak* [96] improves subject-driven protection under input transformations and online service evaluation, and *AntiPure* [101] directly models diffusion purification followed by customization. This adaptation pressure cuts across both artist-facing and privacy-facing deployments: protected images must remain disruptive after purification, compression, and service-side processing.

4.5 Discussion

Countermeasures. Countermeasure studies show that UE effectiveness depends on adversary capabilities and deployment paths. Some countermeasures recover learnability before training through compression, projection, or variational autoencoder purification [195], [196], [197]. Others bypass protection during training by changing objectives, such as prompt learning with cross-modal alignment [198], or by detecting and filtering UE perturbations [199]. For generative UE, media transformations, purification-customization pipelines, resilience evaluation, and service-leakage purification further show that perturbations may be weakened before or during personalization [200], [201]. Together, these countermeasures motivate cost-of-learning metrics for UE evaluation: beyond whether unauthorized learning is completely prevented, evaluation should measure the additional data, adaptation, compute, or query costs required to recover usable models.

Open problems and future directions. Emerging reuse scenarios further challenge UE by shifting protection from released samples to reusable task experience. In GUI-agent customization, workflow screenshots and action traces may be reused to fine-tune agents, exposing interface states, proprietary procedures, and action policies. UE could limit such unauthorized policy learning while preserving authorized customization. In embodied intelligence, shared robot demonstrations or egocentric observations may reveal manipulation routines and spatial layouts. UE could protect such videos so that collaborators can inspect task execution while making the same data harder to reuse for imitation learning or policy training. These scenarios motivate two directions for future UE research. First, emerging reuse targets workflows, policies, and skills rather than isolated samples, calling for dataset-level, concept-level, or capability-level UE through coordinated data-side interventions. Second, controlled learnability could make UE more practical by moving beyond blanket disruption toward conditional authorization, recovery, model-conditional learnability, target steering, and provenance-aware accountability [202], [203], [204], [205].

5 Proactive Generative Safeguards

Proactive generative safeguards protect visual assets before release so that an existing generative pipeline cannot use them reliably as source images or conditioning references. A protector applies a visually restrained transformation to an image or a small reference set, preserving its visual utility for ordinary viewing and sharing; after release, the asset may be re-encoded, resized, or deliberately purified before an image editor, identity adapter, face swapper, video animator, or talking-head generator processes it. Protection succeeds if the pipeline can no longer produce its intended edit, identity-consistent synthesis, or subject-preserving output under the evaluated release-to-generation path. Whereas unlearnable examples limit what an unauthorized training process can learn from released data, generative safeguards limit what an already available model can produce from protected content during the generation process.

Table 4 summarizes the surveyed methods by **transferability**, **adaptability**, **deployment readiness**, and target model family. This section includes a method only when it evaluates fixed-weight generation; five methods that also evaluate training-based customization are included for their fixed-weight branches [117], [119], [128], [132], [134].

The literature develops along two directions. (1) *Editing immunization* protects an individual released image against direct semantic, identity, or motion manipulation. (2) *Subject personalization safeguards* protect an identity or subject that a fixed-weight pipeline derives from one or more reference images during generation. Face manipulation lies at their boundary: direct modification of a particular image is instance-level editing, whereas extraction of a reusable identity representation is subject personalization.

5.1 Editing Immunization

Editing immunization modifies a released image so that a fixed-weight generator fails to perform the requested manipulation. Its **transferability** evidence spans direct optimization against a known target, transfer within a public diffusion or face-processing lineage, and transfer from surrogates to unseen targets. Its **adaptability** evidence must be read separately: model, checkpoint, or editor changes establish transferability, whereas JPEG compression, resizing, and re-encoding establish routine evidence, and protection-aware purification, removal, or retraining establishes adaptive evidence. Its **deployment readiness** remains largely at the laboratory level; several methods add human perceptual studies, and only a small subset tests external systems. None provides sustained operational evidence.

From fixed editing to controllable manipulation. *PhotoGuard* [102] establishes image immunization against a known editor under a fixed evaluation pipeline. Later work broadens the edited content and the controls available to a user: *EditShield* [107] studies instruction-guided editing, *DiffusionGuard* [104] and *AdvPaint* [119] address masked or inpainting-based manipulation, and *Anti-Inpainting* [120] evaluates changes in masks, prompts, seeds, and routine image transformations. *PromptFlare* [121] instead exploits cross-attention to reduce dependence on a prompt specified during protection construction. These studies broaden the generation controls considered, but variation in a mask, prompt, or seed does not by itself constitute either cross-target transfer or an adaptive countermeasure.

Transfer across generation pipelines. The main transferability question is whether protection constructed on

TABLE 4
Taxonomy of proactive generative safeguards against visual misuse.

Paper	L_1-Transferability	L_2-Adaptability	L_3-Deployment Readiness	Target Model
Editing Immunization				
PhotoGuard[ICML'23] [102]	White-box	Static	Laboratory	Image Editor
PCA[TIFS'25] [103]	Gray-box	Adaptive	Laboratory	Image Editor
DiffusionGuard[ICLR'25] [104]	Gray-box	Adaptive	External	Image Editor
DiffVax[ICLR'26] [105]	Gray-box	Adaptive	External	Image Editor
Distraction[CVPR'24] [106]	Gray-box	Static	Laboratory	Image Editor
EditShield[ECCV'24] [107]	White-box	Routine	External	Image Editor
DCT-Shield[ICCV'25] [108]	Gray-box	Adaptive	External	Image Editor
DeContext[arXiv'25] [109]	White-box	Static	External	Image Editor
DIA[ICCV'25] [110]	Gray-box	Adaptive	Laboratory	Image Editor
Edit Away[CVPR'25] [111]	White-box	Adaptive	Laboratory	Image Editor
Pixel Is Not a Barrier[AAAI'25] [112]	Gray-box	Routine	Laboratory	Image Editor
PSFD[ICME'25] [113]	Gray-box	Static	Laboratory	Image Editor
Semantic Mismatch[arXiv'25] [114]	Black-box	Static	External	Image Editor
TarPro[AAAI'26] [115]	White-box	Adaptive	External	Image Editor
Universal Image Immunization[arXiv'26] [116]	Black-box	Adaptive	External	Image Editor
BlurGuard[NeurIPS'25] [117]	Black-box	Adaptive	Laboratory	Image Editor
Transferable Defense[TPAMI'26] [118]	Gray-box	Static	Laboratory	Image Editor
AdvPaint[ICLR'25] [119]	Black-box	Adaptive	Laboratory	Inpainter
Anti-Inpainting[arXiv'25] [120]	Gray-box	Routine	Laboratory	Inpainter
PromptFlare[ACM MM'25] [121]	Gray-box	Adaptive	Laboratory	Inpainter
My Face Is Mine, Not Yours[arXiv'25] [122]	Gray-box	Routine	Laboratory	Face Swapper
Beauty and the Beast[arXiv'26] [123]	Gray-box	Routine	Laboratory	Face Swapper
Cascading Pathway Disruption[arXiv'26] [124]	Gray-box	Routine	External	Face Swapper
I2VGuard[CVPR'25] [125]	White-box	Adaptive	Laboratory	Video Animator
Anti-I2V[CVPR'26] [126]	Black-box	Adaptive	Laboratory	Video Animator
Vid-Freeze[arXiv'25] [127]	White-box	Routine	Laboratory	Video Animator
DORMANT[USENIX'25] [128]	Black-box	Adaptive	External	Video Animator
Silence Is Golden[CVPR'25] [129]	Gray-box	Adaptive	Laboratory	Talking-Head Generator
SyncBreaker[arXiv'26] [130]	White-box	Adaptive	Laboratory	Talking-Head Generator
Subject Personalization Safeguards				
IDProtector[CVPR'25] [131]	Black-box	Routine	External	ID Adapter
DLADiff[arXiv'25] [132]	Gray-box	Static	External	Subject Generator
AIR[ICME'25] [133]	Black-box	Routine	External	Face Swapper
Targeted Ensemble Defense[InfFus'25] [134]	Gray-box	Adaptive	Laboratory	Subject Generator
Adapter Shield[CVPR'26] [135]	White-box	Adaptive	Laboratory	ID Adapter

one pipeline survives when the eventual target differs. *PCA* [103], *Pixel Is Not a Barrier* [112], and *PSFD* [113] evaluate transfer across related editors, checkpoints, or diffusion components, supporting gray-box evidence when the source and target retain a known lineage. *Semantic Mismatch* [114], *Universal Image Immunization* [116], *BlurGuard* [117], *AdvPaint* [119], and *Anti-I2V* [126] report direct transfer to architecturally distinct targets. *Transferable Defense* [118] also studies source-to-target transfer within a related editor ecosystem. Across these methods, a target-model change belongs to transferability even when the original paper describes it as robustness to *shift*.

Post-release transformations and informed countermeasures. Routine transformations and adaptive removal impose different evidentiary burdens. *EditShield*, *Pixel Is Not a Barrier*, and *Anti-Inpainting* evaluate protection-agnostic operations such as compression, resizing, cropping, or quantization [107], [112], [120]. By contrast, *PCA*, *DiffusionGuard*, *DiffVax*, *DCT-Shield*, *DIA*, *Edit Away*, *TarPro*, *BlurGuard*, and *PromptFlare* explicitly test purification, restoration, or signal removal intended to weaken the protection [103], [104], [105], [108], [110], [111], [115], [117], [121]. Frequency-domain objectives do not automatically imply post-release robustness: *DCT-Shield* evaluates adaptive purification and routine transformations, whereas *PSFD*'s audited evidence supports cross-pipeline transfer but not a qualifying post-release countermeasure [108], [113]. An Adaptive label therefore records that an informed countermeasure was evaluated; it does not assert that protection remained equally strong after that countermeasure.

Scalability and controlled failure. Per-image optimization can make protection expensive to apply at release time. *Distraction* [106] reduces the memory cost of image-specific optimization, *DiffVax* [105] amortizes protection across images, and *Universal Image Immunization* [116] aims to broaden the reuse of a single protection pattern. A complementary line controls how generation fails. *TarPro* [115] steers an edit toward a selected protection outcome, *DeContext* [109] disrupts contextual dependencies used by diffusion-transformer editors, and *DIA* [110] targets inversion-based editing. These design choices improve construction cost or failure control, but neither property alone raises deployment readiness without external evidence.

Identity editing, video, and talking-head generation. Editing immunization now extends beyond static semantic editing. Facial safeguards include *Edit Away* [111], *My Face Is Mine, Not Yours* [122], *Beauty and the Beast* [123], and *Cascading Pathway Disruption* [124]; we place them here when the evaluated misuse directly modifies a supplied face image. For image-to-video generation, *I2VGuard* [125] targets image-conditioned video diffusion, *Anti-I2V* [126] combines color- and frequency-domain objectives, and *Vid-Freeze* [127] disrupts temporal evolution through the protected image. *DORMANT* [128] covers pose-driven human animation and tests six commercial animation services, providing external evidence for those services rather than evidence of sustained deployment. *Silence Is Golden* [129] and *SyncBreaker* [130] extend protection to audio-conditioned talking-head generation. Human evaluations reported by several methods establish external perceptual or utility evidence, but they likewise do not establish operational use.

5.2 Subject Personalization Safeguards

Within this chapter, subject personalization safeguards protect a reusable identity or subject representation that a fixed-weight generation pipeline derives from reference images. The protected reference remains recognizable to people, but an identity adapter, zero-shot subject generator, or face-swapping pipeline should produce identity-inconsistent or otherwise unusable results. In this defining branch, the protected images do not update model parameters; training-based personalization remains within UE unless the same paper also evaluates fixed-weight generation. Their **transferability** evidence ranges from target-specific construction

to transfer across related adapters and independent external systems. Their **adaptability** ranges from fixed evaluation to routine transformations and informed purification or authentication bypass. Their **deployment readiness** is limited to specified services or human studies, and no method demonstrates operational deployment.

Encoder- and adapter-based identity conditioning. *IDProtector* [131] uses a learned protection encoder and evaluates transfer to several identity-conditioning pipelines, including specified closed services. These tests support black-box and external evidence within the evaluated service set, not a claim about all commercial generators. *Adapter Shield* [135] adds a paired authentication mechanism so authorized generation can recover while unauthorized use fails. Its random-password bypass experiment provides limited adaptive evidence, but the authentication design is evaluated as a local prototype and therefore remains at the laboratory level.

Extensions to training-based customization. *DLADiff* [132] evaluates fixed-weight FaceID/Instance-ID generation and further extends protection to DreamBooth/LoRA customization. It also reports transfer across related Stable Diffusion settings and a 10-volunteer mean-opinion-score study. *Targeted Ensemble Defense* [134] similarly evaluates fixed-weight IP-Adapter Plus and PhotoMaker generation alongside DreamBooth/LoRA customization. Its transfer experiments remain within a known Stable Diffusion/SDXL lineage, and its adversarial-purification experiments show that protection degrades under purification. These studies demonstrate that safeguards within the scope of this section can also extend to training-based customization.

Face-swapping boundary. *AIR* [133] constructs protection with a surrogate face-recognition representation and evaluates direct transfer to independent face-swapping targets. An AWS Rekognition measurement of the swapped outputs and a human perceptual study provide two forms of external evidence; the former is an evaluation service rather than an external face-swapping target. In contrast, *My Face Is Mine, Not Yours*, *Beauty and the Beast*, and *Cascading Pathway Disruption* are grouped under editing immunization because their primary evaluated pipeline directly modifies a supplied image [122], [123], [124]. The distinction is therefore determined by the evaluated misuse pipeline: reusable identity extraction is subject personalization, whereas direct image modification is editing immunization.

5.3 Discussion

Countermeasures. Generative safeguards inherit a first-mover disadvantage: the protector commits a signal before release, while a later user can transform the asset or change the generation pipeline. Protection-agnostic compression and resizing constitute routine pressure. Purification, restoration, or learned removal can deliberately attenuate the signal before generation or customization and therefore constitute adaptive pressure [200], [206], [207]. Changing the target editor, adapter, or backbone instead tests transferability unless it is coupled to a protection-aware operation. A defense-aware fine-tuning experiment on protected inputs, as evaluated by *DORMANT* [128], is a cross-stage adaptive countermeasure for that branch; the mere presence of a parallel DreamBooth/LoRA branch in another hybrid study is not, by itself, adaptive evidence.

Open problems and future directions. Three gaps follow directly from the evidence in Table 4. First, transfer studies should report the source–target relationship explicitly, distinguishing shared checkpoint or component lineage from direct transfer to architecturally distinct or independently operated targets. Second, adaptability studies should separate routine sharing operations from informed removal, and should report the residual protection–utility trade-off rather than treating the presence of a purification experiment as proof of resistance. Third, deployment evaluation must move beyond laboratory pipelines, one-time human studies, and a small number of commercial-service tests. Only three methods report external evidence beyond human studies: *DORMANT* and *IDProtector* test specified commercial services, and *AIR* uses a third-party recognition service to score generated outputs; none provides sustained operational evidence. Progress therefore requires evaluations that jointly measure protection strength, visual utility, computational cost, release-platform transformations, service updates, and long-term availability without inflating prototype integration into deployment.

6 Adversarial CAPTCHAs

Adversarial CAPTCHAs apply adversarial learning to human interaction proofs that distinguish legitimate users from automated requests [208], [209], [210], [211]. Whereas generative safeguards protect already-released content from misuse by deployed models, adversarial CAPTCHAs guard the access gate itself: automated systems must pass the platform's verification before they can reach the visual content that the preceding stages protect. As solvers have developed from rule-based recognition to machine learning and foundation models, adversarial CAPTCHA research has expanded beyond character transcription to semantic image understanding and multimodal reasoning [212], [213], [214], [215].

Table 5 summarizes the surveyed methods by **transferability**, **adaptability**, **deployment readiness**, and target model families.

The development of adversarial CAPTCHAs can be organized into three directions. (1) *Character-based CAPTCHAs* strengthen text verification against automated recognition. (2) *Image-based CAPTCHAs* extend protection to semantic object recognition through adversarial and generative designs. (3) *Reasoning-based CAPTCHAs* use semantic, spatial, or logical tasks to challenge multimodal models and agents.

6.1 Character-Based CAPTCHAs

Character-based CAPTCHAs protect text verification against automated transcription. Their **transferability** evidence spans optimization against known recognizers, transfer across solver architectures, and evaluation with third-party OCR software. Their **adaptability** ranges from fixed-pipeline evaluation to resistance against preprocessing and solver adaptation. Their **deployment readiness** remains divided between laboratory studies and external evaluation, with no sustained operational evidence in the surveyed methods.

TABLE 5
Taxonomy of adversarial CAPTCHA methods for visual human verification.

Paper	L_1-Transferability	L_2-Adaptability	L_3-Deployment Readiness	Target Model
Character-Based CAPTCHAs				
rCAPTCHA[TMM'20] [136]	Black-box	Adaptive	External	CNN
Spatial Smoothing[GLOBECOM'21] [137]	White-box	Routine	Laboratory	CNN
MBAGP[Electronics'21] [138]	White-box	Routine	Laboratory	CNN
AECAPTCHA[JPCS'21] [139]	White-box	Static	Laboratory	CNN
aCAPTCHA[TCYB'22] [140]	Gray-box	Routine	Laboratory	CNN
RTC[BigData'22] [141]	Gray-box	Adaptive	External	CNN, OCR
ACG[TDSC'26] [142]	Gray-box	Routine	Laboratory	CNN
Image-Based CAPTCHAs				
DeepCAPTCHA[TIFS'17] [143]	White-box	Adaptive	External	CNN
Robust CAPTCHA Generator[ICAICTA'20] [144]	White-box	Routine	Laboratory	CNN
Capture-the-bot[IEEE Intell. Syst.'20] [145]	Gray-box	Static	External	CNN
Diff-CAPTCHA[arXiv'23] [146]	White-box	Static	External	CNN
DAC[TDSC'25] [147]	Gray-box	Adaptive	Laboratory	CNN
Reasoning-Based CAPTCHAs				
TICS[Vis. Comput.'22] [148]	White-box	Static	Laboratory	CNN
zxCAPTCHA[KST'23] [149], [150]	Gray-box	Routine	External	CNN
IllusionCAPTCHA[WWW'25] [151]	Black-box	Static	External	MLLM
Next-Gen CAPTCHAs[arXiv'26] [152]	Black-box	Adaptive	External	MLLM, GUI Agent

Baseline adversarial perturbation. *AECAPTCHA* [139] applies gradient-based perturbations against a known CNN-based character recognizer without evaluating preprocessing or solver adaptation, establishing model-directed perturbation as an early protection strategy.

Robustness to preprocessing and solver variation. *Spatial Smoothing* [137] targets a known CNN-based character recognizer and is designed to resist spatial smoothing. *rCAPTCHA* [136] evaluates transfer to held-out CNN-based recognizers through a segmentation–recognition pipeline. *aCAPTCHA* [140] extends this evidence across different solver architectures and standard preprocessing filters, while *ACG* [142] evaluates transfer across multiple character recognizers and resistance to preprocessing. *MBAGP* [138] evaluates preprocessing resistance with author-controlled CNN-based recognizers. Together, these studies move evaluation beyond a single fixed recognizer.

Adaptive and external evaluation. *RTC* [141] evaluates a heterogeneous set of character-recognition models, spanning shallow classifiers, neural networks, and OCR models, together with preprocessing, solver adaptation, and human usability. It therefore provides both adaptive and external evidence. Across the character-based literature, however, external evaluation does not yet amount to sustained operation in a deployed verification service.

6.2 Image-Based CAPTCHAs

Image-based CAPTCHAs shift the protected task from character transcription to semantic image recognition. Their **transferability** evidence ranges from known-model optimization to transfer across visual recognizers. Their **adaptability** includes fixed-pipeline evaluation, routine transformations, and informed defenses. Their **deployment readiness** is supported by both laboratory experiments and human-facing external evaluation, but not by operational deployment.

Perturbation-based protection. *DeepCAPTCHA* [143] optimizes adversarial noise against a known CNN-based image recognizer and evaluates resistance to noise removal. *Robust CAPTCHA Generator* [144] optimizes against a known image recognizer while accounting for geometric and photometric transformations. *Capture-the-bot* [145] evaluates localized adversarial patterns across several CNN-based image recognizers without a targeted solver countermeasure. These methods broaden perturbation-based evaluation from a fixed recognizer to transformations, signal removal, and cross-architecture transfer.

Generative protection and stronger solver tests. *Diff-CAPTCHA* [146] incorporates adversarial protection into diffusion-based generation and evaluates the resulting challenges with author-controlled CNN-based image recognizers. *DAC* [147] evaluates its challenges across multiple CNN-based image recognizers and several defensive preprocessing methods. These studies distinguish the mechanism used to generate a challenge from the system used to solve it.

Human-facing evaluation. Human-facing studies accompany the machine-side evaluations of *DeepCAPTCHA*, *Capture-the-bot*, and *Diff-CAPTCHA* [143], [145], [146]. These studies provide external evidence through usability or perceptual-quality assessment, but they do not provide operational deployment evidence.

6.3 Reasoning-Based CAPTCHAs

Reasoning-based CAPTCHAs move the solver bottleneck from recognition alone toward semantic, spatial, or logical interpretation. The surveyed methods include designs evaluated with CNN-based image recognizers and later challenges evaluated with proprietary vision–language systems and GUI-agent systems. Their **transferability** therefore ranges from white-box to black-box evaluation, their **adaptability** ranges from static tests to informed solving strategies, and their **deployment readiness** remains laboratory or external rather than operational.

Recognition-oriented designs. *TICS* [148] is evaluated with an author-trained CNN-based image recognizer under a fixed laboratory setting. *zxCAPTCHA* [149], [150] extends evaluation across several CNN-based image recognizers, preprocessing defenses, and human users. These studies retain recognition-based solvers while increasing the semantic structure of the challenge.

Vision–language and agentic solvers. *IllusionCAPTCHA* [151] uses visual illusions and evaluates stock proprietary vision–language systems together with human users. *Next-Gen CAPTCHAs* [152] evaluates commercial vision–language systems and GUI-agent systems on spatial and multi-attribute tasks. It also tests an informed, family-aware hint strategy, providing adaptive evidence beyond fixed-model evaluation.

From recognition failure to solving cost. Reasoning-oriented evaluation considers human performance along-

side the computational effort required by vision–language and GUI-agent systems [149], [150], [151], [152]. This extends evaluation beyond whether one fixed solver succeeds, while retaining the central requirement that legitimate users can complete the challenge.

6.4 Discussion

Countermeasures. The surveyed evidence identifies three pressures on adversarial CAPTCHA protection. Routine preprocessing can weaken perturbations before recognition, restoration or defensive preprocessing can target the protective signal, and adapted solvers can retrain or change their solving strategy [137], [140], [144], [147], [141], [152]. Results obtained against a fixed recognizer therefore do not establish robustness against an informed solver.

Open problems and future directions. Three gaps remain. Transferability evaluation should distinguish known targets, cross-architecture transfer, and genuinely external solvers more consistently. Adaptability evaluation should test both routine preprocessing and informed solver adjustment, especially as vision–language and GUI-agent systems replace conventional recognizers. Deployment evidence should also move beyond laboratory and one-time user studies toward long-term operation, while reporting human usability together with solver resistance. These requirements follow the same three-axis framework used throughout this survey and avoid treating success against a single solver as evidence of general protection.

7 Adversarial Provenance and Accountability

When privacy filters, unlearnable data, or other prevention-oriented defenses fail, visual assets may already have entered training corpora, model services, or public content streams. **Adversarial provenance and accountability** address this downstream stage: they ask how a data owner, model developer, platform, or user can later prove data use, model copying, content origin, or media manipulation. We focus on evidence that is adversarially constructed, verified against an active counterparty, or stress-tested by attacks, including dataset traces, ownership triggers, diffusion watermarks, and watermark red teaming [153], [158], [171]. The goal is not to promise an indestructible mark, but to make misuse harder to deny under realistic disputes.

Table 6 summarizes representative work through three layers. **L1 Transferability** asks whether verification can move from internal access to owner-keyed, black-box, or content-only disputes. **L2 Adaptability** asks whether evidence remains meaningful under model adaptation, removal, forgery, or semantic manipulation. **L3 Deployment Readiness** asks whether the evidence stays in benchmarks or approaches service, platform, and arbitration workflows. We organize this section around three application scenarios: *(1) adversarial training-asset tracing*, which audits whether visual data were used for training; *(2) adversarial model ownership verification*, which supports claims over copied or adapted model components; and *(3) attack-resilient generated-media provenance*, which traces AIGC images and videos after circulation and manipulation [154], [159], [175].

7.1 Adversarial Training-Asset Tracing

Adversarial training-asset tracing is the post-hoc counterpart of data protection: instead of stopping unauthorized learning, it tries to show that a released visual dataset, annotation set, or sensitive collection contributed to a trained model [153], [154]. The scenario has become practical because web-scale vision and vision-language pipelines mix scraped images, licensed datasets, synthetic samples, and private collections before an external auditor can inspect the model. The application story is ordered by how much control the owner has over the data path: licensed release, web-scale scraping, and mixed-corpus attribution.

Licensed dataset release and ownership disputes. When a dataset owner intentionally releases or licenses visual data, the key application question is how to keep a later ownership claim possible. This is the cleanest tracing setting because the owner can prepare the release. *Radioactive Data* [153] shows that small optimized changes to released images can leave a statistical trace in later models. Dataset watermarking then makes the claim more operational: *DVBW* [155] uses backdoor-style marks so an owner can query a suspected model, while clean-label variants such as *SSCL-BW* [156] and *X-Mark* [157] study stronger evidence under sample modification or medical-domain constraints.

Web-scale scraping and post-hoc audits. For data already circulating on the web, the owner may have no chance to premark the release. *Dataset Inference* [154] instead tests whether a model's black-box behavior reflects a particular dataset, shifting the dispute from a secret mark to statistical evidence. This branch fits post-hoc audits of scraped training data, but it also makes the claim more probabilistic and sensitive to model adaptation.

Mixed-corpus attribution under weak evidence. The least controlled setting is a real training pipeline, where the suspected data may be mixed with public pretraining data, duplicates, synthetic samples, and later fine-tuning sets. Data-use claims must therefore account for deduplication and copyright-evasion strategies [186], [187]. Thus the main deployment challenge is not merely detecting use, but distinguishing misuse from common pretraining, duplicated web images, and distributional similarity.

7.2 Adversarial Model Ownership Verification

Adversarial model ownership verification asks whether a developer can prove that a deployed model, API, adapter, or derivative service came from their original visual model. This scenario appears when checkpoints are copied, APIs are extracted, models are fine-tuned, or components are repackaged in model hubs and service markets. The disputed object has expanded from a single classifier to image-processing networks, GANs, diffusion models, VLMs, VLAs, prompts, and LoRA modules [159], [161], [165], [169]. The application story follows where ownership disputes now occur, moving from whole-model copying to derivative services and then to component reuse in VLM/VLA and adapter ecosystems.

Model marketplace and API-copying disputes. In model marketplaces or hosted APIs, the suspicious artifact may be a copied checkpoint or extracted service rather than a file the owner can inspect. Early model watermarks turn

TABLE 6
Taxonomy of adversarial provenance and accountability methods.

Paper	L_1-Transferability	L_2-Adaptability	L_3-Deployment Readiness	Target Model
Training Asset Tracing				
Radioactive Data[ICML'20] [153]	Gray-box	Media Transformations	Benchmark-scale	Vision DNN
Dataset Inference[arXiv'21] [154]	Black-box Query	Model Adaptation	Benchmark-scale	Vision DNN
DVBW[TIFS'23] [155]	Gray-box	No Adaptive Eval.	Benchmark-scale	Vision DNN
SSCL-BW[arXiv'25] [156]	Gray-box	Evidence Manipulation	Benchmark-scale	Vision DNN
X-Mark[arXiv'26] [157]	Gray-box	Evidence Manipulation	Benchmark-scale	Vision DNN
Model Ownership Verification				
DNN Watermark[AsiaCCS'18] [158]	Gray-box	No Adaptive Eval.	Lab-only	Vision DNN
IPN Watermark[AAAI'20] [159]	Gray-box	Model Adaptation	Benchmark-scale	Vision DNN
Deep Watermarking[TPAMI'21] [160]	Gray-box	Model Adaptation	Benchmark-scale	Vision DNN
Wide-Flat GAN WM[TIFS'24] [161]	Gray-box	Model Adaptation	Benchmark-scale	Image/Video Generator
Free Fine-tuning WM[ACM MM'23] [162]	Gray-box	Evidence Manipulation	Benchmark-scale	Vision DNN
CycleGAN Watermark[TDSC'24] [163]	Gray-box	Model Adaptation	Lab-only	Image/Video Generator
PlugMark[ICCV'25] [164]	Gray-box	Model Adaptation	Benchmark-scale	Image/Video Generator
VLA-Mark[EMNLP'25] [165]	Gray-box	Evidence Manipulation	Lab-only	VLM/MLLM
SWAP[arXiv'25] [166]	Gray-box	No Adaptive Eval.	Benchmark-scale	VLM/MLLM
Cert-LAS[arXiv'26] [167]	Gray-box	Evidence Manipulation	Benchmark-scale	Image/Video Generator
VLA Backdoor Ownership[arXiv'26] [168]	Gray-box	Model Adaptation	Benchmark-scale	VLM/MLLM
LoRA-Key[arXiv'26] [169]	Gray-box	Model Adaptation	Benchmark-scale	Image/Video Generator
SIF[arXiv'26] [170]	Gray-box	Model Adaptation	Benchmark-scale	VLM/MLLM
Media Provenance Verification				
ROBIN[NeurIPS'24] [171]	Gray-box	Evidence Manipulation	Benchmark-scale	Image/Video Generator
ConceptWM[arXiv'24] [172]	Gray-box	Evidence Manipulation	Benchmark-scale	Image/Video Generator
Traceable Adv. Examples[TCSVT'24] [173]	Content-only	Media Transformations	Benchmark-scale	Image/Video Generator
Invisible Adv. WM[TOMM'24] [174]	Gray-box	No Adaptive Eval.	Lab-only	Vision DNN
ZoDiac[NeurIPS'24] [175]	Gray-box	Evidence Manipulation	Benchmark-scale	Image/Video Generator
NoisePrints[arXiv'25] [176]	Gray-box	Evidence Manipulation	Benchmark-scale	Image/Video Generator
Video Signature[arXiv'25] [177]	Gray-box	Evidence Manipulation	Benchmark-scale	Image/Video Generator
BitMark[NeurIPS'26] [178]	Gray-box	Evidence Manipulation	Benchmark-scale	Image/Video Generator
Info-Theoretic AIGI Detector[arXiv'25] [179]	Content-only	Evidence Manipulation	Benchmark-scale	Image/Video Generator
CSGuard[arXiv'26] [180]	Gray-box	Evidence Manipulation	Lab-only	Image/Video Generator
RWP[Neural Networks'26] [181]	Gray-box	Evidence Manipulation	Benchmark-scale	Image/Video Generator
AEON[WACV'26] [182]	Gray-box	Evidence Manipulation	Lab-only	Image/Video Generator
Dual-Guard[arXiv'26] [183]	Content-only	Evidence Manipulation	Benchmark-scale	Image/Video Generator
Robust Content WM[arXiv'26] [184]	Gray-box	Evidence Manipulation	Benchmark-scale	Image/Video Generator
Evidence Manipulation Attacks				
Watermark Radioactivity Eval.[ICLR'25 Workshop] [185]	Gray-box	Model Adaptation	Lab-only	Image/Video Generator
Dataset Copyright Evasion[TIFS'25] [186]	Gray-box	Evidence Manipulation	Benchmark-scale	Image/Video Generator
FT-Traceability Benchmark[CVPR'26] [187]	Gray-box	Evidence Manipulation	Benchmark-scale	Image/Video Generator
MarkSweep[ICASSP'26] [188]	Gray-box	Evidence Manipulation	Benchmark-scale	Image/Video Generator
Forensic-Stealth WM Removal[arXiv'26] [189]	Gray-box	Evidence Manipulation	Benchmark-scale	Image/Video Generator
Fragile Reconstruction[arXiv'26] [190]	Black-box Query	Evidence Manipulation	Lab-only	Image/Video Generator
RAVEN[arXiv'26] [191]	Gray-box	Evidence Manipulation	Benchmark-scale	Image/Video Generator
Frequency-Domain WM Attack[arXiv'26] [192]	Gray-box	Evidence Manipulation	Benchmark-scale	Image/Video Generator

ownership into a behavioral test: the DNN watermarking framework [158] uses trigger-like inputs for remote claims, while *IPN Watermarking* [159] and deep output-barrier watermarking [160] make stolen image-processing networks inherit owner-visible output patterns. Free fine-tuning watermarking later targets disputes where the accused model has been modified or overwritten before verification [162].

Derivative generative services. The next setting is a GAN or diffusion service, where copying is harder to define because the suspect may be a fine-tuned generator, a merged checkpoint, or a model that only preserves part of the original behavior. *Wide-flat GAN watermarking* [161] uses adversarial parameter noise to keep ownership evidence stable under model updates, CycleGAN watermarking studies adapted image generators [163], and diffusion ownership work moves toward derivative verification and certified removal resistance [164], [167]. Red-team evaluation further shows that watermark radioactivity can be weakened by model adaptation [185]. In application terms, the evidence must survive the same operations that make derivative services commercially useful: fine-tuning, pruning, extraction, overwriting, and surrogate training.

Component reuse in model ecosystems. The most modular setting is a modern AI service, where the disputed asset may be a reusable component rather than a whole model. *PlugMark* [164] verifies diffusion derivatives through decision-boundary zero-watermarks, while VLM/VLA and adapter work attaches ownership to cross-modal outputs, soft prompts, embodied observations, and LoRA modules [165], [166], [168], [169]. *SIF* [170] further studies semantically in-distribution fingerprints for VLMs. The remaining application tension is clear: owner secrets help resist evasion, but public arbitration needs evidence that a neutral party can reproduce without leaking reusable triggers or keys.

7.3 Attack-Resilient Generated-Media Provenance

Attack-resilient generated-media provenance addresses the public side of accountability: after an AI-generated image or video leaves the generator, it may be edited, regenerated, screenshotted, compressed, reposted, or stripped of evidence [175], [178], [177]. The core question is no longer whether a clean watermark can be decoded, but whether a platform, creator, journalist, or victim can still support an origin claim after adversarial circulation. The application story follows the media lifecycle: a creator first needs a claim after circulation, an adversary may then create false attribution, and a platform finally has to review partial evidence.

Creator-side claims after content circulation. Creators and generation services need provenance after media have passed through editing tools, social platforms, recompression, reposting, and adversarial watermark erasure. *ROBIN* [171] and *ZoDiac* [175] optimize diffusion watermarks against removal and regeneration, while traceable adversarial examples and invisible adversarial watermarks turn perturbations into copyright evidence [173], [174]. The practical principle is that provenance should be designed for the path media will actually take, not only for clean decoding immediately after generation.

False attribution and ownership disputes. After circulation, the next dispute is adversarial attribution: attackers may not only remove provenance but also forge or transfer it. *ConceptWM* [172] protects visual-concept watermarks under purification and fine-tuning, *NoisePrints* [176] binds authorship to secret diffusion seeds, and *CSGuard* [180] targets forgery-resistant watermark recovery. These methods shift the goal from benign robustness to dispute safety: a useful signal should resist both disappearance and false attribution.

Platform moderation and forensic review. At the end of the lifecycle, platforms, journalists, or moderators often only possess the circulated image or video, not the original prompt, model, or owner key. *BitMark* [178] embeds bit-level evidence in autoregressive generation, video signatures treat temporal tampering as a first-class threat [177], and adaptive watermarking studies cumulative attacks and removal-forgery tradeoffs [182], [184]. Content-only methods such as adversarially robust AIGI detection and *DualGuard* [179], [183] support moderation or forensic review without generator access. At deployment time, these signals may trigger labeling, escalation, creator notification, or human review rather than an automatic verdict, especially when watermark erasure tools are also available [181], [191].

7.4 Discussion

Countermeasures and limitations. The main limitation of adversarial provenance is that evidence becomes another attack surface. Dataset traces may be diluted by dataset mixing or evasion [186], [187]; model watermarks may be weakened by fine-tuning, extraction, or removal [185], [167]; and media watermarks can fail under no-box removal, forensic-stealth attacks, reconstruction, or frequency-domain filtering [188], [189], [190], [192]. Existing countermeasures, such as certified verification, author-proof designs, anti-forgery constraints, and robust content watermarking, protect useful parts of the chain [176], [180], [184], but they rarely provide end-to-end accountability across data owners, model hubs, service providers, platforms, and victims. A protocol remains hard to deploy if it cannot explain who may run the test, what secrets must be revealed, how uncertainty is reported, and how the accused party can contest the result.

Open problems and future directions. The useful lesson for AIGC and agent-era safety is that adversarial thinking should shape the evidence workflow, not only the watermark algorithm. Future provenance systems should connect data-use traces, model-component ownership, adapter reuse, tool calls, generated media, and platform logs into evidence chains [165], [168], [169], [181]. This is especially relevant for native multimodal systems, world-models, vision-language-actions, and multimodal agents: a GUI agent may learn from screenshots and action logs, a robot policy may inherit visual demonstrations, and a video-generation service may reuse private adapters or scene priors. For such systems, provenance cannot treat media files as isolated artifacts; it must bind perception, decision, tool use, and generation records into an auditable workflow. Stronger **L1 Transferability** requires verification that works for auditors without exposing reusable secrets; stronger **L2 Adaptability** requires composed attacks that combine model adaptation, regeneration, false attribution, and trace tampering; stronger **L3 Deployment Readiness** requires API versioning, registries, content-platform workflows, and third-party arbitration. The long-term goal is not a permanent signal, but a calibrated body of evidence whose uncertainty remains interpretable when adversaries attack the weakest link.

8 Conclusion

This survey has followed a single idea across five research communities: the perturbations that expose the fragility of learned models can be turned, by the owners of visual content, into protection. Read in lifecycle order, adversarial privacy filters, unlearnable examples, generative safeguards, adversarial CAPTCHAs, and provenance mechanisms are not isolated literatures but successive answers to the same structural asymmetry, applied at the moments of sharing, release, generation, access, and dispute (Sections 3–7). The shared axes of transferability, adaptability, and deployment readiness (L_1–L_3) let their threat models and robustness claims be weighed on one scale.

Cross-stage countermeasures. Seen side by side, the five families face the same three counterattacks. Adversaries purify, restore, or re-encode a protected asset until a near-clean copy re-enters the pipeline; they swap the pipeline itself, changing backbones, training objectives, or customization protocols until the assumptions behind a protection no longer hold; and once a protective signal can be detected, they strip it or, in the provenance setting, forge it. Beneath all three lies an asymmetry that the protective inversion itself creates: the protector commits a transformation at release time and cannot revise it, while the adversary moves second, with the asset in hand and unlimited attempts. Static validation therefore flatters every mechanism in this survey. Robustness claims are meaningful only against informed adversaries, and they remain conditional on the capability the adversary is assumed to have.

Cross-stage open problems. Four problems recur across the five stages and will decide how far the paradigm carries. First, evaluation is still incommensurable in practice: each community reports robustness against its own surrogates, datasets, and attack suites; adaptive evaluation is the exception; and evidence seldom matures beyond the laboratory. Shared benchmarks that instantiate L_1–L_3 with informed attackers are the most immediate need. Second, the field should retire the promise of absolute prevention. Where removal is possible given enough capability, the honest measures of protection are the cost it imposes, the authorization it can grant or withhold, and the evidence it leaves behind; prevention up front and provenance behind it are layers of one design rather than competing goals. Third, the lifecycle view exposes a question no single community can ask: how protective signals compose. A photograph may need to defeat recognition, resist training, disrupt personalization, and still carry a verifiable mark, all within one imperceptibility budget, yet almost nothing is known about how such signals interfere or how the budget should be divided among them. Fourth, the target is moving. Every stage reports the same shift, away from fixed recognizers and editors toward multimodal models and autonomous agents that can re-perceive, reason about, and retry a protected asset;

protection designed to survive one inference pass must now hold against a compositional stack.

Outlook. The reason to expect this paradigm to endure is the same reason adversarial examples were never engineered away: the perceptual gap between human observers and learned models is a structural property of gradient-trained systems, and every pipeline that automates the use of visual content inherits it. Whoever controls an asset before release can therefore reach into pipelines they will never see, and each new class of systems, multimodal or agentic, renews the opportunity along with the risk. The paradigm's name is meant in both of its senses: these are attacks turned to good ends, and, for as long as AI is built this way, attacks that are here for good.

REFERENCES

[1] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *Proceedings of ICLR*, 2014.

[2] V. Asnani, X. Yin, and X. Liu, "Proactive schemes: A survey of adversarial attacks for social good," *IJCV*, vol. 134, no. 4, p. 186, 2026.

[3] S. Al-Maliki, A. Qayyum, H. Ali, M. Abdallah, J. Qadir, D. T. Hoang, D. Niyato, and A. Al-Fuqaha, "Adversarial machine learning for social good: Reframing the adversary as an ally," *IEEE TAI*, vol. 5, no. 9, pp. 4322–4343, 2024.

[4] L. Laishram, M. Shaheryar, J. T. Lee, and S. K. Jung, "Toward a privacy-preserving face recognition system: A survey of leakages and solutions," *ACM CSUR*, vol. 57, no. 6, Feb. 2025. [Online]. Available: https://doi.org/10.1145/3673224

[5] J. Li, Y. Chen, Y. Xing, Y. Gu, and X. Lan, "A survey on unlearnable data," 2025. [Online]. Available: https://arxiv.org/abs/2503.23536

[6] J. Deng, C. Lin, Z. Zhao, S. Liu, Z. Peng, Q. Wang, and C. Shen, "A survey of defenses against AI-generated visual media: Detection, disruption, and authentication," *ACM CSUR*, 2025.

[7] S. A. Alsuhibany, "A survey on adversarial perturbations and attacks on captchas," *Applied Sciences*, vol. 13, no. 7, 2023. [Online]. Available: https://www.mdpi.com/2076-3417/13/7/4602

[8] X. Zhao, S. Gunn, M. Christ, J. Fairoze, A. Fabrega, N. Carlini, S. Garg, S. Hong, M. Nasr, F. Tramer, S. Jha, L. Li, Y.-X. Wang, and D. Song, "Sok: Watermarking for ai-generated content," 2025. [Online]. Available: https://arxiv.org/abs/2411.18479

[9] E. Wenger, S. Shan, H. Zheng, and B. Y. Zhao, "Sok: Anti-facial recognition technology," in *Proceedings of IEEE S&P*. IEEE, 2023, pp. 864–881.

[10] H.-H. Nguyen-Le, V.-T. Tran, T. Nguyen, and N.-A. Le-Khac, "A survey on proactive deepfake defense: Disruption and watermarking," *ACM CSUR*, vol. 58, no. 5, pp. 1–37, 2025.

[11] Z. Wu, Z. Wang, Z. Wang, and H. Jin, "Towards privacy-preserving visual recognition via adversarial training: A pilot study," in *Proceedings of ECCV*, 2018, pp. 606–624.

[12] S. Shan, E. Wenger, J. Zhang, H. Li, H. Zheng, and B. Y. Zhao, "Fawkes: Protecting privacy against unauthorized deep learning models," in *Proceedings of USENIX Security*, 2020, pp. 1589–1604.

[13] J. Zhang, J. Sang, X. Zhao, X. Huang, Y. Sun, and Y. Hu, "Adversarial privacy-preserving filter," in *Proceedings of ACM MM*, 2020, pp. 1423–1431.

[14] V. Cherepanova, M. Goldblum, H. Foley, S. Duan, J. P. Dickerson, G. Taylor, and T. Goldstein, "Lowkey: Leveraging adversarial attacks to protect social media users from facial recognition," in *Proceedings of ICLR*, 2021.

[15] M. Xue, S. Sun, Z. Wu, C. He, J. Wang, and W. Liu, "Socialguard: An adversarial example based privacy-preserving technique for social images," *Journal of Information Security and Applications*, vol. 63, p. 102993, 2021.

[16] J. Zhang, Q. Yi, D. Lu, and J. Sang, "Low-mid adversarial perturbation against unauthorized face recognition system," *Information Sciences*, vol. 648, p. 119566, 2023.

[17] R. Wu, Y. Wang, H. Shi, Z. Yu, Y. Wu, and D. Liang, "Towards prompt-robust face privacy protection via adversarial decoupling augmentation framework," *arXiv preprint arXiv:2305.03980*, 2023.

[18] W. Zhu, Y. Sun, J. Liu, Y. Cheng, X. Ji, and W. Xu, "Campro: Camera-based anti-facial recognition," in *Proceedings of the Network and Distributed System Security Symposium (NDSS)*, 2024.

[19] X. Liu, Y. Zhong, W. Deng, H. Shi, X. Cui, Y. Yin, and D. Wen, "Enhancing generalization of invisible facial privacy cloak via gradient accumulation," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 5290–5294.

[20] K. Zhang, H. Zhou, J. Zhang, W. Zhou, W. Zhang, and N. Yu, "Transferable facial privacy protection against blind face restoration via domain-consistent adversarial obfuscation," in *Forty-first International Conference on Machine Learning*, 2024.

[21] H. F. Meftah, W. Hamidouche, S. A. Fezza, and O. D'eforges, "Vip: Visual information protection through adversarial attacks on vision-language models," *arXiv preprint arXiv:2507.08982*, 2025.

[22] J. Zhang, C. Wang, Y. Cao, L. Huang, and W. Y. B. Lim, "Disrupting hierarchical reasoning: Adversarial protection for geographic privacy in multimodal reasoning models," *arXiv preprint arXiv:2512.08503*, 2025.

[23] X. Liu, X. Jia, Y. Xun, S. Qin, and X. Cao, "Geoshield: Safeguarding geolocation privacy from vision-language models via adversarial perturbations," in *Proceedings of AAAI*, vol. 40, no. 42, 2026, pp. 35 653–35 661.

[24] Z. Kuang, H. Liu, J. Yu, A. Tian, L. Wang, J. Fan, and N. Babaguchi, "Effective de-identification generative adversarial network for face anonymization," in *Proceedings of the 29th ACM international conference on multimedia*, 2021, pp. 3182–3191.

[25] B. Yin, W. Wang, T. Yao, J. Guo, Z. Kong, S. Ding, J. Li, and C. Liu, "Adv-makeup: A new imperceptible and transferable attack on face recognition," *arXiv preprint arXiv:2105.03162*, 2021.

[26] F. Shamshad, M. Naseer, and K. Nandakumar, "Clip2protect: Protecting facial privacy using text-guided makeup via adversarial latent search," in *Proceedings of CVPR*, 2023, pp. 20 595–20 605.

[27] J. Liu, C. P. Lau, Z. Guo, Y. Guo, Z. Wang, and R. Chellappa, "Diffprotect: Generate adversarial examples with diffusion models for facial privacy protection," *arXiv preprint arXiv:2305.13625*, 2023.

[28] Z. Wang, H. Wang, S. Jin, W. Zhang, J. Hu, Y. Wang, P. Sun, W. Yuan, K. Liu, and K. Ren, "Privacy-preserving adversarial facial features," in *Proceedings of CVPR*, 2023, pp. 8212–8221.

[29] Y. Lyu, Y. Jiang, Z. He, B. Peng, Y. Liu, and J. Dong, "3d-aware adversarial makeup generation for facial privacy protection," *IEEE TPAMI*, vol. 45, no. 11, pp. 13 438–13 453, 2023.

[30] J. Cao, B. Liu, Y. Wen, R. Xie, and L. Song, "Achieving privacy-preserving multi-view consistency with advanced 3d-aware face de-identification," in *Proceedings of the 5th ACM International Conference on Multimedia in Asia*, 2023, pp. 1–7.

[31] M. Li, J. Wang, H. Zhang, Z. Zhou, S. Hu, and X. Pei, "Transferable adversarial facial images for privacy protection," in *Proceedings of ACM MM*, 2024, pp. 10 649–10 658.

[32] D. Liu, X. Wang, C. Peng, N. Wang, R. Hu, and X. Gao, "Adv-diffusion: imperceptible adversarial face identity attack via latent diffusion model," in *Proceedings of AAAI*, vol. 38, no. 4, 2024, pp. 3585–3593.

[33] Y. Sun, L. Yu, H. Xie, J. Li, and Y. Zhang, "Diffam: Diffusion-based adversarial makeup transfer for facial privacy protection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 24 584–24 594.

[34] X. He, M. Zhu, D. Chen, N. Wang, and X. Gao, "Diff-privacy: Diffusion-based face privacy protection," *IEEE TCSVT*, vol. 34, no. 12, pp. 13 164–13 176, 2024.

[35] F. Shamshad, M. Naseer, and K. Nandakumar, "Makeup-guided facial privacy protection via untrained neural network priors," in *Proceedings of ECCV*. Springer, 2024, pp. 227–246.

[36] M.-H. Le and N. Carlsson, "Styleadv: a usable privacy framework against facial recognition with adversarial image editing," *Proceedings on Privacy Enhancing Technologies*, 2024.

[37] J. An, W. Zhang, D. Wu, Z. Lin, J. Gu, and W. Wang, "Sd4privacy: exploiting stable diffusion for protecting facial privacy," in *2024 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2024, pp. 1–6.

[38] S. Yang, B. Zhu, and Z. Yan, "Adversarial 3d generation based on diffusion models for anti-facial recognition," in *2025 International Conference on Information and Automation (ICIA)*. IEEE, 2025, pp. 271–276.

[39] X. Liu, Y. Zhong, X. Cui, Y. Zhang, P. Li, and W. Deng, "Advcloak: Customized adversarial cloak for privacy protection," *Pattern Recognition*, vol. 158, p. 111050, 2025.

[40] J. Zhou, J. Zhang, W. Zhou, C. Yi, and B. Song, "Crfd: A novel face privacy preservation via fine-grained controllable and reversible de-identification," *Expert Systems with Applications*, p. 130386, 2025.

[41] B. M. Le and S. S. Woo, "Machine pareidolia: Protecting facial image with emotional editing," in *Proceedings of AAAI*, vol. 40, no. 42, 2026, pp. 35 580–35 588.

[42] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter, "Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition," in *Proceedings of the 2016 acm sigsac conference on computer and communications security*, 2016, pp. 1528–1540.

[43] E. Chatzikyriakidis, C. Papaioannidis, and I. Pitas, "Adversarial face de-identification," in *2019 IEEE International conference on image processing (ICIP)*. IEEE, 2019, pp. 684–688.

[44] X. Yang, Y. Dong, T. Pang, H. Su, J. Zhu, Y. Chen, and H. Xue, "Towards face encryption by generating adversarial identity masks," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 3897–3907.

[45] E. Lyko and M. Kedziora, "Adversarial attacks on face detection algorithms using anti-facial recognition t-shirts," in *International Conference on Computational Collective Intelligence*. Springer, 2021, pp. 266–277.

[46] Y. Wen, B. Liu, J. Cao, R. Xie, L. Song, and Z. Li, "Identitymask: Deep motion flow guided reversible face video de-identification," *IEEE TCSVT*, vol. 32, no. 12, pp. 8353–8367, 2022.

[47] Y. Zhong and W. Deng, "Opom: Customized invisible cloak towards face privacy protection," *IEEE TPAMI*, vol. 45, no. 3, pp. 3590–3603, 2022.

[48] S. Hu, X. Liu, Y. Zhang, M. Li, L. Y. Zhang, H. Jin, and L. Wu, "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*, 2022, pp. 15 014–15 023.

[49] Z. Pan, J. Sun, X. Li, X. Zhang, and H. Bai, "Collaborative face privacy protection method based on adversarial examples in social networks," in *International Conference on Intelligent Computing*. Springer, 2023, pp. 499–510.

[50] K.-H. Chow, S. Hu, T. Huang, F. Ihan, W. Wei, and L. Liu, "Diversity-driven privacy protection masks against unauthorized face recognition," *Proceedings on Privacy Enhancing Technologies (PoPETs)*, vol. 2024, no. 4, pp. 381–392, 2024.

[51] P. Ying, Z. Li, M. Wei, and X. Xu, "Reversible privacy preserving on vision-language models via adversarial multimodal key," in *Proceedings of ACM MM*, 2025, pp. 3380–3389.

[52] S. Shen, Y. Zhang, D. Ye, X. Shi, L. Tang, H. Duan, Y. Shang, and Z. Tian, "Erasablemask: A robust and erasable privacy protection scheme against black-box face recognition models," *IEEE TMM*, 2025.

[53] F. Zhang, J. Zhang, C. Wang, X. Sun, Y. Hao, G. Guan, W. Li, L. Huang, and W. Y. B. Lim, "Dualtap: A dual-task adversarial protector for mobile mllm agents," *arXiv preprint arXiv:2511.13248*, 2025.

[54] L. Fowl, M. Goldblum, P.-Y. Chiang, J. Geiping, W. Czaja, and T. Goldstein, "Adversarial examples make strong poisons," in *Proceedings of NeurIPS*, vol. 34, 2021, pp. 30 339–30 351.

[55] H. Huang, X. Ma, S. M. Erfani, J. Bailey, and Y. Wang, "Unlearnable examples: Making personal data unexploitable," in *Proceedings of ICLR*, 2021.

[56] Z. Liu, Z. Zhao, A. Kolmus, T. Berns, T. van Laarhoven, T. Heskes, and M. Larson, "Going grayscale: The road to understanding and improving unlearnable examples," *arXiv preprint arXiv:2111.13244*, 2021.

[57] S. Fu, F. He, Y. Liu, L. Shen, and D. Tao, "Robust unlearnable examples: Protecting data privacy against adversarial learning," in *Proceedings of ICLR*, 2022.

[58] B. Fang, B. Li, S. Wu, S. Ding, R. Yi, and L. Ma, "Re-thinking data availability attacks against deep neural networks," in *Proceedings of CVPR*, 2024, pp. 12 215–12 224.

[59] Y. Liu, K. Xu, X. Chen, and L. Sun, "Stable unlearnable example: Enhancing the robustness of unlearnable examples via stable error-minimizing noise," in *Proceedings of AAAI*, vol. 38, no. 4, 2024, pp. 3783–3791.

[60] X. Gong, Y. Wang, Y. Chen, H. Dong, Y. Li, M. Sun, S. Li, and Q. Wang, "Armor: Shielding unlearnable examples against data augmentation," *IEEE TPAMI*, 2026.

[61] Y. Liu, H. Ye, K. Zhang, and L. Sun, "Securing biomedical images from unauthorized training with anti-learning perturbation," *arXiv preprint arXiv:2303.02559*, 2023, also accepted as an NDSS 2023 poster. [Online]. Available: https://www.ndss-symposium.org/wp-content/uploads/2023/02/NDSS2023Poster_paper_9173.pdf

[62] S. Chen, G. Yuan, X. Cheng, Y. Gong, M. Qin, Y. Wang, and X. Huang, "Self-ensemble protection: Training checkpoints are good data protectors," in *Proceedings of ICLR*, 2023.

[63] C.-H. Yuan and S.-H. Wu, "Neural tangent generalization attacks," in *Proceedings of ICML*. PMLR, 2021, pp. 12 230–12 240.

[64] R. Wen, Z. Zhao, Z. Liu, M. Backes, T. Wang, and Y. Zhang, "Is adversarial training really a silver bullet for mitigating data poisoning?" in *Proceedings of ICLR*, 2023.

[65] R. Meng, C. Yi, Y. Yu, S. Yang, B. Shen, and A. C. Kot, "Semantic deep hiding for robust unlearnable examples," *IEEE TIFS*, vol. 19, pp. 6545–6558, 2024.

[66] Y. Zhu, Y. Miao, Y. Dong, and X.-S. Gao, "Why do unlearnable examples work: A novel perspective of mutual information," in *Proceedings of ICLR*, 2026.

[67] J. Ren, H. Xu, Y. Wan, X. Ma, L. Sun, and J. Tang, "Transferable unlearnable examples," in *Proceedings of ICLR*, 2023.

[68] Y. Wang, Y. Zhu, and X.-S. Gao, "Efficient availability attacks against supervised and contrastive learning simultaneously," in *Proceedings of NeurIPS*, vol. 37, 2024, pp. 72 872–72 900.

[69] J. Zhang, X. Ma, Q. Yi, J. Sang, Y.-G. Jiang, Y. Wang, and C. Xu, "Unlearnable clusters: Towards label-agnostic unlearnable examples," in *Proceedings of CVPR*, 2023, pp. 3984–3993.

[70] C. Chen, J. Zhang, Y. Li, and Z. Han, "One for all: A universal generator for concept unlearnability via multi-modal alignment," in *Proceedings of ICML*, 2024.

[71] X. Liu, X. Jia, Y. Xun, S. Liang, and X. Cao, "Multimodal unlearnable examples: Protecting data against multimodal contrastive learning," in *Proceedings of ACM MM*, 2024, pp. 8024–8033.

[72] Y. Sun, H. Zhang, T. Zhang, X. Ma, and Y.-G. Jiang, "Unseg: One universal unlearnable example generator is enough against all image segmentation," in *Proceedings of NeurIPS*, vol. 37, 2024, pp. 79 168–79 193.

[73] X. Ma, H. Huang, T. Song, Y. Sun, Y. Gao, and Y.-G. Jiang, "T2ue: Generating unlearnable examples from text descriptions," in *Proceedings of ACM MM*, 2025, pp. 12 257–12 265.

[74] Z. Li, J. Cai, G. Xu, H. Zheng, Q. Li, F. Zhou, S. Yang, C. Ling, and B. Wang, "Versatile transferable unlearnable example generator," in *Proceedings of NeurIPS*, vol. 38, 2025, pp. 17 495–17 522.

[75] S. Liu, Y. Wang, and X.-S. Gao, "Game-theoretic unlearnable example generator," in *Proceedings of AAAI*, vol. 38, no. 19, 2024, pp. 21 349–21 358.

[76] Z. Li, G. Xu, J. Cai, R. Fang, D. Wu, Q. Lao, C. Ling, and B. Wang, "When priors backfire: On the vulnerability of unlearnable examples to pretraining," in *Proceedings of ICLR*, 2026.

[77] D. Wang, M. Xue, B. Li, S. Camtepe, and L. Zhu, "Provably unlearnable data examples," *Proceedings 2025 Network and Distributed System Security Symposium*, 2025.

[78] Z. Zhang, J. Zhang, K. Zhang, W. Zhou, T. Xu, D. Gao, Z. Guo, Q. Guo, W. Zhang, and N. Yu, "Segue: side-information guided generative unlearnable examples for facial privacy protection in real world," in *Proceedings of ICASSP*. IEEE, 2025, pp. 1–5.

[79] D. Yu, H. Zhang, W. Chen, J. Yin, and T.-Y. Liu, "Availability attacks create shortcuts," in *Proceedings of KDD*, 2022, pp. 2367–2376.

[80] P. Sandoval-Segura, V. Singla, J. Geiping, M. Goldblum, T. Goldstein, and D. W. Jacobs, "Autoregressive perturbations for data poisoning," in *Proceedings of NeurIPS*, vol. 35, 2022, pp. 27 374–27 386.

[81] S. Wu, S. Chen, C. Xie, and X. Huang, "One-pixel shortcut: On the learning preference of deep neural networks," in *Proceedings of ICLR*, 2023.

[82] V. S. Sadasivan, M. Soltanolkotabi, and S. Feizi, "Cuda: Convolution-based unlearnable datasets," in *Proceedings of CVPR*, 2023, pp. 3862–3871.

[83] Y. Huang, J. Styborski, M. Lyu, F. Wang, and A. Kong, "Leveraging imperfect restoration for data availability attack," in *Proceedings of ECCV*. Springer, 2024, pp. 69–86.

[84] J. Li, Y. Chen, Y. Xing, Y. Gu, and X. Lan, "K-space bispectrum steganography for robust unlearnable data," in *Proceedings of ACM MM*, 2025, pp. 11 492–11 501.

[85] X. Lin, Y. Yu, S. Xia, J. Jiang, H. Wang, Z. Yu, Y. Liu, Y. Fu, S. Wang, W. Tang, and A. C. Kot, "Safeguarding medical image segmentation datasets against unauthorized training via contour-and texture-aware perturbations," *arXiv preprint arXiv:2403.14250*, 2024.

[86] Q. Wu, Y. Yu, C. Kong, Z. Liu, J. Wan, H. Li, A. C. Kot, and A. B. Chan, "Temporal unlearnable examples: Preventing personal video data from unauthorized exploitation by object tracking," in *Proceedings of ICCV*, 2025, pp. 11 110–11 121.

[87] C. Liang, X. Wu, Y. Hua, J. Zhang, Y. Xue, T. Song, Z. Xue, R. Ma, and H. Guan, "Adversarial example does good: preventing painting imitation from diffusion models via adversarial examples," in *Proceedings of ICML*, 2023, pp. 20 763–20 786.

[88] S. Shan, J. Cryan, E. Wenger, H. Zheng, R. Hanocka, and B. Y. Zhao, "Glaze: Protecting artists from style mimicry by {Text-to-Image} models," in *Proceedings of USENIX Security*, 2023, pp. 2187–2204.

[89] S. Shan, W. Ding, J. Passananti, S. Wu, H. Zheng, and B. Y. Zhao, "Nightshade: Prompt-specific poisoning attacks on text-to-image generative models," in *Proceedings of IEEE S&P*, 2024, pp. 807–825.

[90] H. Xue, C. Liang, X. Wu, and Y. Chen, "Toward effective protection against diffusion-based mimicry through score distillation," in *Proceedings of ICLR*, 2024.

[91] Y. Li, W. Zhang, X. Lyu, Y. Liu, and B. Xiao, "Styleguard: Preventing text-to-image-model-based style mimicry attacks by style perturbations," in *Proceedings of NeurIPS*, 2025.

[92] Q. Tang, J. Krinsky, and A. Bharati, "Styleprotect: Safeguarding artistic identity in finetuned diffusion models," in *Proceedings of CVPR*, 2026, pp. 10 759–10 769.

[93] N. Ahn, K. Yoo, W. Ahn, D. Kim, and S.-H. Nam, "Nearly zero-cost protection against mimicry by personalized diffusion models," in *Proceedings of CVPR*, 2025, pp. 28 801–28 810.

[94] T. Van Le, H. Phung, T. H. Nguyen, Q. Dao, N. N. Tran, and A. Tran, "Anti-dreambooth: Protecting users from personalized text-to-image synthesis," in *Proceedings of ICCV*, 2023, pp. 2116–2127.

[95] H. Liu, Z. Sun, and Y. Mu, "Countering personalized text-to-image generation with influence watermarks," in *Proceedings of CVPR*, 2024, pp. 12 257–12 267.

[96] Y. Liu, C. Fan, Y. Dai, X. Chen, P. Zhou, and L. Sun, "Metacloak: Preventing unauthorized subject-driven text-to-image diffusion-based synthesis via meta-learning," in *Proceedings of CVPR*, 2024, pp. 24 219–24 228.

[97] X. Xu, S. Kamath, M. A. Butt, and B. Raducanu, "An h-space based adversarial attack for protection against few-shot personalization," in *Proceedings of ACM MM*, 2025, pp. 4904–4913.

[98] L. Xu, J. Wang, H. Hao, H. Qin, J. Zhao, and X. Liu, "Harnessing global-local collaborative adversarial perturbation for anti-customization," in *Proceedings of CVPR*, 2025, pp. 13 414–13 423.

[99] Y. Liu, J. An, W. Zhang, D. Wu, J. Gu, Z. Lin, and W. Wang, "Disrupting diffusion: Token-level attention erasure attack against diffusion-based customization," in *Proceedings of ACM MM*, 2024, pp. 3587–3596.

[100] B. Zheng, C. Liang, and X. Wu, "Targeted attack improves protection against unauthorized diffusion customization," in *Proceedings of ICLR*, 2025.

[101] W. Yang, J. Cao, J. Duan, and R. He, "Towards robust defense against customization via protective perturbation resistant to diffusion-based purification," in *Proceedings of ICCV*, 2025, pp. 19 290–19 300.

[102] H. Salman, A. Khaddaj, G. Leclerc, A. Ilyas, and A. Mądry, "Raising the cost of malicious ai-powered image editing," in *Proceedings of ICML*, 2023, pp. 29 894–29 918.

[103] Z. Guo, C. T. Lei, L. Fang, S. Zhao, Y. Qian, J. Lin, Z. Wang, C. Chen, O. Arandjelović, and C. P. Lau, "A gray-box attack against latent diffusion model-based image editing by posterior collapse," *IEEE TIFS*, vol. 20, pp. 12 918–12 933, 2025.

[104] W. J. S. Choi, K. Lee, J. Jeong, S. Xie, J. Shin, and K. Lee, "Diffusionguard: A robust defense against malicious diffusion-based image editing," in *Proceedings of ICLR*, vol. 2025, 2025, pp. 27 134–27 180.

[105] T. C. Ozden, O. Kara, O. Akcin, K. Zaman, S. Srivastava, S. P. Chinchali, and J. M. Rehg, "Diffvax: Optimization-free image immunization against diffusion-based editing," in *Proceedings of ICLR*, 2026. [Online]. Available: https://openreview.net/forum?id=QEJaKJYOIn

[106] L. Lo, C. Y. Yeo, H.-H. Shuai, and W.-H. Cheng, "Distraction is all you need: Memory-efficient image immunization against diffusion-based image editing," in *Proceedings of CVPR*, 2024, pp. 24 462–24 471.

[107] R. Chen, H. Jin, Y. Liu, J. Chen, H. Wang, and L. Sun, "Editshield: Protecting unauthorized image editing by instruction-guided diffusion models," in *Proceedings of ECCV*. Springer, 2024, pp. 126–142.

[108] A. Bala, R. Chowdhury, R. Jaiswal, and S. Roheda, "Dct-shield: A robust frequency domain defense against malicious image editing," in *Proceedings of ICCV*, 2025, pp. 18 876–18 884.

[109] L. Shen, M. Cui, and X. Yang, "Decontext as defense: Safe image editing in diffusion transformers," *arXiv preprint arXiv:2512.16625*, 2025.

[110] S. Hong, G. Son, J. Lee, and S. S. Woo, "Dia: The adversarial exposure of deterministic inversion in diffusion models," in *Proceedings of ICCV*, 2025, pp. 17 994–18 003.

[111] H. Wang, Y. Zhang, R. Bai, Y. Zhao, S. Liu, and Z. Tu, "Edit away and my face will not stay: Personal biometric defense against malicious generative editing," in *Proceedings of CVPR*, 2025, pp. 23 806–23 816.

[112] C.-Y. Shih, L.-X. Peng, J.-W. Liao, E. Chu, C.-F. Chou, and J.-C. Chen, "Pixel is not a barrier: An effective evasion attack for pixel-domain diffusion models," in *Proceedings of AAAI*, vol. 39, no. 7, 2025, pp. 6905–6913.

[113] L. Zeng, X. Mo, M. Xie, H. Zhang, Y. Liu, Y. Peng, and Y. Li, "Psfd: Proactive spatial-frequency defense against malicious exemplar-guided image editing," in *Proceedings of ICME*. IEEE, 2025, pp. 1–6.

[114] S. Dong, J. Zhang, G. Zhao, S. Shan, and X. Chen, "Semantic mismatch and perceptual degradation: A new perspective on image editing immunity," *arXiv preprint arXiv:2512.14320*, 2025.

[115] K. Shen, R. Quan, J. Miao, and J. Xiao, "Tarpro: Targeted protection against malicious image editing," in *Proceedings of AAAI*, vol. 40, no. 11, 2026, pp. 8896–8904.

[116] C. Lee, S. Shin, D. Choi, H.-g. Jeon, and J. Son, "Universal image immunization against diffusion-based image editing via semantic injection," *arXiv preprint arXiv:2602.14679*, 2026.

[117] J. Kim, Y. Nam, M. Kim, S. Kim, and J. Jeong, "Blurguard: A simple approach for robustifying image protection against ai-powered editing," in *Proceedings of NeurIPS*, vol. 38, 2026, pp. 28 664–28 706.

[118] J. Zhang, S. Dong, S. Shan, and X. Chen, "Towards transferable defense against malicious image edits," *IEEE TPAMI*, 2026.

[119] J. Jeon, W. J. Kim, S. Ha, S. Son, and S.-e. Yoon, "Advpaint: Protecting images from inpainting manipulation via adversarial attention disruption," in *Proceedings of ICLR*, Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu, Eds., vol. 2025, 2025, pp. 76 927–76 940. [Online]. Available: https://proceedings.iclr.cc/paper_files/paper/2025/file/bf67feee86c1711d63cb4e6dc9998702-Paper-Conference.pdf

[120] Y. Guo, Z. Qu, W. Lu, and X. Luo, "Anti-inpainting: A proactive defense approach against malicious diffusion-based inpainters under unknown conditions," *arXiv preprint arXiv:2505.13023*, 2025.

[121] H. Na, S. Hong, and S. S. Woo, "Promptflare: Prompt-generalized defense via cross-attention decoy in diffusion-based inpainting," in *Proceedings of ACM MM*, 2025, pp. 10 544–10 553.

[122] H. M. Yam, Z. Guo, and C. P. Lau, "My face is mine, not yours: Facial protection against diffusion model face swapping," *arXiv preprint arXiv:2505.15336*, 2025.

[123] Y. Huang and S. Li, "Beauty and the beast: Imperceptible perturbations against diffusion-based face swapping via directional attribute editing," *arXiv preprint arXiv:2601.22744*, 2026.

[124] L. Wang, Q. Hu, W. Lu, and X. Luo, "Safeguarding facial identity against diffusion-based face swapping via cascading pathway disruption," *arXiv preprint arXiv:2601.14738*, 2026.

[125] D. Gui, X. Guo, W. Zhou, and Y. Lu, "I2vguard: Safeguarding images against misuse in diffusion-based image-to-video models," in *Proceedings of CVPR*, 2025, pp. 12 595–12 604.

[126] D. Vu, A. Nguyen, C. Tran, and A. Tran, "Anti-i2v: Safeguarding your photos from malicious image-to-video generation," in *Proceedings of CVPR*, 2026, pp. 37 621–37 631.

[127] R. Chowdhury, A. Bala, R. Jaiswal, and S. Roheda, "Vid-freeze: Protecting images from malicious image-to-video generation via temporal freezing," *arXiv preprint arXiv:2509.23279*, 2025.

[128] J. Zhou, M. Wang, T. Li, G. Meng, and K. Chen, "Dormant: Defending against pose-driven human image animation," in *Proceedings of USENIX Security*, 2025, pp. 5209–5228.

[129] Y. Gan, J. Miao, Y. Wang, and Y. Yang, "Silence is golden: Leveraging adversarial examples to nullify audio control in ldm-based talking-head generation," in *Proceedings of CVPR*, 2025, pp. 13 434–13 444.

[130] W. Zhang, X. Shi, S. Zhao, X. Chen, G. Cheng, Y. Xu, T. Xu, and Y. Liao, "Syncbreaker: Stage-aware multimodal adversarial attacks on audio-driven talking head generation," *arXiv preprint arXiv:2604.08405*, 2026.

[131] Y. Song, P. Yang, H. Ci, and M. Z. Shou, "Idprotector: An adversarial noise encoder to protect against id-preserving image generation," in *Proceedings of CVPR*, 2025, pp. 3019–3028.

[132] J. Jia, H. Miao, Y. Zhou, L. Cao, Y. Jiang, W. Zhou, D. Zhu, H. Yang, W. Sun, X. Min *et al.*, "Dladiff: A dual-layer defense framework against fine-tuning and zero-shot customization of diffusion models," *arXiv preprint arXiv:2511.19910*, 2025.

[133] M. Lyu, Y. Huang, J. Xie, Z. Zhao, H. Xu, and K. W.-K. Adams, "Transferable attack against face swapping in an extended space," in *Proceedings of ICME*. IEEE, 2025, pp. 1–6.

[134] M. Hu, Y. Tu, D. Tu, and L. Wang, "Targeted ensemble defense against unauthorized text-to-image identity customization," *Information Fusion*, p. 103696, 2025.

[135] J. Jia, H. Miao, Y. Zhou, W. Zhou, J. Zhang, L. Cao, D. Zhu, H. Yang, X. Min, W. Sun *et al.*, "Adapter shield: A unified framework with built-in authentication for preventing unauthorized zero-shot image-to-image generation," in *Proceedings of CVPR*, 2026, pp. 30 120–30 129.

[136] J. Zhang, J. Sang, K. Xu, S. Wu, X. Zhao, Y. Sun, Y. Hu, and J. Yu, "Robust captchas towards malicious ocr," *IEEE TMM*, vol. 23, pp. 2575–2587, 2020.

[137] Y. Matsuura, H. Kato, and I. Sasase, "Adversarial text-based captcha generation method utilizing spatial smoothing," in *Proceedings of GLOBECOM*. IEEE, 2021, pp. 1–6.

[138] S. Dankwa and L. Yang, "Securing iot devices: A robust and efficient deep learning with a mixed batch adversarial generation process for captcha security verification," *Electronics*, vol. 10, no. 15, p. 1798, 2021.

[139] S. Wang, G. Zhao, and J. Liu, "Text captcha defense algorithm based on overall adversarial perturbations," in *Journal of Physics: Conference Series*, vol. 1744, no. 4. IOP Publishing, 2021, p. 042243.

[140] C. Shi, X. Xu, S. Ji, K. Bu, J. Chen, R. Beyah, and T. Wang, "Adversarial captchas," *IEEE TCYB*, vol. 52, no. 7, pp. 6095–6108, 2022.

[141] R. Shao, Z. Shi, J. Yi, P.-Y. Chen, and C.-J. Hsieh, "Robust text captchas using adversarial examples," in *2022 ieee international conference on big data (big data)*. IEEE, 2022, pp. 1495–1504.

[142] G. Sun, Y. Fu, H. Yang, J. Huang, R. Zhang, and H. Wang, "Enhancing the security of large character set captchas using transferable adversarial examples," *IEEE TDSC*, vol. 23, no. 2, pp. 3898–3915, 2026.

[143] M. Osadchy, J. Hernandez-Castro, S. Gibson, O. Dunkelman, and D. Pérez-Cabo, "No bot expects the deepcaptcha! introducing immutable adversarial examples, with applications to captcha generation," *IEEE TIFS*, vol. 12, no. 11, pp. 2640–2653, 2017.

[144] N. B. Ardhita and N. U. Maulidevi, "Robust adversarial example as captcha generator," in *2020 7th International conference on advance informatics: concepts, theory and applications (ICAICTA)*. IEEE, 2020, pp. 1–4.

[145] D. Hitaj, B. Hitaj, S. Jajodia, and L. V. Mancini, "Capture the bot: Using adversarial examples to improve captcha robustness to bot attacks," *IEEE Intelligent Systems*, vol. 36, no. 5, pp. 104–112, 2021.

[146] R. Jiang, S. Zhang, L. Liu, and Y. Peng, "Diff-captcha: An image-based captcha with security enhanced by denoising diffusion model," *arXiv preprint arXiv:2308.08367*, 2023.

[147] X. Du, X. Liu, J. Zhou, Z. Lin, C.-m. Pun, C. Wu, T. Li, Z. Chen, W. Ni, and J. Luo, "Defensive adversarial captcha: A semantics-driven framework for natural adversarial example generation," *IEEE TDSC*, vol. 23, no. 2, pp. 3423–3435, 2026.

[148] X. Jia, J. Xiao, and C. Wu, "Tics: Text–image-based semantic captcha synthesis via multi-condition adversarial learning," *The Visual Computer*, vol. 38, no. 3, pp. 963–975, 2022.

[149] N. D. Trong, T. H. Huong, and V. T. Hoang, "New cognitive deep-learning captcha," *Sensors*, vol. 23, no. 4, p. 2338, 2023.

[150] N. Dinh, T. Nguyen, and V. Truong, "zxcaptcha: new security-enhanced captcha," in *2023 15th International Conference on Knowledge and Smart Technology (KST)*. IEEE, 2023, pp. 1–6.

[151] Z. Ding, G. Deng, Y. Liu, J. Ding, J. Chen, Y. Sui, and Y. Li, "Illusioncaptcha: A captcha based on visual illusion," in *Proceedings of the ACM on Web Conference 2025*, 2025, pp. 3683–3691.

[152] J. Liu, Y. Luo, J. Cui, X. Shang, X. Zhao, and Z. Shen, "Next-gen captchas: Leveraging the cognitive gap for scalable and diverse gui-agent defense," in *Proceedings of the Forty-Third International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 306, 2026. [Online]. Available: https://openreview.net/forum?id=wPKGirrtyz

[153] A. Sablayrolles, M. Douze, C. Schmid, and H. Jégou, "Radioactive data: tracing through training," in *Proceedings of ICML*. PMLR, 2020, pp. 8326–8335.

[154] P. Maini, M. Yaghini, and N. Papernot, "Dataset inference: Ownership resolution in machine learning," *arXiv preprint arXiv:2104.10706*, 2021.

[155] Y. Li, M. Zhu, X. Yang, Y. Jiang, T. Wei, and S.-T. Xia, "Black-box dataset ownership verification via backdoor watermarking," *IEEE TIFS*, vol. 18, pp. 2318–2332, 2023.

[156] Y. Wang, T. Qiao, X. Liu, C. Li, S. Wu, and J. Li, "Sscl-bw: Sample-specific clean-label backdoor watermarking for dataset ownership verification," *arXiv preprint arXiv:2510.26420*, 2025.

[157] P. Kulkarni, J. Guo, and H. Huang, "X-mark: Saliency-guided robust dataset ownership verification for medical imaging," *arXiv preprint arXiv:2602.09284*, 2026.

[158] J. Zhang, Z. Gu, J. Jang, H. Wu, M. P. Stoecklin, H. Huang, and I. Molloy, "Protecting intellectual property of deep neural networks with watermarking," in *Proceedings of the 2018 on Asia conference on computer and communications security*, 2018, pp. 159–172.

[159] J. Zhang, D. Chen, J. Liao, H. Fang, W. Zhang, W. Zhou, H. Cui, and N. Yu, "Model watermarking for image processing networks," in *Proceedings of AAAI*, vol. 34, no. 07, 2020, pp. 12 805–12 812.

[160] J. Zhang, D. Chen, J. Liao, W. Zhang, H. Feng, G. Hua, and N. Yu, "Deep model intellectual property protection via deep watermarking," *IEEE TPAMI*, vol. 44, no. 8, pp. 4005–4020, 2021.

[161] J. Fei, Z. Xia, B. Tondi, and M. Barni, "Wide flat minimum watermarking for robust ownership verification of gans," *IEEE TIFS*, vol. 19, pp. 8322–8337, 2024.

[162] R. Wang, J. Ren, B. Li, T. She, W. Zhang, L. Fang, J. Chen, and L. Wang, "Free fine-tuning: A plug-and-play watermarking scheme for deep neural networks," in *Proceedings of ACM MM*, 2023, pp. 8463–8474.

[163] D. Lin, B. Tondi, B. Li, and M. Barni, "A cyclegan watermarking method for ownership verification," *IEEE TDSC*, vol. 22, no. 2, pp. 1040–1054, 2024.

[164] P. Chen, Y. Liu, X. Gu, E. Liu, Z. Shang, X. Ji, and W. Liu, "Plugmark: A plug-in zero-watermarking framework for diffusion models," in *Proceedings of ICCV*, 2025, pp. 17 335–17 345.

[165] S. Liu, Z. Qi, J. J. Xu, Y. Yan, J. Zhang, H. Geng, A. Liu, P. Jiang, J. Liu, Y.-C. Tam *et al.*, "Vla-mark: A cross modal watermark for large vision-language alignment models," in *Proceedings of EMNLP*, 2025, pp. 26 420–26 438.

[166] W. Yang, Y. Sun, C. Chen, Z. Chu, J. Zhang, Y. Li, and D. Tao, "Swap: Towards copyright auditing of soft prompts via sequential watermarking," *arXiv preprint arXiv:2511.04711*, 2025.

[167] L. Qi, Y. Li, S. Liang, Z. Tu, and D. Tao, "Cert-las: Toward certified model ownership verification for text-to-image diffusion models via layer-adaptive smoothing," *arXiv preprint arXiv:2605.29809*, 2026.

[168] M. Sun, R. Wang, X. Yu, L. Jing, H. Du, Z. Wan, X. Pan, and I. Tsang, "Towards backdoor-based ownership verification for vision-language-action models," *arXiv preprint arXiv:2605.09005*, 2026.

[169] Y. Wang, Q. Wang, Z. Wang, H. Xu, J. Du, Q. Wang, J.-L. Yin, and K. Ren, "Lora-key: User-centric lora watermarking for text-to-image diffusion models," *arXiv preprint arXiv:2605.29569*, 2026.

[170] Y. Zhao, Q. Lou, and M. Zheng, "Sif: Semantically in-distribution fingerprints for large vision-language models," 2026. [Online]. Available: https://arxiv.org/abs/2604.17041

[171] H. Huang, Y. Wu, and Q. Wang, "Robin: Robust and invisible watermarks for diffusion models with adversarial optimization," in *Proceedings of NeurIPS*, vol. 37, 2024, pp. 3937–3963.

[172] L. Lei, K. Gai, J. Yu, L. Zhu, and Q. Wu, "Watermarking visual concepts for diffusion models," *arXiv preprint arXiv:2411.11688*, 2024.

[173] M. Li, Z. Yang, T. Wang, Y. Zhang, and W. Wen, "Dual protection for image privacy and copyright via traceable adversarial examples," *IEEE TCSVT*, vol. 34, no. 12, pp. 13 401–13 412, 2024.

[174] J. Wang, H. Wang, J. Zhang, H. Wu, X. Luo, and B. Ma, "Invisible adversarial watermarking: A novel security mechanism for enhancing copyright protection," *ACM TOMM*, vol. 21, no. 2, pp. 1–22, 2024.

[175] L. Zhang, X. Liu, A. V. Martin, C. X. Bearfield, Y. Brun, and H. Guan, "Attack-resilient image watermarking using stable diffusion," in *Proceedings of NeurIPS*, vol. 37, 2024, pp. 38 480–38 507.

[176] N. Goren, O. Katzir, A. Nakarmi, E. Ronen, M. Sharif, and O. Patashnik, "Noiseprints: Distortion-free watermarks for authorship in private diffusion models," *arXiv preprint arXiv:2510.13793*, 2025.

[177] Y. Huang, J. Chen, S. Liu, H. Li, J. Li, Q. Zheng, A. Liu, Y. R. Fung, and X. Hu, "Video signature: Implicit watermarking for video diffusion models," *arXiv preprint arXiv:2506.00652*, 2025.

[178] L. Kerner, M. Meintz, B. Zhao, F. Boenisch, and A. Dziedzic, "Bitmark: Watermarking bitwise autoregressive image generative models," in *Proceedings of NeurIPS*, vol. 38, 2026, pp. 99 606–99 640.

[179] R. Zhang, H. Wang, Z. Zhao, Z. Guo, X. Yang, Y. Diao, and M. Wang, "Adversarially robust ai-generated image detection for free: An information theoretic perspective," *arXiv preprint arXiv:2505.22604*, 2025.

[180] J. Lai, L. Zhang, C. Tang, P. Sun, Z. Zhang, Y. Wang, and H. Jin, "Csguard: Toward forgery-resistant watermarking in diffusion models via compressed sensing constraint," *arXiv preprint arXiv:2605.01479*, 2026.

[181] Z. Liu, J. Zhang, Y. Dong, B. Song, and W. Zhou, "Rwp: A robust watermarking plugin for attribution and protection in stable diffusion models," *Neural Networks*, p. 108626, 2026.

[182] M. S. Muneer and S. S. Woo, "Aeon: Adaptive embedding optimized noise for robust watermarking in diffusion models," in *Proceedings of WACV*, 2026, pp. 5406–5415.

[183] J. Xie, C. Ou, P. Yu, X. Zhou, D. Huang, J. Fei, Z. Shen, and Z. Xia, "Dual-guard: Dual-channel latent watermarking for provenance and tamper localization in diffusion images," *arXiv preprint arXiv:2604.19090*, 2026.

[184] Y. Zhu, Y. Wang, and X.-S. Gao, "Towards robust content watermarking against removal and forgery attacks," 2026. [Online]. Available: https://arxiv.org/abs/2604.06662

[185] J. Dubiński, M. Meintz, F. Boenisch, and A. Dziedzic, "Are watermarks for diffusion models radioactive?" in *The 1st Workshop on GenAI Watermarking (WMARK), co-located with ICLR*, 2025.

[186] K. Gao, Y. Zhu, Y. Li, J. Bai, Y. Yang, Z. Li, and S.-T. Xia, "Toward dataset copyright evasion attack against personalized text-to-image diffusion models," *IEEE TIFS*, vol. 21, pp. 725–740, 2025.

[187] X. Wang, H. Sun, W. Sun, K. Xue, W. Zhou, J. Zhang, W. Sun, D. Zhu, X. Min, J. Jia *et al.*, "Evaluating dataset watermarking for fine-tuning traceability of customized diffusion models: A comprehensive benchmark and removal approach," in *Proceedings of CVPR*, 2026, pp. 2230–2239.

[188] J. Cao, Z. Zhang, Q. Li, and J. Ni, "Marksweep: A no-box removal attack on ai-generated image watermarking via noise intensification and frequency-aware denoising," in *Proceedings of ICASSP*. IEEE, 2026, pp. 13 932–13 936.

[189] Y. N. Goonatilake and G. Ateniese, "Removing the watermark is not enough: Forensic stealth in generative-ai watermark removal," *arXiv preprint arXiv:2605.09203*, 2026.

[190] H. Jiang, M. Yi, S. Zhang, J. Cai, Q. Liu, X. Chen, and J. Fan, "Fragile reconstruction: Adversarial vulnerability of reconstruction-based detectors for diffusion-generated images," *arXiv preprint arXiv:2604.12781*, 2026.

[191] F. Shamshad, N. Lukas, and K. Nandakumar, "Raven: Erasing invisible watermarks via novel view synthesis," *arXiv preprint arXiv:2601.08832*, 2026.

[192] C. Wang, B. Qu, X. Wang, Z. Xia, S. Zhang, Y. Liu, and Q. Li, "Breaking watermarks in the frequency domain: A modulated diffusion attack framework," *arXiv preprint arXiv:2604.22220*, 2026.

[193] N. Papernot, P. McDaniel, and I. Goodfellow, "Transferability in machine learning: from phenomena to black-box attacks using adversarial samples," *arXiv preprint arXiv:1605.07277*, 2016.

[194] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann, "Shortcut learning in deep neural networks," *Nat. Mach. Intell.*, vol. 2, no. 11, pp. 665–673, 2020.

[195] Z. Liu, Z. Zhao, and M. Larson, "Image shortcut squeezing: Countering perturbative availability poisons with compression," in *Proceedings of ICML*. PMLR, 2023, pp. 22 473–22 487.

[196] P. Sandoval-Segura, V. Singla, J. Geiping, M. Goldblum, and T. Goldstein, "What can we learn from unlearnable datasets?" in *Proceedings of NeurIPS*, vol. 36, 2023, pp. 75 372–75 391.

[197] Y. Yu, Y. Wang, S. Xia, W. Yang, S. Lu, Y.-P. Tan, and A. C. Kot, "Purify unlearnable examples via rate-constrained variational autoencoders," in *Proceedings of ICML*. PMLR, 2024, pp. 57 678–57 702.

[198] X. Wang, X. Gao, D. Liao, T. Qin, Y.-L. Lu, and C.-Z. Xu, "A^3: Few-shot prompt learning of unlearnable examples with cross-modal adversarial feature alignment," in *Proceedings of CVPR*, 2025, pp. 9507–9516.

[199] Y. Zhu, L. Yu, and X.-S. Gao, "Detection and defense of unlearnable examples," in *Proceedings of AAAI*, vol. 38, no. 15, 2024, pp. 17 211–17 219.

[200] B. Cao, C. Li, T. Wang, J. Jia, B. Li, and J. Chen, "IMPRESS: Evaluating the resilience of imperceptible perturbations against unauthorized data usage in diffusion-based generative ai," in *Proceedings of NeurIPS*, vol. 36, 2023, pp. 10 657–10 677.

[201] Y. Wang, Y. Lu, X.-S. Gao, G. Kamath, and Y. Yu, "BridgePure: Limited protection leakage can break black-box data protection," in *Proceedings of NeurIPS*, vol. 38, 2026, pp. 28 175–28 209.

[202] W. Peng and J. Chen, "Learnability lock: Authorized learnability control through adversarial invertible transformations," in *Proceedings of ICLR*, 2022.

[203] J. Ye and X. Wang, "Ungeneralizable examples," in *Proceedings of CVPR*, 2024, pp. 11 944–11 953.

[204] H. Lee, M. Koo, Y. Song, and N. Kwak, "Targeted data protection for diffusion model by matching training trajectory," in *Proceedings of AAAI*, vol. 40, no. 7, 2026, pp. 5854–5862.

[205] B. Wang, J. Tian, X. Wang, X. Yuan, and J. Li, "Reversible unlearnable examples: Towards the copyright protection in deep learning era," *IEEE TCSVT*, 2025.

[206] Z. Zhao, J. Duan, K. Xu, C. Wang, R. Zhang, Z. Du, Q. Guo, and X. Hu, "Can protective perturbation safeguard personal data from being exploited by stable diffusion?" in *Proceedings of CVPR*, 2024, pp. 24 398–24 407.

[207] Q. Zhao, S. Zhai, X. Bai, Q. Shen, Q. Lin, Y. Gao, and Z. Wu, "Purify once, edit freely: Breaking image protections under model mismatch," *arXiv preprint arXiv:2603.13028*, 2026.

[208] M. D. Lillibridge, M. Abadi, K. Bharat, and A. Z. Broder, "Method for selectively restricting access to computer systems," Feb. 27 2001, uS Patent 6,195,698.

[209] M. Naor, "Verification of a human in the loop or identification via the turing test," *Unpublished draft from http://www. wisdom. weizmann. ac. il/˜ naor/PAPERS/human abs. html*, 1996.

[210] L. Von Ahn, M. Blum, N. J. Hopper, and J. Langford, "Captcha: Using hard ai problems for security," in *International conference on the theory and applications of cryptographic techniques*. Springer, 2003, pp. 294–311.

[211] L. Von Ahn, M. Blum, and J. Langford, "Telling humans and computers apart automatically," *CACM*, vol. 47, no. 2, pp. 56–60, 2004.

[212] K. Chellapilla and P. Simard, "Using machine learning to break visual human interaction proofs (hips)," in *Proceedings of NeurIPS*, vol. 17, 2004.

[213] J. Yan and A. S. El Ahmad, "A low-cost attack on a microsoft captcha," in *Proceedings of ACM CCS*, 2008, pp. 543–554.

[214] E. Bursztein, M. Martin, and J. Mitchell, "Text-based captcha strengths and weaknesses," in *Proceedings of ACM CCS*, 2011, pp. 125–138.

[215] H. Gao, W. Wang, J. Qi, X. Wang, X. Liu, and J. Yan, "The robustness of hollow captchas," in *Proceedings of ACM CCS*, 2013, pp. 1075–1086.