DiffUE: Enhancing Utility-Unlearnability Trade-off of Unlearnable Examples via Diffusion Autoencoders

Syed Irfan Ali Meerza[1,4,*,†], Oktay Ozturk[1,†], Amir Sadovnik[2], and Jian Liu[1,3]

[1] University of Tennessee, Knoxville, TN, USA
[2] Oak Ridge National University, Knoxville, TN, USA
[3] University of Georgia, Athens, GA, USA
[4] Virginia Commonwealth University, Richmond, VA, USA

Abstract. AI models are increasingly trained on personal images scraped from social media and public platforms, often without consent, leading to serious privacy violations, such as unauthorized facial recognition and targeted advertising. To counter this, researchers have developed unlearnable examples (UEs), images modified with imperceptible noise to prevent AI models from extracting meaningful information. However, existing UE methods primarily rely on pixel-space noise, which can be bypassed by relearning strategies such as adversarial training, image transformation, and compression. While some techniques improve robustness, they often come at the expense of significant degradation in image utility and perceptual quality. In this paper, we introduce DiffUE to overcome these limitations by injecting noise into the semantic space of images instead of the pixel space. Instead of corrupting pixel values, DiffUE modifies high-level semantic features of images, ensuring robust unlearnability while preserving visual quality and utility. By leveraging a diffusion-based autoencoder framework to manipulate semantic features, DiffUE generates purposeful, natural-looking modifications that effectively resist advanced relearning strategies. Extensive experiments on four datasets, CIFAR-10, CIFAR-100, CelebA-HQ, and ImageNet, as well as a subjective user study, demonstrate that DiffUE significantly enhances the trade-off between image quality and unlearnability, offering a more robust and effective solution for safeguarding personal data in an increasingly exploitative AI landscape.

1 Introduction

The success of state-of-the-art deep learning models relies heavily on vast datasets, often sourced from *free-to-use* online platforms without explicit consent. This unauthorized acquisition poses significant risks, as seen in cases where personal data, including facial images and medical records, have been exploited in commercial AI models [7,3]. Public concern about these practices has spurred interest in privacy protections, reinforced by regulations such as GDPR [33] and

* Work done while at the University of Tennessee, Knoxville.
† These authors contributed equally.

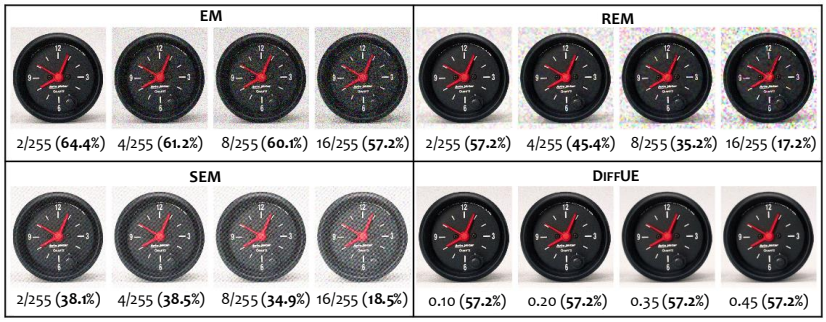


Fig. 1: Usability vs. unlearnability comparison under adversarial training with an adversarial perturbation radius of $\rho_a = 4/255$. We compared DIFFUE with three baselines: EM [9], REM [4] and SEM [20]. The number below each image represents the defensive noise radii (ρ_u) applied (the defensive noise of DIFFUE is in the semantic space), with the value in parentheses indicating the test accuracy achieved. Lower accuracy reflects greater unlearnability. Note that we intentionally selected images with a light background to better visualize perturbation artifacts in existing UE methods.

CCPA [10]. However, these regulations focus primarily on data collection and governance rather than addressing the core issue: once data are acquired and integrated into AI training pipelines, few mechanisms exist to prevent its continued use. To bridge this gap, recent research has introduced the concept of making data unlearnable to AI models. Huang *et al.* [9] proposed unlearnable examples (UE), a technique that adds imperceptible noise to the data, causing models to fail in extracting useful information. This approach uses *error-minimizing* (EM) noise to yield near-zero loss during training, tricking the model into "learning" nothing from the protected data.

Built on this foundation, subsequent research has focused on enhancing the robustness of UEs. For instance, Fu *et al.* [4] proposed robust unlearnable examples (REM), resilient to adversarial training techniques that could otherwise make standard UEs learnable. Similarly, Liu *et al.* [20] further improved the stability of these robust UEs by designing defensive noise that resists random perturbations, thus eliminating the need for complex adversarial perturbation. Expanding on this idea, Ren *et al.* [27] enhanced the transferability of error-minimizing noise using the Classwise Separability Discriminant to maximize model confusion across different data categories. Moreover, application-specific adaptations of unlearnable examples have emerged, targeting areas such as diffusion models [40], time series data [13], and medical image segmentation [19]. Yet, despite their promise, UEs still struggle with two critical challenges that hinder their effectiveness:

(1) Fragility of UEs: Recent studies have exposed significant vulnerabilities in the robustness of UEs. Existing methods rely on error-minimizing noise to render images uninformative to AI models. However, researchers have proposed different relearning strategies based on adversarial training [32] or image transformations like resizing, compression, and grayscale conversion [21,26]. These methods can

significantly weaken the effectiveness of error-minimization noise, rendering these defenses unreliable for long-term data security.

(2) Poor Usability for Real-World Applications: While some approaches have attempted to enhance the robustness of UEs by introducing adversarial perturbation [4] or by stabilizing the UEs using random perturbation [20], they often lead to significant image degradation as unlearnability increases. Stronger noise improves protection but makes images unusable for photographers, artists, and social media users who rely on high-quality visuals. Grainy, artifact-ridden images are impractical for professional work and unappealing for casual sharing. It is evident from Fig. 1 that the robustness comes with the cost of perceptual quality for these methods. Moreover, pixel-space noise disrupts format consistency, leading to compression artifacts and post-processing errors.

DiffUE. To address the aforementioned limitations, we propose DiffUE, a robust defense framework that achieves greatly improved trade-offs between data perceptual quality and unlearnability. Our method is particularly effective against various state-of-the-art data relearning strategies, such as adversarial training/augmentations [4], grayscale conversion [21], data purification [12], and image transformation [26]. Unlike conventional methods that apply pixel-space noise, DiffUE operates in the semantic space, leveraging a pre-trained diffusion autoencoder [25]. Specifically, it encodes an input image x into a high-level semantic representation z_{sem} and a stochastic code x_T, then optimizes a defensive semantic noise z_δ to make the image unlearnable. The final image x_u is reconstructed through a Conditional Denoising Diffusion Implicit Model (DDIM) [30] decoding process. This optimization process ensures that x_u induces a near-zero loss in a surrogate model, ensuring that models fail to extract meaningful features. To maintain both robustness and visual quality, we conduct empirical studies to determine the noise radius ρ_u ensuring that semantic space perturbations translate into effective pixel domain disruptions, changing the image's intrinsic features. By aligning the image's intrinsic features with model gradient space, DiffUE enhances resistance against adversarial retraining, image transformations, and lossy compression. Unlike unnatural, high-frequency "static" artifacts [41] introduced by pixel-based noise, our method guides semantic modifications along the learned image manifold, creating coherent variations such as gentle lighting shifts or subtle tone adjustments. These semantic changes align closely with human perception, appearing natural [35].

Extensive experiments with multiple UE baselines, various relearning strategies, and diverse settings across four datasets demonstrate the superior effectiveness of our system, particularly in challenging relearning scenarios.

2 Related Work

Unlearnable Examples. Unlike poisoning attacks [1,16,29], which degrade model performance by injecting malicious data, unlearnable examples (UEs) [9] defensively add imperceptible perturbations to render data uninformative for training, thereby limiting generalization. Early work explored diverse UE mechanisms, including channel/appearance manipulations [21] and robustness to ad-

versarial training via Robust Unlearnable Examples (REM) [4]. Subsequent efforts improved transferability and stability across models and training pipelines [27,20] and extended UEs to generative diffusion models [40].

More recent work has shifted toward stronger robustness, guarantees, and broader modalities. Meng *et al.* [24] proposed semantic Deep Hiding using invertible neural networks to produce more resilient UE perturbations, while Zhu *et al.* [42] studied the detectability of UEs and corresponding defenses, highlighting a growing arms race. Wang *et al.* [34] provided a certification framework and constructed provably unlearnable examples, offering formal learnability guarantees. Beyond images, Meerza *et al.* [23] introduced HarmonyCloak, extending unlearnable examples to the music domain by leveraging psychoacoustic masking and semantic guidance to protect audio content from generative music models while preserving perceptual quality. Wu *et al.* [37] introduced Temporal Unlearnable Examples for protecting personal video data against tracking-based exploitation, and Li *et al.* [18] investigated versatile transferable UE generators across challenging shifts (e.g., architectures and resolutions). Despite these advances, achieving strong robustness against adaptive retraining while preserving perceptual quality and maintaining broad real-world effectiveness remains challenging.

Relearning from UEs. Recent research has explored countermeasures that attempt to restore learnability to UEs [2,12,26]. One of the earliest approaches by Tao *et al.* [32] showed that Adversarial Training (AT) can weaken UE protections by introducing adversarial perturbations during training. Building on this idea, Qin *et al.* proposed UEraser [26], which combines data augmentation with adversarial training to improve model generalization on protected data. Other methods adopt alternative strategies. Liu *et al.* [21] demonstrated that grayscale transformations can neutralize channel-specific UE noise, improving model robustness. Diffusion-based purification has also gained attention: Jian *et al.* proposed LUE [12], which uses joint-conditional diffusion to restore learnability, while Dolatabadi *et al.* introduced AVATAR [2], a diffusion-based approach that removes defensive noise from UEs. These countermeasures highlight the ongoing arms race between privacy-preserving defenses and adaptive learning strategies, underscoring the need for more resilient UE techniques.

3 Preliminaries

3.1 Diffusion Autoencoders

Diffusion autoencoders unify the strengths of denoising diffusion probabilistic models (DPMs) [8] and Denoising Diffusion Implicit Models (DDIMs) [30] to enable high-fidelity image reconstructions from semantically structured latent codes. DPMs progressively corrupt images with Gaussian noise and use a U-Net denoiser to learn the reverse process, while DDIMs offer a faster, deterministic variant.

Building on this foundation, diffusion autoencoders [25] introduce a semantically structured latent space that enhances controllability and interpretability. The encoder decomposes an image x_0 into two latent variables: a high-level semantic code $z_{sem} = \mathcal{E}(x_0)$ and a low-level stochastic latent x_T. The semantic

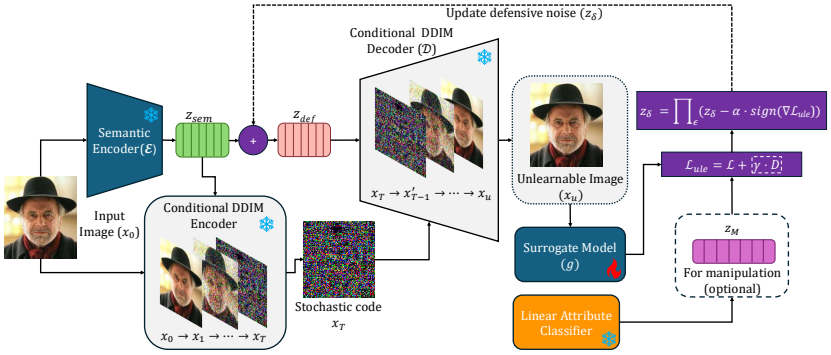


Fig. 2: Overview of DIFFUE: DIFFUE leverages a diffusion autoencoder to encode the input image into a semantic code z_{sem} and a stochastic code x_T. It then iteratively optimizes a defensive semantic code z_δ so that the resulting unlearnable image x_u, generated by the conditional DDIM decoding process, tricks the surrogate model g_θ into believing there is nothing to learn.

code z_{sem}, a compact 512-dimensional vector, captures global, non-spatial semantics akin to the style vector in StyleGAN [15], while x_T preserves local and textural details via reverse DDIM encoding. During decoding, the model reconstructs the image using a DDIM-based generative process conditioned on both z_{sem} and x_T:

$$p_\theta(x_{0:T}|z_{sem}) = p(x_T)\prod_{t=1}^{T} p_\theta(x_{t-1}|x_t, z_{sem}), \tag{1}$$

where $p_\theta(x_{t-1}|x_t, z_{sem}) = \mathcal{N}(f_\theta(x_t, t, z_{sem}), 0)$ for $t = 1$, and follows the DDIM update for other t. The conditional mean function f_θ integrates the semantics z_{sem} to guide denoising, enabling the model to generate coherent and diverse outputs consistent with the encoded structure.

This dual-latent formulation, comprising semantic z_{sem} and stochastic x_T, enables disentangled control over content and style, supporting tasks such as semantic interpolation, content-preserving manipulation, and reconstruction, while preserving the high-fidelity synthesis capabilities enabled by diffusion models.

3.2 Unlearnable Examples

Conventionally, the process of crafting UEs for classification models involves modifying each data sample to make the model's *classification loss* as close to zero as possible. This *zero-loss* approach essentially tricks the model into perceiving these samples as trivial or uninformative, thereby suppressing any gradients that could guide effective learning [9]. To achieve this, consider a data sample x with label y, the defender can craft *error-minimizing* noise δ that, when added to x, minimizes the model's learning potential for this sample. The noise δ is generated by solving a bi-level optimization problem:

$$\arg\min_\theta \quad \mathbb{E}_{x,y}[\min_\delta \mathcal{L}(g_\theta(x+\delta), y)] \quad s.t. \quad \|\delta\|_p \leq \rho, \tag{2}$$

where g_θ denotes the classification model, $\mathcal{L}$ is the cross-entropy loss, and the noise magnitude $\|\delta\|$ is bounded by ρ.

4 Design of DiffUE

4.1 Overview of DiffUE

The overall pipeline of DiffUE is depicted in Fig. 2. DiffUE utilizes a diffusion-based autoencoder architecture [25], comprising a semantic encoder $\mathcal{E}$ and a conditional DDIM, which contains a stochastic encoder and an image decoder $\mathcal{D}$. In this process, an input image x_0 is encoded into a high-level semantic representation z_{sem} and a stochastic code x_T that captures low-level variations. The objective is to optimize a defensive semantic noise z_δ such that the perturbed semantic representation $z_{def} = z_{sem} + z_\delta$ produces an unlearnable image x_u, causing the surrogate model g to experience near-zero loss. Given x_0, we can generate the defensive noise z_δ by solving the bi-level optimization problem:

$$\begin{aligned} \arg\min_{\theta} \mathbb{E}_{(x_0,y)} \left[\min_{z_\delta} \mathcal{L}(g_\theta(\mathcal{D}(\mathcal{E}(x_0) + z_\delta, x_T)), y) \right], \\ s.t. ||z_\delta||_p \leq \rho_u, \end{aligned} \tag{3}$$

where $z_{sem} = \mathcal{E}(x_0)$, $x_u = \mathcal{D}(\mathcal{E}(x_0) + z_\delta, x_T)$ and ρ_u is the noise bound. This formulation is a min-min bi-level optimization problem where the inner minimization seeks to find z_δ, constrained by an L_p-norm, that minimizes the model's classification loss, while the outer minimization optimizes the model parameters θ to further reduce the loss.

To generate effective unlearnable examples, it is crucial to understand the model's learning dynamics, specifically the gradient trajectory of parameters θ during training. Thus, θ must be partially trained initially to reflect the model's typical behavior. This enables δ to be optimized in alignment with learning dynamics, guiding the model toward low training loss. Unlike adversarial examples, which induce prediction failures, here δ is crafted to create the illusion of successful learning. Specifically, we optimize θ for Q steps before updating δ. This process repeats until the model achieves an error rate below a predefined threshold λ, terminating the training.

Additionally, the conditional DDIM is conditioned on the defensive semantic code $z_{def} = z_{sem} + z_\delta$, with the noise prediction network $f(x_T, T, z_{def})$ estimating noise at each step, where T indicates the number of noise addition and removal steps. During the decoding phase, the deterministic generative process reconstructs the final image x_u based on the conditioned input.

4.2 Defensive Noise in the Semantic Space

In DiffUE, we focus on generating defensive noise within the semantic space of diffusion autoencoders. To achieve this, we inject noise into the semantic space using the first-order optimization method, Projected Gradient Descent (PGD) [22]. PGD works by iteratively adjusting the input in the direction of the gradient of a loss function, ensuring that each step stays within a defined perturbation limit. The constrained inner minimization problem is solved with the following update rule:

$$z_{\delta(t+1)} = \Pi_{\rho_u}(z_{\delta(t)} - \eta \cdot \text{sign}(\nabla_{z_\delta} \mathcal{L}_{ule})), \tag{4}$$

where $\mathcal{L}_{ule}$ represents the loss function defined as $\mathcal{L}_{ule} = \mathcal{L}(g_\theta(\mathcal{D}(\mathcal{E}(x_0) + z_\delta, x_T)), y)$, where t is the current step of the perturbation, and η is the step

size. The function Π_{ρ_u} is a projection function that ensures the noise z_δ remains within the allowed perturbation range ρ_u by clipping it if it exceeds the limit. In our framework, PGD iteratively refines the noise z_δ, to hinder the surrogate model g_θ's learning.

Benefits of Injecting Semantic Noises. Semantic perturbations offer a unique advantage over pixel-level noise by creating subtle adjustments that alter an image's underlying attributes without introducing detectable spatial patterns. Human perception, which processes high-level attributes non-linearly, can naturally integrate and overlook these subtle semantic changes, making them imperceptible [35]. In contrast, pixel-space perturbations introduce noticeable patterns perceived as foreign elements [31]. Neural networks, however, respond sensitively even to minor semantic adjustments, directly impacting the linear feature space of model predictions. Thus, effective unlearnable behavior can be achieved with a smaller perturbation budget, reducing pixel-space distortion and allowing our method to outperform previous approaches in visual quality and robustness.

Moreover, semantic perturbations inherently improve robustness by altering high-level attributes instead of pixel-level noise. This makes them less vulnerable to pixel-domain augmentations [31] and filtering techniques [39] commonly used to counteract defensive effects. Since these semantic changes do not present fixed spatial patterns, isolating or removing them becomes considerably more challenging. In the domain of adversarial examples [38], photorealistic perturbations that closely align with the natural characteristics of image objects have proven to be resilient to adversarial training. Using semantic-space transformations, DiffUE improves the trade-off between perceptual quality and unlearnability while enhancing robustness against pixel-space attacks.

4.3 UEs via Controlled Modifications

Adding noise in the semantic space allows for subtle, meaningful adjustments to an image that is generally more acceptable to human observers than pixel-level noise. This enables controlled modifications, such as attribute or style adjustments, allowing users to share visually enhanced images online. In this work, we focus on two modification types: attribute manipulation (e.g., adding a smile) and style manipulation (e.g., adjusting tone or contrast), where defensive noise z_δ alters specific image features.

By manipulating semantic features, our approach can also leverage targeted semantic directions to create purposeful, visible modifications, ensuring both unlearnability and perceptual changes. To achieve this, we adjust the defensive semantic code z_{def} in the direction of the desired modification, determined by the weight vector of a linear classifier trained to distinguish positive and negative instances of the attribute or style. For example, to add a smile, we use a classifier that takes semantic features of an image as input and distinguishes between smiling and non-smiling faces. The classifier's weight vector, which captures the relationship between the semantic features associated with smiling, is then used to guide the defensive noise in that direction. Similarly, for style changes, such as applying certain filter effects, the classifier weights capture the characteristic patterns of these effects in semantic space. The modified loss function incorpo-

rating these controlled changes for effective unlearnability is as follows:

$$\mathcal{L}_{ule} = \mathcal{L}(g_\theta(\mathcal{D}(\mathcal{E}(x_0) + z_\delta, x_T)), y) + \gamma \cdot (1 - cos(z_\delta, z_M)), \tag{5}$$

where z_M is the desired semantic direction calculated by normalizing the weight vector of the relevant classifier $w_{cls}/||w_{cls}||$, $cos(.)$ denotes the cosine similarity, γ is a scaling factor that modulates the degree of desired modifications, and the cosine similarity is defined as D in Fig. 2. Cosine similarity aligns the noise vector in the target direction (direction of the semantics of the desired effect z_M), allowing for controlled, intentional adjustments within the semantic space. This approach enhances practical usability while maintaining resistance to unauthorized usage.

5 Experiments

5.1 Experimental Setup

Datasets and Model Architecture. We conducted experiments on three major vision benchmark datasets: CIFAR-10, CIFAR-100 [17], and the first 100 classes of ImageNet [28]. For data augmentation, we applied random cropping and horizontal flipping. For the experiment of controlled modifications, we used the CelebA-HQ dataset [14], and ImageNet [28]. The proposed method and baselines were evaluated across various vision tasks using ResNet18 [5] as a surrogate model. We use a pre-trained DiffAE [25] model trained on different datasets. Following prior UE works, DiffUE employs an iterative bi-level optimization procedure with 10 PGD inner steps per surrogate model update, and optimization continues until the surrogate model reaches over 99% training accuracy.

UE Baselines. We compare **DiffUE** with state-of-the-art approaches for generating UEs. Specifically, we evaluate against error-minimizing noise (**EM**) [9], which is the first work in the field of UEs and directly applies error minimization noise to the images. We also include robust error-minimizing noise (**REM**) [4], which adds noise designed to reduce adversarial training loss, and stable error-minimizing noise (**SEM**) [20], which generates unlearnable noise that is resilient to random perturbations. These baselines represent the foundational UE generation paradigms widely adopted in prior literature, and many subsequent methods build upon or extend these error-minimization frameworks.

Techniques Used for Robustness Evaluation. To assess the robustness of our method, we apply Mean and Gaussian filters to test the resilience of our images against noise-filtering approaches. We evaluate our approach using recent state-of-the-art relearning techniques designed to counter UEs and restore their learnability. Specifically, we tested **Adversarial Training** [32], which introduces adversarial noise to bypass defensive noise and recover learnability in UEs. We also used **GrayScale** [21], a method that converts UEs to grayscale to neutralize color-channel noise, **UEraser** [26], which employs adversarial augmentations to reintroduce learnability, as well as **LUE** [12], which uses a diffusion purification model to eliminate unlearnable noise from images. Additionally, we evaluate the effectiveness of our approach under common lossy image compression formats, such as JPEG, WebP, and HEIC.

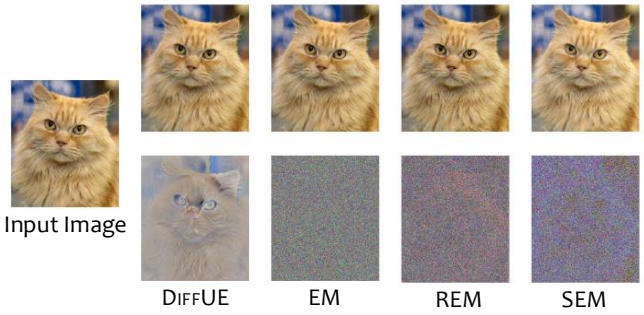


Fig. 3: Visualization of noise types and crafted examples for the ImageNet Subset.

5.2 Semantic Noise Radii Selection

Unlike existing UE methods that constrain noise levels in the pixel space (e.g., 2/255, and 4/255), DiffUE introduces defensive noise in the semantic space, where the noise boundary (i.e., ρ_u in Equation 3) cannot be directly mapped to the pixel space due to non-linear transformations between the two spaces. Moreover, defensive noise in the semantic space generates more photorealistic pixel-level changes, as illustrated in Fig. 3. Even with the same amount of pixel modifications, DiffUE can more effectively conceal perturbations within the original images. Therefore, instead of relying on pixel-level perturbation bounds to assess perceptual impact, as in existing studies, we use widely adopted metrics to quantify image perceptual quality (see section 5.5 for details).

To ensure the semantic noise radius is set to a reasonable value, we empirically measure the maximum pixel changes induced by a given noise level for each dataset. Specifically, we perturb 50 of the 512 dimensions of the semantic code and analyze their impact on the reconstructed image in the pixel space. Based on the analysis on 2,000 samples per dataset, we set the semantic noise radius to 0.20 for CIFAR-10/100 and 0.35 for ImageNet, corresponding to a maximum pixel shift of 4/255.

While increasing the noise radius can further enhance unlearnability, it may also distort the image's semantics. Although the utility-unlearnability trade-off could be further optimized through fine-tuning, we did not adjust the radius further, as our chosen setting already achieves significantly stronger unlearnability and robustness against relearning, even when compared to baselines with perturbation radius of 8/255, while also achieving higher perceptual quality.

5.3 Effectiveness of DiffUE

We applied various UE methods to the entire training set, setting the defensive perturbation radius ρ_u to 8/255 for all baselines. For DiffUE, semantic noise radii (ρ_u) were determined empirically for each dataset: 0.2 for CIFAR-10 and CIFAR-100, and 0.35 for the ImageNet Subset, targeting a maximum pixel-level change of 4/255, corresponding to a 50% smaller noise budget than the baselines. REM and SEM jointly optimize defensive perturbations with an additional noise component, namely adversarial perturbation (radius ρ_a) in REM and random perturbation (radius ρ_r) in SEM. We evaluate different ρ_a and ρ_r

Table 1: Test accuracy (%) of models trained on data protected by different noise strategies under adversarial training with varying perturbation radii. The defensive perturbation radius ρ_u is fixed for all the methods. The training adversarial perturbation radius ρ_a and ρ_r for REM and SEM are adjusted. A lower test accuracy indicates better protective effectiveness.

Dataset	Adv.Train. ρ_t	Clean	EM	DiffUE	REM			SEM		
					$\rho_a = 0$	2/255	4/255	$\rho_r = 0$	2/255	4/255
CIFAR-10	0	95.64	11.98	**10.79**	13.22	23.91	31.27	18.92	18.51	25.14
	2/255	89.24	20.96	**12.59**	68.54	35.37	37.75	20.82	23.97	26.65
	4/255	86.92	76.95	**17.11**	86.55	75.47	64.64	73.57	46.19	46.85
CIFAR-100	0	85.32	**1.76**	2.17	2.11	9.82	19.3	9.81	4.09	18.61
	2/255	76.76	67.29	**3.78**	51.81	10.44	18.35	52.52	10.44	12.56
	4/255	72.19	64.59	**4.92**	65.93	61.29	32.22	60.64	47.52	26.19
ImageNet Subset	0	82.54	**1.89**	2.11	3.45	9.13	17.74	4.15	8.14	10.11
	2/255	73.22	70.67	**12.45**	48.56	30.22	31.40	35.21	33.22	36.22
	4/255	70.23	67.86	**17.20**	65.58	57.23	45.49	57.32	38.11	38.57

Table 2: Test accuracy (%) of various methods trained on the unlearnable CIFAR-10 dataset, evaluated under different filtering techniques and relearning strategies.

Method	Filters		Relearning strategies		
	Mean	Gaussian	GrayScale	UEraser	LUE
Clean	90.65	91.26	91.65	92.78	91.17
EM	37.87	32.71	83.23	90.23	90.33
REM	29.92	28.53	63.87	87.39	89.57
SEM	28.55	24.22	34.23	81.56	89.29
DiffUE	**13.32**	**12.43**	**13.99**	**14.19**	**33.97**

values during unlearnable image generation and subsequently perform adversarial training with varying adversarial radii ρ_t. Table 1 reports the resulting test accuracies on clean data. As ρ_t increases, models trained on baseline unlearnable datasets achieve higher test accuracy, indicating that stronger adversarial training makes these methods less effective. For example, with $\rho_t = 4/255$, test accuracy reaches 76.95%, 64.64%, and 46.85% for EM, REM ($\rho_a = 4/255$), and SEM ($\rho_r = 4/255$), respectively, on CIFAR-10. In contrast, DiffUE maintains a low accuracy of 17.11%, demonstrating greater resistance to adversarial training. A similar trend appears on the ImageNet Subset, where DiffUE remains at 17.20% while REM and SEM increase to 45.49% and 38.57%, respectively. Fig. 3 illustrates the generated perturbations. Unlike pixel-based noise, DiffUE produces perceptually realistic perturbations that preserve visual fidelity while making the protective signal more difficult for adversarial training and filtering methods to remove.

5.4 Robustness Evaluation

We further evaluate the robustness of DiffUE and baseline methods against real-world transformations that could compromise data protection. Table 2 com-

Table 3: Impact of image compression on different defensive noise evaluated on different datasets. Table shows the test accuracy (%) of various methods.

Method	CIFAR-10				ImageNet			
	No Comp.	JPEG	WebP	HEIC	No Comp.	JPEG	WebP	HEIC
EM	11.98	45.22	55.57	35.62	**1.89**	34.21	41.98	37.57
REM	13.22	38.12	41.78	34.95	3.45	33.45	43.22	37.12
SEM	18.92	32.92	35.96	23.89	4.15	18.33	29.45	17.22
DiffUE	**10.79**	**17.22**	**18.05**	**18.11**	2.11	**4.12**	**7.88**	**5.11**

Table 4: Comparison of perceptual metrics with state-of-the-art methods on CelebA-HQ and ImageNet. The defensive perturbation radius for DiffUE is set to $\rho_u = 0.25$ for CelebA-HQ and $\rho_u = 0.35$ for ImageNet, while EM, REM, and SEM use a perturbation radius of $\rho_u = 8/255$ and $\rho_a = 4/255$ for REM and $\rho_r = 4/255$ for SEM.

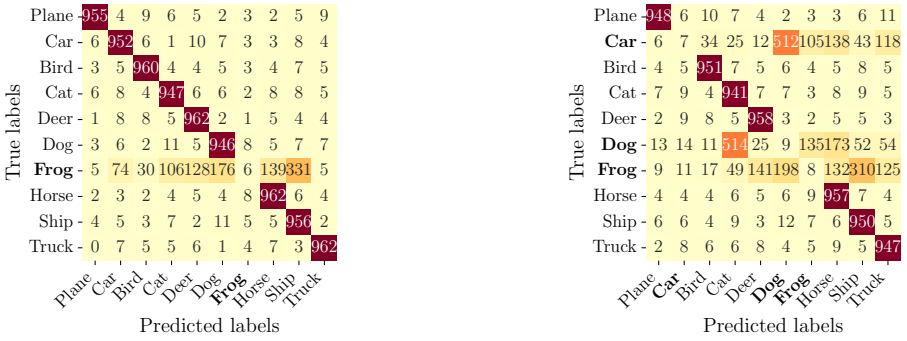
Dataset	Method	EM	REM	SEM	DiffUE
CelebA-HQ	**FID**(↓)	11.612	10.971	10.145	**5.355**
	SSIM(↑)	0.704	0.561	0.611	**0.814**
	PSNR(↑)	30.112	29.411	29.871	**32.438**
	Test Acc. (↓)	55.98%	54.27%	52.14%	**50.22%**
ImageNet	**FID**(↓)	10.583	11.226	11.874	**5.492**
	SSIM(↑)	0.682	0.389	0.576	**0.781**
	PSNR(↑)	29.091	23.295	29.044	**32.934**
	Test Acc. (↓)	**1.89**%	17.74%	10.11%	2.11%

pares the resilience of various unlearnable noise methods under different filtering techniques and relearning strategies. The table shows that DiffUE achieves the lowest test accuracy across all filtering techniques (Mean: 13.32%, Gaussian: 12.43%) and remains more resistant to relearning strategies such as UEraser (14.19%) and LUE (33.97%), compared to the next-best method (SEM: 81.56% under UEraser, 89.29% under LUE). This highlights the robustness of DiffUE against pixel-domain filtering and augmentation.

Given that most images undergo lossy compression before online sharing, we also assess the impact of compression on unlearnable noise in Table 3. DiffUE exhibits minimal degradation when compressed with JPEG (17.22% CIFAR-10, 4.12% ImageNet) or WebP (18.05% CIFAR-10, 7.88% ImageNet), whereas pixel-based baselines degrade significantly (e.g., EM: 45.22% CIFAR-10, 34.21% ImageNet for JPEG). By avoiding pixel-space noise, DiffUE maintains unlearnability under real-world transformations.

5.5 Perceptual Quality Evaluation

We argue that UEs should maintain high perceptual quality so users can use them without compromises. To assess perceptual quality, we use three metrics: Frechet Inception Distance (FID) [6] for naturalness, Structural Similarity Index Measure (SSIM) [36] for similarity to the clean image, and Peak Signal-to-Noise Ratio (PSNR) [11] for image quality under noise. Table 4 compares



(a) Single Class (b) Multiple Classes

Fig. 4: Classwise unlearnability analysis. In (a), we have perturbed only one class and kept the other classes clean, and in (b), we have perturbed multiple classes.

unlearnable examples generated by various methods on CelebA-HQ and ImageNet. On CelebA-HQ, DiffUE achieves the lowest FID (40–50% reduction) with the highest PSNR (32.438) and the highest SSIM (0.814), indicating natural unlearnable examples with low perceptual noise. On ImageNet, DiffUE shows similar trends, with a 40–50% FID improvement and higher SSIM and PSNR than baselines.

5.6 Class-wise Unlearnability Analysis

We evaluate DiffUE's class-specific effect by applying it solely to the 'Frog' class in CIFAR-10, keeping other classes untouched. A ResNet-18 model trained on this partially protected dataset shows poor classification on 'Frog' test samples, mispredicting them as 'Horse', 'Dog', or 'Cat' (Fig. 4a). Meanwhile, performance on untouched classes remains stable, confirming that DiffUE selectively impairs learning without inducing domain shift. Extending this to three classes ('Car', 'Dog', 'Frog'), the confusion matrix in Fig. 4b shows similar degradation for protected classes, while others retain high accuracy. This demonstrates DiffUE 's effectiveness under multi-class settings and its fine-grained control over unlearnability.

5.7 Defensive Noise as Controlled Modifications

DiffUE uses a diffusion autoencoder to guide defensive perturbations toward realistic photo edits, enhancing both unlearnability and perceptual quality. Unlike untargeted noise, which is dispersed in semantic space, control modification aligns perturbations with specific attributes for visible changes. We train a classifier on CelebA-HQ semantic codes (z_{sem}) to identify attributes, using its weights to steer the noise. For implementation, we use a binary gender classifier and set the perturbation radius ρ_u to 0.35, which is higher than the radius used for untargeted noise (0.25).

DiffUE uses a diffusion autoencoder to guide defensive perturbations toward realistic photo edits, enhancing both unlearnability and perceptual quality.

Fig. 5: Defensive noise as attribute modification

Table 5: Test accuracy on CIFAR-10/100 with different model architectures.

Dataset	Model	Clean	EM	REM	SEM	DiffUE
CIFAR-10	VGG-16	85.34	81.50	68.10	54.55	**25.33**
	RN-18	86.92	76.95	64.64	46.85	**17.11**
	RN-50	87.24	79.80	71.22	52.00	**20.05**
	DN-121	84.85	81.70	72.15	55.60	**21.75**
CIFAR-100	VGG-16	61.45	59.77	54.45	59.11	**11.65**
	RN-18	72.19	64.59	32.22	26.19	**4.92**
	RN-50	67.45	69.42	62.78	29.41	**9.03**
	DN-121	52.65	59.89	66.68	50.70	**10.98**

Unlike untargeted noise, which is dispersed in semantic space, controlled modification aligns perturbations with specific attributes for visible changes. We train a classifier on CelebA-HQ semantic codes (z_{sem}) to identify attributes, using its weights to steer the noise. For implementation, we use a binary gender classifier and set the perturbation radius ρ_u to 0.35, which is higher than the radius used for untargeted noise (0.25). The controlled semantic edits are an optional demonstration of semantic-space controllability and are not used in the default DiffUE configuration.

Fig. 5 illustrates the impact of attribute and style manipulations on clean test accuracy. For example, curly hair (51.8%) and black hair (52.1%) modifications reduce accuracy to near-random levels, demonstrating strong unlearnability. Beyond experimental validation, this method helps protect personal images from unauthorized AI training. Attribute modifications can subtly alter profile pictures without degrading quality. By balancing privacy and image integrity, DiffUE enables users to share visually appealing content while minimizing AI exploitation risks.

5.8 Transferability Analysis

To investigate the transferability of different noise methods across neural architectures, we selected ResNet-18 as the source model for generating unlearnable examples (UEs). The impact of the generated noise was evaluated on target mod-

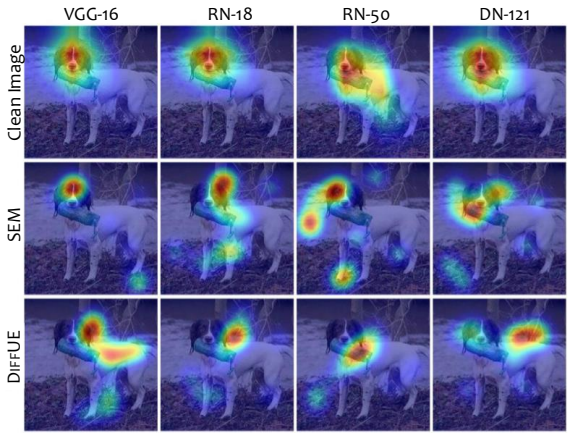


Fig. 6: Comparison of the spatial regions in the input image that contribute to classification for each model when trained on clean and unlearnable examples.

els, including VGG-16, ResNet-50, and DenseNet-121. All transferability experiments were conducted on the CIFAR-10 and CIFAR-100 datasets, as shown in Table 5. DiffUE demonstrated superior performance in generating highly transferable UEs, achieving the lowest test accuracies in all cases. This highlights its effectiveness in surpassing existing methods such as REM and SEM.

Transferring UEs to models with different architectures reduces their effectiveness due to differences in feature representations between the source and target models. As architectural disparity increases, UEs become less effective. For example, ResNet-50, which is structurally similar to ResNet-18, retains higher performance with UEs than VGG-16, which differs significantly in design. This highlights the role of architectural similarity in UE transferability. To further analyze this effect, we examine Grad-CAM visualizations of models trained on clean images and unlearnable examples. Fig. 6 presents results for the same input image using models trained under both settings.

When trained on clean images, all models consistently focus on the same discriminative region, the dog's face. In contrast, when trained on UEs generated using DiffUE, the models shift their attention to other regions unrelated to the dog's face. Notably, even though the unlearnable examples were generated using a ResNet-18 surrogate, the altered attention patterns are consistent across all tested architectures, demonstrating the strong transferability of DiffUE. By comparison, models trained on unlearnable datasets generated by SEM show less consistent attention shifts, different architectures focus on different regions, and in many cases still attend to the dog's face, indicating that SEM exhibits weaker transferability than DiffUE.

5.9 User Study

We conducted a user study to assess the perceptual quality and user preference of DiffUE compared to baseline methods. A total of 56 participants were recruited

Table 6: User study comparison of DiffUE with baselines across perceptual metrics.

Method	Naturalness(↑)	Acceptability (↑)	Artifact(↓)	Preference(↓)
Clean	4.90 ± 0.10	98.5%	1.10 ± 0.15	-
EM	3.76 ± 0.05	56.5%	3.97 ± 0.12	2.86
REM	2.91 ± 0.19	19.2%	4.23 ± 0.07	3.79
SEM	3.11 ± 0.35	61.5%	3.31 ± 0.10	2.13
DiffUE	4.65 ± 0.12	98.5%	1.22 ± 0.10	1.22

via online flyers posted in photography, digital art, and content creation forums (e.g., Reddit), including 26 photographers, 13 digital artists, and 17 social media content creators. Participants independently completed the study on their own devices.

Each participant was shown 20 randomized image sets containing portraits (3 sets contain the participant's own portraits) and landscape images, with each set containing a clean image and four unlearnable versions (EM, REM, SEM with $\rho_u = 4/255$ and DiffUE with $\rho_u = 0.35$). They rated each image using a 5-point Likert scale on naturalness and artifact visibility, and responded to a yes/no question regarding its usability. They were also asked to rank the four UE images from each set based on overall preference. As shown in Table 6, DiffUE achieved a naturalness score of 4.65 ± 0.12, closely matching the clean images (4.90 ± 0.10), and maintained high usability at 98.5%. In contrast, baseline methods such as REM and EM received significantly lower acceptability scores of 19.2% and 56.5%, respectively, along with higher artifact ratings. DiffUE also ranked best in overall preference with an average rank of 1.22, clearly outperforming the other baselines. These results indicate that DiffUE delivers near-clean visual quality while preserving robust privacy characteristics.

6 Conclusion

To confront the challenge of developing robust unlearnable examples, we introduce DiffUE, a defense framework that operates at the semantic level to preserve both unlearnability and visual realism. Built on a diffusion autoencoder, DiffUE injects controlled perturbations into the semantic space, producing images that appear natural yet remain unlearnable even against advanced adversarial relearning strategies. Experimental evaluations demonstrate that DiffUE consistently achieves higher robustness than existing methods against state-of-the-art relearning attacks while preserving data usability across various applications, establishing a reliable path toward privacy-preserving data in adversarial environments.

7 Acknowledgment

We would like to thank our anonymous reviewers and shepherd for their insightful feedback. This work is supported in part by NSF CNS-2114161, ECCS-2132106, CBET-2130643, and CNS-2403529.

References

1. Biggio, B., Nelson, B., Laskov, P.: Poisoning attacks against support vector machines. arXiv preprint arXiv:1206.6389 (2012)
2. Dolatabadi, H.M., Erfani, S., Leckie, C.: The devil's advocate: Shattering the illusion of unexploitable data using diffusion models. In: 2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML). pp. 358–386. IEEE (2024)
3. Edwards, B.: Artist finds private medical record photos in popular AI training data set. https://arstechnica.com/information-technology/2022/09/artist-finds-private-medical-record-photos-in-popular-ai-training-data-set/ (2022), [Accessed 22-06-2026]
4. Fu, S., He, F., Liu, Y., Shen, L., Tao, D.: Robust unlearnable examples: Protecting data against adversarial learning. In: International Conference on Learning Representations (2022)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
6. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
7. Hill, K.: The secretive company that might end privacy as we know it. In: Ethics of Data and Analytics, pp. 170–177. Auerbach Publications (2020)
8. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. NeurIPS (2020)
9. Huang, H., Ma, X., Erfani, S.M., Bailey, J., Wang, Y.: Unlearnable examples: Making personal data unexploitable. In: ICLR (2021)
10. Illman, E., Temple, P.: California consumer privacy act. The Business Lawyer **75**(1), 1637–1646 (2019)
11. Instruments, N.: Peak signal-to-noise ratio as an image quality metric (2013)
12. Jiang, W., Diao, Y., Wang, H., Sun, J., Wang, M., Hong, R.: Unlearnable examples give a false sense of security: Piercing through unexploitable data with learnable examples. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 8910–8921 (2023)
13. Jiang, Y., Ma, X., Erfani, S.M., Bailey, J.: Unlearnable examples for time series. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining. pp. 213–225. Springer (2024)
14. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of gans for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196 (2017)
15. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4401–4410 (2019)
16. Koh, P.W., Liang, P.: Understanding black-box predictions via influence functions. In: International conference on machine learning. pp. 1885–1894. PMLR (2017)
17. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
18. Li, Z., Cai, J., Xu, G., Zheng, H., Li, Q., Zhou, F., Yang, S., Ling, C., Wang, B.: Versatile transferable unlearnable example generator. Advances in Neural Information Processing Systems **38**, 17495–17522 (2026)
19. Lin, X., Yu, Y., Xia, S., Jiang, J., Wang, H., Yu, Z., Liu, Y., Fu, Y., Wang, S., Tang, W., et al.: Safeguarding medical image segmentation datasets against

unauthorized training via contour-and texture-aware perturbations. arXiv preprint arXiv:2403.14250 (2024)
20. Liu, Y., Xu, K., Chen, X., Sun, L.: Stable unlearnable example: Enhancing the robustness of unlearnable examples via stable error-minimizing noise. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 3783–3791 (2024)
21. Liu, Z., Zhao, Z., Kolmus, A., Berns, T., van Laarhoven, T., Heskes, T., Larson, M.: Going grayscale: The road to understanding and improving unlearnable examples. arXiv preprint arXiv:2111.13244 (2021)
22. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: International Conference on Learning Representations (2018)
23. Meerza, S.I.A., Sun, L., Liu, J.: Harmonycloak: Making music unlearnable for generative ai. In: 2025 IEEE Symposium on Security and Privacy (SP). pp. 430–448. IEEE (2025)
24. Meng, R., Yi, C., Yu, Y., Yang, S., Shen, B., Kot, A.C.: Semantic deep hiding for robust unlearnable examples. IEEE Transactions on Information Forensics and Security (2024)
25. Preechakul, K., Chatthee, N., Wizadwongsa, S., Suwajanakorn, S.: Diffusion autoencoders: Toward a meaningful and decodable representation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10619–10629 (2022)
26. Qin, T., Gao, X., Zhao, J., Ye, K., Xu, C.Z.: Learning the unlearnable: Adversarial augmentations suppress unlearnable example attacks. arXiv preprint arXiv:2303.15127 (2023)
27. Ren, J., Xu, H., Wan, Y., Ma, X., Sun, L., Tang, J.: Transferable unlearnable examples. arXiv preprint arXiv:2210.10114 (2022)
28. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International journal of computer vision **115**(3), 211–252 (2015)
29. Shafahi, A., Huang, W.R., Najibi, M., Suciu, O., Studer, C., Dumitras, T., Goldstein, T.: Poison frogs! targeted clean-label poisoning attacks on neural networks. In: NeurIPS (2018)
30. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: International Conference on Learning Representations (2021)
31. Song, Y., Shu, R., Kushman, N., Ermon, S.: Constructing unrestricted adversarial examples with generative models. Advances in neural information processing systems **31** (2018)
32. Tao, L., Feng, L., Yi, J., Huang, S.J., Chen, S.: Better safe than sorry: Preventing delusive adversaries with adversarial training. In: Advances in neural information processing systems. pp. 16209–16225 (2021)
33. Voigt, P., Von dem Bussche, A.: The eu general data protection regulation (gdpr). A Practical Guide, 1st Ed., Cham: Springer International Publishing **10**(3152676), 10–5555 (2017)
34. Wang, D., Xue, M., Li, B., Camtepe, S., Zhu, L.: Provably unlearnable data examples. arXiv preprint arXiv:2405.03316 (2024)
35. Wang, S., Chen, S., Chen, T., Nepal, S., Rudolph, C., Grobler, M.: Generating semantic adversarial examples via feature manipulation in latent space. IEEE Transactions on Neural Networks and Learning Systems (2023)
36. Wang, Z., Bovik, A.C., Sheikh, H.R.: Structural similarity based image quality assessment. In: Digital Video image quality and perceptual coding, pp. 225–242. CRC Press (2017)

37. Wu, Q., Yu, Y., Kong, C., Liu, Z., Wan, J., Li, H., Kot, A.C., Chan, A.B.: Temporal unlearnable examples: Preventing personal video data from unauthorized exploitation by object tracking. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11110–11121 (2025)
38. Xiao, C., Li, B., Zhu, J., He, W., Liu, M., Song, D.: Generating adversarial examples with adversarial networks. In: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI). pp. 3905–3911 (2018)
39. Xie, C., Wu, Y., Maaten, L.v.d., Yuille, A.L., He, K.: Feature denoising for improving adversarial robustness. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 501–509 (2019)
40. Zhao, Z., Duan, J., Hu, X., Xu, K., Wang, C., Zhang, R., Du, Z., Guo, Q., Chen, Y.: Unlearnable examples for diffusion models: Protect data from unauthorized exploitation. arXiv preprint arXiv:2306.01902 (2023)
41. Zhu, P., Osada, G., Kataoka, H., Takahashi, T.: Frequency-aware gan for adversarial manipulation generation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4315–4324 (2023)
42. Zhu, Y., Yu, L., Gao, X.S.: Detection and defense of unlearnable examples. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 17211–17219 (2024)