Synthetic Forgetting Without Access: A Few-shot Zero-glance Framework for Machine Unlearning

Qipeng Song[1], Nan Yang[1], Ziqi Xu[2], Yue Li[1]*, Wei Shao[3], Feng Xia[2]
[1]Xidian University
[2]Royal Melbourne Institute of Technology
[3]Data61, CSIRO
{qpsong, Liyue}@xidian.edu.cn, nanyang@stu.xidian.edu.cn
{ziqi.xu,feng.xia}@rmit.edu.au ,Phdweishao@gmail.com

Abstract

Machine unlearning aims to eliminate the influence of specific data from trained models to ensure privacy compliance. However, most existing methods assume full access to the original training dataset, which is often impractical. We address a more realistic yet challenging setting: *few-shot zero-glance*, where only a small subset of the retained data is available and the forget set is entirely inaccessible. We introduce GFOES, a novel framework comprising a Generative Feedback Network (GFN) and a two-phase fine-tuning procedure. GFN synthesises Optimal Erasure Samples (OES), which induce high loss on target classes, enabling the model to forget class-specific knowledge without access to the original forget data, while preserving performance on retained classes. The two-phase fine-tuning procedure enables aggressive forgetting in the first phase, followed by utility restoration in the second. Experiments on three image classification datasets demonstrate that GFOES achieves effective forgetting at both logit and representation levels, while maintaining strong performance using only 5% of the original data. Our framework offers a practical and scalable solution for privacy-preserving machine learning under data-constrained conditions.

Code — https://github.com/sheltparkle/OES-Unlearning
Extended version — https://arxiv.org/abs/2511.13116

1 Introduction

In recent years, growing awareness of personal data rights has led to the introduction of various privacy regulations, including the General Data Protection Regulation (GDPR) (Voigt and Von dem Bussche 2017), the California Consumer Privacy Act (CCPA) (Goldman 2020), and China's Personal Information Protection Law (PIPL) (Calzada 2022). These laws enshrine the *Right to be Forgotten* (Villaronga, Kieseberg, and Li 2018), which grants individuals the right to request data deletion. For Machine Learning as a Service (MLaaS) providers, complying with this right requires not only permanently deleting personal data from storage systems, but also removing any knowledge derived from it and embedded in trained models. This challenge has given rise to a new research direction: *machine unlearning* (Cao and Yang 2015).

*Corresponding author

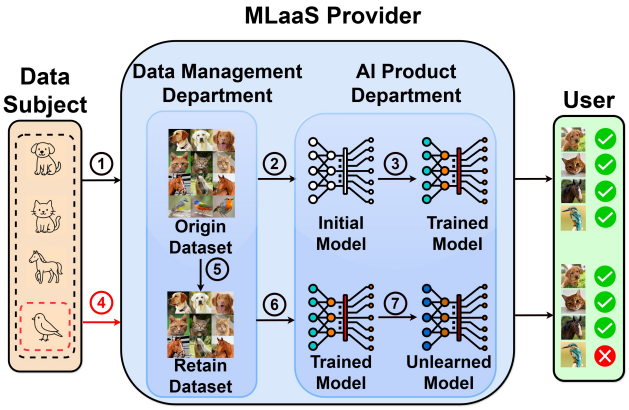

Figure 1: Overview of how an MLaaS provider processes data deletion requests under the *Right to be Forgotten*. After data collection (①) and initial model training (②–③), a deletion request (④) triggers removal of the specified data (e.g., the bird image) from storage (⑤). The updated dataset is used for unlearning (⑥–⑦), and the resulting model is then served to users.

To enforce the *Right to be Forgotten*, MLaaS providers must ensure that both user data and its learned influence are thoroughly removed. As illustrated in Figure 1, the unlearning process involves two key objectives: (i) deleting the user's data from storage, and (ii) removing its impact from the trained model. Importantly, data deletion must precede unlearning to avoid unauthorised retention or reprocessing. This order aligns with Article 17(1) of the GDPR, which requires that personal data be erased *without undue delay* once conditions such as consent withdrawal or data irrelevance are met. Performing unlearning before deletion risks caching or duplicating the data during intermediate steps, thereby violating GDPR principles like storage limitation and potentially leading to unlawful processing.

From both legal and technical perspectives, it is therefore essential that unlearning is performed *after* irreversible data deletion. This order ensures that the learning system cannot access or replicate deleted data, thereby reinforcing privacy-preserving compliance and fostering trust in responsible AI systems. Furthermore, MLaaS providers are eco-

nomically motivated to minimise the amount of data used in the unlearning process, as such compliance efforts incur operational costs without generating direct revenue. These constraints give rise to a new and practical unlearning setting, which we refer to as the *few-shot zero-glance*: *few-shot* reflects the minimal reliance on retained training data, while *zero-glance* denotes the complete lack of access to the deleted data throughout the unlearning process.

Despite rapid progress in machine unlearning, existing methods fall short of meeting the stringent demands of the *few-shot zero-glance* setting. Prior work generally falls into two categories: exact unlearning (Wu, Dobriban, and Davidson 2020; Yan et al. 2022) and approximate unlearning (Liu et al. 2022a; Shaik et al. 2023; Tanno et al. 2022). However, most methods in both categories assume full access to either the entire training set or the data to be forgotten. While a few methods (Guo et al. 2020) only consider the *few-shot* setting, they still require access to the target data during unlearning and thus violate the *zero-glance* setting. This highlights an urgent need for unlearning strategies that enforce both data minimisation and strict non-access guarantees, while preserving strong predictive performance.

To tackle the challenges of the *few-shot zero-glance* setting, we propose a novel unlearning framework that removes class-specific knowledge without accessing the forget set. We first train a Generative Feedback Network (GFN) on a small subset of retained data to generate Optimal Erasure Samples (OES), which are synthetic instances labelled as the target class. These samples are crafted to interfere with forgotten-class knowledge while preserving the decision boundaries of retained classes. The GFN is trained with a stabilised joint objective that promotes forgetting through gradient ascent on target-class predictions and encourages retention through gradient descent on preserved-class outputs. To further enhance forgetting, we introduce a two-phase fine-tuning procedure: the first phase uses a large learning rate on both OES and retained data to overwrite class-specific representations, while the second phase uses a smaller learning rate and only retained data to refine decision boundaries and recover utility.

Our main contributions are as follows:

- We introduce GFOES, a practical framework for machine unlearning in the *few-shot zero-glance* setting, enabling class-specific forgetting with no access to the forget set and minimal use of retained data.
- Our framework leverages synthetic OES and a two-phase fine-tuning procedure to balance effective forgetting and utility preservation, even under severe data constraints.
- Extensive experiments on CIFAR-10, CIFAR-100, and Fashion-MNIST demonstrate that GFOES consistently outperforms state-of-the-art methods in both forgetting effectiveness and model retention quality.

2 Related Work

Exact Unlearning Exact unlearning aims to reproduce retraining results after data removal with lower cost. The SISA framework (Bourtoule et al. 2019) enables selective retraining via data partitioning, inspiring methods like DaRE (Brophy and Lowd 2021), GraphEraser (Chen et al. 2022), and federated unlearning (Su and Li 2023). Other variants use memory augmentation (Yan et al. 2022), feature scores (Cao and Yang 2015), or Hessian guidance (Liu et al. 2022b).Despite their theoretical appeal, exact unlearning methods remain impractical in real-world scenarios due to their computational and memory overhead.

Approximate Unlearning To address the limitations of exact unlearning, approximate methods aim to remove the influence of target data without replicating full retraining. Gradient-based approaches include gradient ascent to erase backdoors (Liu et al. 2022a) and selective parameter tuning (Fan et al. 2024). Influence function-based methods, such as Certified Removal (Guo et al. 2020) and its scalable extensions (Tanno et al. 2022; Suriyakumar and Wilson 2022), estimate per-sample contributions. Hybrid strategies incorporate linear approximations (Izzo et al. 2021), Fisher information masking (Golatkar et al. 2021), and UNSIR (Tarun et al. 2023), which introduced the zero-glance setting. Despite their flexibility, these methods face two key challenges: (1) utility degradation due to catastrophic forgetting, and (2) reliance on auxiliary datasets to maintain accuracy (Parisi et al. 2019; Chundawat et al. 2023a). While UNSIR operates without access to the forget set, our analysis indicates its limited effectiveness in removing representation-level knowledge under strict data constraints.

Few-shot Unlearning As a subset of approximate unlearning, few-shot unlearning removes the influence of target data using only a small portion of the original training set, reducing storage costs and reflecting realistic post-deletion scenarios (Yoon et al. 2024). Representative methods include model inversion for proxy data reconstruction (Yoon et al. 2024) and manifold mixup for vertical federated learning (Gu et al. 2024). A special case is zero-shot unlearning (Chundawat et al. 2023b; Zhang et al. 2025), which relies on knowledge distillation to unlearn without any original data. While zero-shot unlearning addresses a stricter challenging setting, it faces two key issues in practical few-shot scenarios: (1) lack of source data causes utility degradation even when some data remain; and (2) distillation often incurs high computational overhead, limiting scalability to large models and datasets.

In practice, MLaaS providers often retain a small portion of the original dataset, which, if effectively utilised, can significantly improve utility preservation during unlearning. These insights underscore the need for a framework that operates in the *few-shot zero-glance* setting, achieving complete removal of class-specific knowledge while efficiently preserving utility with minimal retained data.

3 Methodology

3.1 Problem Formulation

We formalise the task of machine unlearning in the *few-shot zero-glance* setting as follows: Consider a K-class classification task defined on a labelled dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$, where $x_i \in \mathcal{X} \subseteq \mathbb{R}^d$ represents the d-dimensional input features, and $y_i \in \mathcal{Y} = \{1, 2, \ldots, K\}$ denotes the

class labels. Let $M_0 = f(\cdot;\theta_0)$ be a predictive model trained on $\mathcal{D}$ by minimising a standard classification loss function $\mathcal{L}$. The optimisation objective is defined as: $\theta_0 = \text{argmin}_\theta \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(y_i, f(x_i;\theta))$.

Given a subset of categories to be forgotten, denoted as $\mathcal{Y}_f \subset \mathcal{Y}$, we partition the original dataset $\mathcal{D}$ into two disjoint subsets. The target dataset is defined as $\mathcal{D}_f = \{(x_i, y_i) \mid y_i \in \mathcal{Y}_f\}$, which contains all instances belonging to the categories selected for forgetting. The retained dataset is defined as $\mathcal{D}_r = \mathcal{D} \setminus \mathcal{D}_f$, which includes all remaining instances associated with the categories to be preserved.

In the ideal case where full access to $\mathcal{D}_f$ and $\mathcal{D}_r$ is available, the unlearning task can be modelled as a multi-objective optimisation:

$$\theta^* = \arg\min_\theta \left[-\lambda \cdot \mathcal{L}(\mathcal{D}_f;\theta) + (1-\lambda)\cdot\mathcal{L}(\mathcal{D}_r;\theta)\right], \quad (1)$$

where $\lambda \in (0,1)$ controls the trade-off between forgetting the target data and preserving the retained knowledge.

However, in the *few-shot zero-glance* setting, two critical constraints apply:

- **Zero-glance constraint:** The dataset $\mathcal{D}_f$ targeted for removal is completely inaccessible due to privacy or regulatory requirements. As a result, $\mathcal{D}_f$ cannot be directly used in the unlearning process.
- **Few-shot constraint:** Only a small subset of the retained dataset, denoted as $\mathcal{D}_{rs} \subset \mathcal{D}_r$, is available for model adjustment, where typically $|\mathcal{D}_{rs}| \ll |\mathcal{D}_r|$.

Consequently, direct optimisation of Eq. 1 becomes infeasible under the few-shot zero-glance constraints. To address this, we substitute $\mathcal{D}_f$ with a set of synthetic samples $\tilde{\mathcal{D}} = \{(\tilde{x}_i, \tilde{y}_i)\}$, where each $\tilde{y}_i \in \mathcal{Y}_f$, and reformulate the unlearning objective as:

$$\theta^* = \arg\min_\theta \left[-\lambda \cdot \mathcal{L}(\mathcal{Y}_f; f(\tilde{x};\theta)) + (1-\lambda)\cdot\mathcal{L}(\mathcal{D}_{rs};\theta)\right],$$

where $\tilde{x}$ denotes synthetic inputs representing the categories in $\mathcal{Y}_f$, and $\mathcal{D}_{rs} \subset \mathcal{D}_r$ is the few-shot retained subset available for model preservation.

Formally, machine unlearning under the *few-shot zero-glance* setting aims to satisfy the following two objectives:

- **Effective Unlearning:** The model should exhibit significantly degraded predictive accuracy on the forgotten classes. This is expressed as:

  $$\mathbb{E}_{(x,y)\sim\mathcal{D}_f}[\mathcal{L}(y, f(x;\theta^*))] \gg \mathbb{E}_{(x,y)\sim\mathcal{D}_f}[\mathcal{L}(y, f(x;\theta_0))],$$

  where θ_0 and θ^* denote the model parameters before and after unlearning, respectively.
- **Performance Preservation:** The model should maintain predictive performance on the retained dataset:

  $$\mathbb{E}_{(x,y)\sim\mathcal{D}_r}[\mathcal{L}(y, f(x;\theta^*))] \approx \mathbb{E}_{(x,y)\sim\mathcal{D}_r}[\mathcal{L}(y, f(x;\theta_0))].$$

This formulation encapsulates the core challenge of *few-shot zero-glance* setting: constructing synthetic data that effectively induces forgetting while preserving model utility, all under strict data access constraints

3.2 Optimal Erasure Samples

Synthesising substitute samples $\tilde{x}$ under the zero-glance constraint, where only the label set $\mathcal{Y}_f$ of the forgotten classes is available, is inherently difficult because label information alone provides no access to the associated feature distributions. To address this, we adopt a reverse design strategy. Instead of approximating the true data distribution, we deliberately generate synthetic samples that deviate from it. These adversarial-like inputs, paired with the target labels, mislead the model into associating irrelevant features with the forgotten classes. As a result, the decision boundaries of the forgotten classes are disrupted with minimal interference to the retained ones.

We refer to these synthetic samples as Optimal Erasure Samples (OES). Fine-tuning the model on OES weakens its ability to recognise the true semantics of classes in $\mathcal{Y}_f$, thereby inducing effective class-level forgetting. Specifically, each sample from a class $y \in \mathcal{Y}_f$ is replaced with an OES labelled as y, while the original data for classes in $\mathcal{Y}_r$ is preserved. This approach reduces the original multi-objective unlearning formulation (Eq. 1) to a single fine-tuning objective as follows:

$$\theta^* = \arg\min_\theta \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}\left(f(x_i';\theta), y_i\right),$$
$$\text{where} \quad x_i' = \begin{cases} x^*, & \text{if } y_i \in \mathcal{Y}_f, \\ x_i, & \text{otherwise.} \end{cases} \quad (2)$$

To construct effective OES under few-shot constraints, two key conditions must be met: (1) when labelled with $\mathcal{Y}_f$, OES should yield high loss under the original model $f(\cdot;\theta_0)$ and low loss under the unlearned model $f(\cdot;\theta^*)$, disrupting the model's prior knowledge of the forgotten classes; (2) fine-tuning on OES should preserve accuracy on the retained set, keeping its loss low and comparable to pre-unlearning performance.

We now present a formal definition for OES that unifies these two constraints:

Definition 1 (Optimal Erasure Samples (OES))**.** *Let $\mathcal{D}^* = \{(x_i^*, y_i^*)\}_{i=1}^{s}$ be a synthetic dataset, where each $y_i^* \in \mathcal{Y}_f$. This dataset qualifies as OES if it minimises the following composite objective:*

$$\mathcal{D}^* = \arg\min_{\mathcal{D}'}\big[-\lambda\cdot\mathbb{E}_{(x,y)\in\mathcal{D}'}\mathcal{L}(y, f(x;\theta_0)) + (1-\lambda)\cdot\mathbb{E}_{(x,y)\in\mathcal{D}_{rs}}\mathcal{L}(y, f(x;\theta^*))\big], \quad (3)$$

where θ^* is a one-epoch updated model defined as:

$$\theta^* = \theta_0 - \eta\cdot\nabla_\theta\mathbb{E}_{(x,y)\in\mathcal{D}'\cup\mathcal{D}_{rs}}\mathcal{L}(y, f(x;\theta))|_{\theta=\theta_0}. \quad (4)$$

Here, $\lambda \in (0,1)$ controls the trade-off between forgetting and retention of model utility. The first term encourages high loss on the forgotten classes to impair their influence, while the second term ensures that performance on the retained data is preserved following fine-tuning. The update step for θ^* simulates the model's adaptation to the synthetic OES without requiring access to the actual data targeted for removal. This unified formulation captures both the destructive effect on the target classes and the stability requirement for the retained classes.

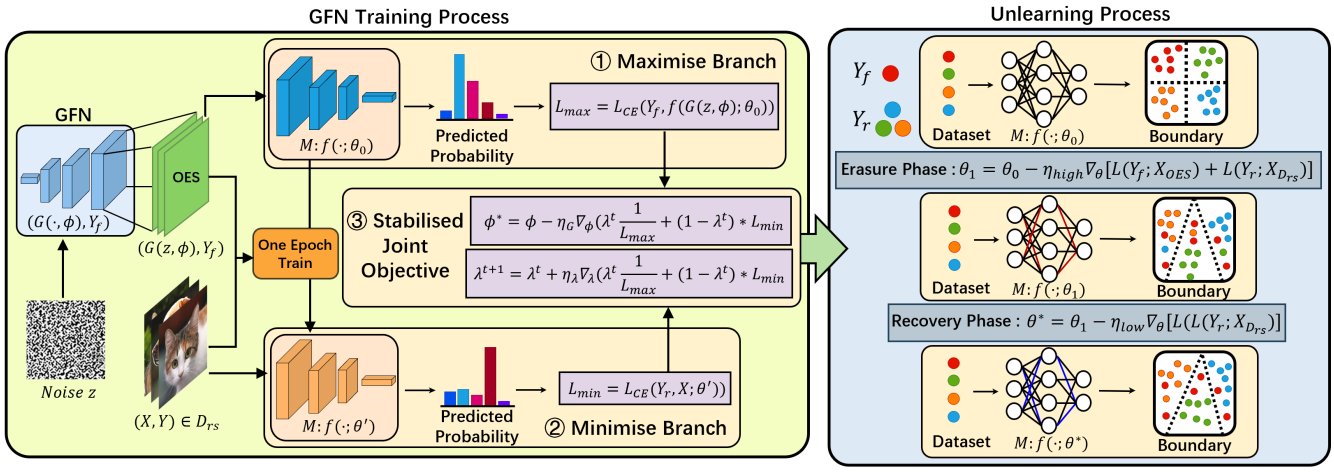

Figure 2: The training process of GFN consists of three components: the *Maximise Branch*, the *Minimise Branch*, and the *Stabilised Joint Objective*, as illustrated on the left. The right side shows the machine unlearning procedure applied to the generated OES.

3.3 Generative Feedback Network

Constructing OES by directly solving the optimisation in Eq. (3) is computationally intractable. To overcome this challenge, we introduce a Generative Feedback Network (GFN), which transforms the OES search into a tractable, learnable optimisation process. As shown on the left side of Figure 2, the GFN decomposes the OES construction into two complementary sub-tasks: (1) maximising the classification loss on forgotten classes to induce forgetting, and (2) minimising the loss on retained classes to preserve utility. To achieve this, the architecture comprises three components: the *Maximise* branch, the *Minimise* branch, and a *Stabilised Joint Objective* module that adaptively balances the two competing goals. These components jointly optimise a generator $G(\cdot,\phi)$, initialised from random noise z, to produce high-quality OES that meet both the forgetting and retention criteria in a stable and adaptive manner.

Maximise Branch This component addresses constraint by encouraging the generator to produce synthetic samples that incur high classification loss under the original model $f(\cdot;\theta_0)$. Formally, it optimises the following objective:

$$\mathcal{L}_{\max}=\mathcal{L}_{\text{CE}}(\mathcal{Y}_f,f(G(z,\phi);\theta_0)), \tag{5}$$

where $G(z,\phi)$ denotes the synthetic sample generated from noise z via generator parameters ϕ, and $\mathcal{L}_{\text{CE}}$ is the cross-entropy loss computed with respect to the target labels $\mathcal{Y}_f$.

Minimise Branch To preserve knowledge of the retained classes, we perform a one-epoch fine-tuning of the original model M_0 using both the generated samples and the few-shot retained dataset $\mathcal{D}_{rs}$. The model parameters are updated as follows:

$$\theta'=\theta_0-\eta_{\text{GFN}}\nabla_\theta\mathcal{L}(\mathcal{Y}_f\cup\mathcal{Y}_r,f(G(z,\phi)\cup\mathcal{D}_{rs};\theta_0)), \tag{6}$$

where η_{GFN} is the learning rate, and $x_r\in\mathcal{D}_{rs}$. The updated model $f(\cdot;\theta')$ is then evaluated on the retained classes to compute:

$$\mathcal{L}_{\min}=\mathcal{L}_{\text{CE}}(\mathcal{Y}_r,f(x_r;\theta')). \tag{7}$$

Stabilised Joint Objective The final training objective of GFN stabilises the conventional bi-objective loss by replacing $\mathcal{L}_{\max}$ with its reciprocal:

$$\mathcal{L}_{\text{GFN}}=\lambda\cdot\frac{1}{\mathcal{L}_{\max}}+(1-\lambda)\cdot\mathcal{L}_{\min},\quad\lambda\in(0,1). \tag{8}$$

This formulation mitigates the influence of large values in $\mathcal{L}_{\max}$ while retaining its gradient direction, as shown by its gradient: $\nabla_\phi\left(\frac{1}{\mathcal{L}_{\max}}\right)=-\frac{1}{\mathcal{L}_{\max}^2}\cdot\nabla_\phi\mathcal{L}_{\max}$.

To ensure both effective forgetting and utility preservation, we define a stabilised joint loss function for the generator:

$$\mathcal{L}_{\text{GFN}}^{(t)}=\lambda_t\cdot\frac{1}{\mathcal{L}_{\max}}+(1-\lambda_t)\cdot\mathcal{L}_{\min},\quad\lambda_t\in(0,1), \tag{9}$$

where the reciprocal form of $\mathcal{L}_{\max}$ mitigates instability arising from unbounded gradients. The coefficient λ_t governs the trade-off between the forgetting and retention objectives at iteration t, and is dynamically adjusted rather than manually selected.

To adaptively balance the competing objectives during training, λ_t is updated via gradient ascent on $\mathcal{L}_{\text{GFN}}^{(t)}$:

$$\lambda_{t+1}=\Pi_{(0,1)}(\lambda_t+\eta\cdot\frac{\partial\mathcal{L}_{\text{GFN}}^{(t)}}{\partial\lambda}), \tag{10}$$

$$\frac{\partial\mathcal{L}_{\text{GFN}}^{(t)}}{\partial\lambda}=\frac{1}{\mathcal{L}_{\max}}-\mathcal{L}_{\min}, \tag{11}$$

where $\eta>0$ denotes the learning rate, and $\Pi_{(0,1)}$ is a projection operator ensuring that λ_t remains within the open interval $(0,1)$. This mechanism obviates the need for manual hyperparameter tuning.

Under standard boundedness and differentiability assumptions, the update rule converges to a fixed point λ^* where $\frac{1}{\mathcal{L}_{\max}(\lambda^*)}=\mathcal{L}_{\min}(\lambda^*)$, indicating a stable equilibrium between objectives. Formal analysis is provided in Appendix.

3.4 Two-Stage Fine-Tuning for Unlearning

Given the synthetic dataset $\mathcal{D}^*$ composed of OES, we adopt a two-phase fine-tuning procedure as shown on the right side of Figure 2, to balance effective forgetting and retention of useful knowledge. This strategy consists of two sequential phases: an *Erasure Phase* that applies a large learning rate to jointly fine-tune the model on $\mathcal{D}^*$ and few-shot retained data $\mathcal{D}_{rs}$, followed by a *Recovery Phase* that uses only $\mathcal{D}_{rs}$ with a smaller learning rate to restore the decision boundaries of preserved classes.

The strategy is motivated by the observation that aggressive updates are needed to disrupt representations of forgotten classes, yet they risk damaging the structure of retained ones. The second stage serves to repair this collateral damage without reintroducing the erased knowledge. While not grounded in formal theory, this design is empirically validated in Section 4.5, which shows its clear advantage over single-phase or uniformly trained alternatives.

Erasure Phase This phase aggressively fine-tunes the model on both the synthetic OES targeting $\mathcal{Y}_f$ and the few-shot retained subset $\mathcal{D}_{rs}$, using a high learning rate η_{high}. The OES, generated adversarially, are designed to disrupt the model's internal representation of the forgotten classes. Meanwhile, the inclusion of $\mathcal{D}_{rs}$ helps to stabilise optimisation and reduce collateral degradation.

The loss function is given by:

$$\mathcal{L}_{\text{erase}} = \mathcal{L}(\mathcal{Y}_f; f(\tilde{x}; \theta_0)) + \mathcal{L}(\mathcal{D}_{rs}; \theta_0), \tag{12}$$

where $\tilde{x} = G(z, \phi^*)$ denotes synthetic inputs produced by the trained generator.

The model is then updated as:

$$\theta_1 = \theta_0 - \eta_{\text{high}} \cdot \nabla_\theta \mathcal{L}_{\text{erase}}. \tag{13}$$

Recovery Phase To mitigate any adverse impact on the retained classes introduced during the erasure phase, we perform a second fine-tuning stage using only the retained subset $\mathcal{D}_{rs}$ and a lower learning rate η_{low}. As the forgotten classes $\mathcal{Y}_f$ are excluded at this point, this stage serves to refine the model's decision boundaries for $\mathcal{Y}_r$ without the risk of relearning discarded knowledge.

The model parameters are updated as:

$$\theta^* = \theta_1 - \eta_{\text{low}} \cdot \nabla_\theta \mathcal{L}(\mathcal{D}_{rs}; \theta_1). \tag{14}$$

A detailed description of the procedure is presented in Appendix.

4 Experiment

We conduct comprehensive experiments across multiple datasets, model architectures, and retained data ratios to evaluate the effectiveness of our method. Our validation assesses both the completeness of forgetting and the preservation of model utility using logit-based and representation-based metrics. We further perform ablation studies to examine the contribution of each key component. Due to space limitations, detailed configurations and experimental results are provided in Appendix. The code is available in the Supplementary Material.

4.1 Experimental Setup

Datasets and Model Architectures We evaluate our framework on three image classification datasets: Fashion-MNIST (Xiao, Rasul, and Vollgraf 2017), CIFAR-10, and CIFAR-100 (Krizhevsky, Hinton et al. 2009). In each setting, a subset of classes is randomly selected as the target for forgetting, and only 5%, 10%, or 20% of samples per class are retained from the original training set.

We adopt three representative models: (1) All-CNN (Springenberg et al. 2014), a lightweight convolutional network used for Fashion-MNIST; (2) ResNet-18 (He et al. 2016), a residual network with skip connections, applied to CIFAR-10; and (3) ResNet-50, a deeper residual architecture used for CIFAR-100.

Comparison Methods As no existing method is explicitly designed for the *few-shot zero-glance* setting, we evaluate a range of representative machine unlearning methods under this constraint. We include the following methods: (1) **Retrain**, a gold-standard baseline that trains a new model from scratch on the retained data only; (2) **NegGrad** (Thudi et al. 2022), which performs gradient ascent to increase the loss on forgotten data; (3) **RL** (Golatkar, Achille, and Soatto 2020), which assigns random labels to target samples for destructive fine-tuning; (4) **Fisher** (Golatkar, Achille, and Soatto 2020), which perturbs parameters based on their Fisher importance; (5) **UNSIR** (Tarun et al. 2023), a zero-glance method that injects adversarial noise to erase target representations; (6) **MI** (Yoon et al. 2024), which reconstructs proxy data via model inversion for fine-tuning; (7) **GKT** (Chundawat et al. 2023b), a knowledge distillation-based method with thresholding to block target class transfer; (8) **SalUn** (Fan et al. 2024), which updates only saliency-sensitive parameters to forget specified data.

Specifically, Retrain, UNSIR, and GKT are naturally compatible with *few-shot zero-glance* setting as they do not rely on access to the forgotten data. For the remaining methods that typically require access to the target data, we substitute the missing data with random noise inputs.

4.2 Logit-based Evaluation

Logit-based evaluation is a widely adopted protocol in machine unlearning (Kim, Cha, and Kim 2025), facilitating an intuitive assessment of unlearning effectiveness. We employ two complementary metrics. The first is the *accuracy on the forgotten dataset* ($A\mathcal{D}_f$), which quantifies the model's accuracy on data that ought to be forgotten; ideally, this value should approach zero, indicating successful erasure of target class knowledge. The second is the *accuracy on the retained dataset* ($A\mathcal{D}_r$), which measures the model's accuracy on data that should be preserved. A higher value reflects better retention of the model's original utility.

As shown in Table 1, GFOES consistently achieves $\mathcal{AD}_f = 0$ across all datasets and settings, indicating complete forgetting of the target classes. Comparable results are observed with Retrain, GKT, and UNSIR, which also attain zero forgetting error. However, only GFOES and UNSIR accomplish this without directly accessing the forget set, thereby satisfying the *zero-glance* constraint. In contrast,

		Fashion-MNIST						CIFAR-10						CIFAR-100					
		$\#y=1$			$\#y=4$			$\#y=1$			$\#y=4$			$\#y=1$			$\#y=10$		
		5%	10%	20%	5%	10%	20%	5%	10%	20%	5%	10%	20%	5%	10%	20%	5%	10%	20%
$\mathcal{AD}_f$ (%) ↓	*Original*	86.80	86.80	86.80	92.03	92.03	92.03	85.82	85.82	85.82	87.15	87.15	87.15	72.94	72.94	72.94	71.38	71.38	71.38
	Retrain	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
	NegGrad	36.87	27.09	18.25	39.39	23.68	20.36	32.14	23.91	11.23	33.72	21.19	13.55	36.87	27.09	18.25	39.39	23.68	20.36
	RL	33.80	26.89	16.27	44.31	28.76	23.58	35.32	25.18	18.01	42.94	30.21	21.76	39.17	32.90	24.64	40.88	32.36	26.93
	Fisher	39.41	33.82	27.65	41.29	35.37	30.08	40.73	32.15	26.09	42.87	34.01	28.46	32.41	23.37	17.02	33.46	26.78	19.12
	UNSIR	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
	MI	25.47	13.85	7.12	27.33	13.92	6.41	27.51	12.09	6.02	27.82	13.45	9.55	29.81	12.89	8.92	26.12	14.85	10.05
	GKT	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
	SalUn	25.81	13.66	6.92	23.44	16.39	4.27	27.92	11.41	8.36	21.13	17.82	2.94	29.31	13.28	9.67	22.76	15.54	5.11
	GFOES	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
$\mathcal{AD}_r$ (%) ↑	*Original*	92.21	92.21	92.21	91.42	91.42	91.42	90.23	90.23	90.23	91.56	91.56	91.56	71.47	71.47	71.47	71.49	71.49	71.49
	Retrain	51.50	57.92	66.37	50.49	58.20	64.03	47.21	58.43	69.38	52.92	59.22	68.06	46.44	49.07	53.50	47.23	50.95	54.24
	NegGrad	61.82	64.55	67.44	61.80	62.43	64.84	59.12	63.04	64.99	59.73	61.85	64.89	48.27	50.56	52.71	47.63	49.03	50.98
	RL	78.35	81.02	83.67	79.03	80.90	83.45	72.45	73.88	75.91	71.55	74.03	77.04	58.01	60.15	62.76	57.41	59.00	61.94
	Fisher	61.45	67.89	73.21	60.56	65.78	71.12	62.34	63.45	69.56	61.23	66.78	71.23	47.89	54.56	59.34	49.01	56.78	59.43
	UNSIR	86.29	87.98	89.33	86.38	87.92	88.03	84.00	84.19	86.94	85.53	86.21	87.67	64.21	65.84	67.22	63.53	65.31	67.78
	MI	78.09	81.12	82.33	79.72	83.92	84.03	77.23	80.15	81.47	78.79	82.84	83.11	62.35	65.21	67.49	63.84	68.12	69.23
	GKT	87.41	87.41	87.41	86.91	86.91	86.91	55.23	55.23	55.23	48.76	48.76	48.76	46.17	46.17	46.17	30.29	30.29	30.29
	SalUn	71.44	72.18	76.21	70.19	74.66	80.17	70.12	73.65	77.43	68.41	73.92	79.18	54.73	58.16	62.29	52.84	57.65	63.49
	GFOES	**88.03**	**89.12**	**90.98**	**89.51**	**89.85**	**90.75**	**86.35**	**87.93**	**88.97**	**86.52**	**88.66**	**89.23**	**66.34**	**67.91**	**69.95**	**67.00**	**68.06**	**69.47**

Table 1: Comprehensive results on three datasets across nine comparison methods and the proposed GFOES. ↑ indicates that higher values are better, while ↓ indicates that lower values are preferable. The *Original* presents model performance prior to unlearning. $\#y$ denotes the number of classes to be unlearned, and e represents the percentage of accessible training data. The best results are shown in **bold**.

methods such as NegGrad, RL, Fisher, MI, and SalUn rely on real data to compute gradients or saliency maps, and consequently struggle under the few-shot zero-glance setting. These methods exhibit substantial residual influence; for example, Fisher reaches 39.41% on Fashion-MNIST with $\#y = 1$ and $e = 5\%$. Notably, SalUn and MI further degrade when e is small, highlighting their sensitivity to noisy input and dependence on sufficient retained data.

In terms of retained utility, GFOES consistently achieves the highest $\mathcal{AD}_r$ across all datasets and settings. For example, on CIFAR-10 with $\#y = 4$ and $e = 20\%$, GFOES retains 89.23% accuracy, surpassing RL (77.04%), MI (83.11%), and UNSIR (87.67%). This strong performance is driven by two key components: an OES generator aligned with the retained class distribution and a two-phase fine-tuning strategy that separates forgetting from recovery. In contrast, Retrain underfits when data access is limited (e.g., 50.49% on CIFAR-10, $\#y = 4$, $e = 5\%$), while GKT fails to generalise on complex datasets (e.g., 30.29% on CIFAR-100, $\#y = 10$, $e = 20\%$). Other baselines show further drops due to their inability to cope with noisy or insufficient forget samples.

4.3 Representation-based Evaluation

While logit-based evaluation is widely used to assess unlearning, it primarily reflects the model's reduced ability to classify forgotten classes and may not indicate whether internal representations related to those classes have been removed. Prior work (Kim, Cha, and Kim 2025) shows

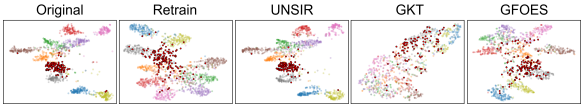

Figure 3: t-SNE visualisation of feature representations for single-class unlearning on CIFAR-10. Red points indicate samples from the forgotten class, while other colours represent the retained classes.

that when the feature extractor retains relevant knowledge, retraining the final layer can restore classification performance. To address this limitation, we include a representation-based evaluation that examines changes in the internal feature space of the model.

To evaluate whether unlearning disrupts class-specific representations, we apply t-SNE (Van der Maaten and Hinton 2008) to visualise feature extractor outputs on CIFAR-10 under the single-class setting. As shown in Figure 3, the original model forms well-separated clusters, while UNSIR preserves the forgotten class cluster, indicating incomplete forgetting. GKT scatters the forgotten class but also distorts retained ones, potentially harming utility. In contrast, our method disperses the forgotten class while maintaining the structure of retained classes, achieving targeted forgetting without compromising performance. As shown in Figure 4, GradCAM (Selvaraju et al. 2017) further reveals that our framework, like Retrain, shifts attention away from seman-

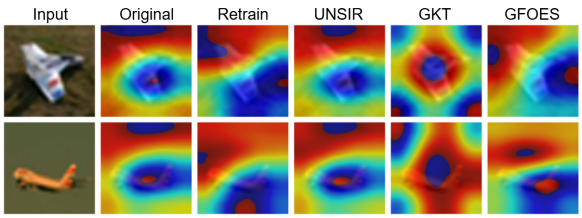

Figure 4: GradCAM visualisation for single-class unlearning on CIFAR-10. Warmer colours highlight regions the model attends to when predicting the forgotten class, while cooler colours indicate lower attention.

	Method	Fashion-MNIST		CIFAR-10		CIFAR-100	
		$\#y = 1$	$\#y = 4$	$\#y = 1$	$\#y = 4$	$\#y = 1$	$\#y = 10$
Time (s) ↓	Retrain	1457.63	1348.66	10816.30	8798.16	16198.26	14974.74
	GKT	5420.43	4801.20	14760.69	12932.40	23623.65	21294.82
	UNSIR	15.52	13.46	26.68	23.69	42.82	38.42
	GFOES	277.53	231.96	619.73	517.18	1165.78	1007.50
	GFN	263.42	218.05	597.56	498.80	1125.68	971.08
	Two-Stage	14.11	13.91	22.17	18.38	40.10	36.42

Table 2: Wall-clock time (seconds) for unlearning methods. GFN denotes the time for generating Optimal Erasure Samples, Two-Stage for the fine-tuning phase, and GFOES for the total (GFN + Two-Stage).

tically meaningful regions, whereas UNSIR retains focus on them. This confirms that GFOES erases both prediction and representation-level dependencies of the forgotten class.

4.4 Time Efficiency Analysis

Efficiency is essential for practical unlearning. To assess computational overhead, we report the wall-clock time of the unlearning process for Retrain, GKT, UNSIR, and GFOES across datasets in Table 2, as only these methods perform well in the *few-shot zero-glance* setting. For GFOES, we report the time separately for GFN training, two-phase fine-tuning, and the overall total. As shown in Table 2, GFOES consistently incurs substantially lower time costs across all datasets and scenarios compared to Retrain and GKT, while maintaining high-quality unlearning. Although GKT supports zero-shot unlearning, its reliance on knowledge distillation leads to the highest overhead. UNSIR is the fastest, but it fails to remove feature representations (see Sec. 4.3), limiting its effectiveness. Overall, GFOES achieves an optimal balance between efficiency and unlearning quality, making it suitable for real-world deployment. Additional time efficiency analyses are provided in Appendix.

4.5 Ablation Study

To evaluate the contributions of core components in the proposed GFOES, we conduct an ablation study focusing on (i) the effectiveness of OES generation and (ii) the role of the two-phase fine-tuning procedure. For data composition,

Setting	Fashion-MNIST		CIFAR-10		CIFAR-100	
	$\mathcal{AD}_f \downarrow$	$\mathcal{AD}_r \uparrow$	$\mathcal{AD}_f \downarrow$	$\mathcal{AD}_r \uparrow$	$\mathcal{AD}_f \downarrow$	$\mathcal{AD}_r \uparrow$
GFOES	**0.00**	89.45	**0.00**	87.54	**0.00**	68.72
OES+D_r+R_l	**0.00**	62.32	**0.00**	58.66	**0.00**	39.12
OES+D_r+R_s	37.29	89.85	33.15	88.19	30.24	**70.54**
OES+R_{ls}	**0.00**	62.16	**0.00**	60.81	**0.00**	40.94
OES+R_l	**0.00**	48.36	**0.00**	45.20	**0.00**	27.26
OES+R_s	34.29	84.94	30.98	80.53	26.23	62.06
D_r+R_{ls}	15.19	79.88	11.30	77.16	12.67	58.93
D_r+R_l	4.30	68.36	5.68	65.01	6.37	46.03
D_r+R_s	42.74	**92.97**	38.70	**88.82**	34.99	70.05

Table 3: Ablation results for single-class unlearning across three datasets. Each setting combines data composition (OES or D_r) and learning rate strategies (R_{ls}, R_l, or R_s). The best results are shown in **bold**.

we compare OES, which uses only OES-generated samples during erasure and 10% retained data in recovery, with D_r, which uses 10% retained data in both phases. For two-stage fine-tuning procedure, we consider three strategies: R_{ls} (0.004 in erasure, 0.004 in recovery), R_l (0.004 in both), and R_s (0.0004 in both). Combining these options yields six configurations, which we evaluate on Fashion-MNIST, CIFAR-10, and CIFAR-100 under single-class unlearning setting to assess the impact of each component.

The ablation results in Table 3 confirm the complementary roles of OES and the two-phase fine-tuning procedure. Configurations lacking OES (e.g., Dr+Rl, Dr+Rs) consistently fail to achieve full forgetting, indicating that retained data alone is insufficient to erase target class knowledge. Conversely, using OES without proper learning rate control (e.g., OES+Rl or OES+Rs) leads to utility degradation or incomplete forgetting. Only the full GFOES configuration (OES+Dr+Rls) consistently achieves optimal performance, balancing complete forgetting ($\mathcal{AD}_f = 0$) with minimal utility loss across all datasets. This unequivocally demonstrates that both synthetic erasure signals and staged fine-tuning are essential for effective and stable unlearning.

5 Conclusion

In this work, we propose GFOES, a novel machine unlearning framework tailored for the *few-shot zero-glance* setting. GFOES introduces Optimal Erasure Samples as targeted adversarial signals to guide forgetting, combined with a two-phase fine-tuning procedure to remove class-specific knowledge while preserving performance on retained classes. Extensive experiments across diverse datasets and architectures demonstrate that GFOES achieves complete forgetting at both the logit and representation levels, while consistently outperforming existing baselines in utility retention under strict constraints. Ablation studies further confirm the importance of both the OES component and the two-phase fine-tuning procedure. Overall, GFOES provides a practical and scalable solution for data deletion compliance in real-world scenarios where direct access to the forget set is unavailable.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant U24A20240, the Fundamental Research Funds for the Central Universities (ZYTS25071), and the '111 Center' (B16037). Ziqi Xu is supported by the research support package from the School of Computing Technologies at RMIT University.

References

Bourtoule, L.; Chandrasekaran, V.; Choquette-Choo, C. A.; Jia, H.; Travers, A.; Zhang, B.; Lie, D.; and Papernot, N. 2019. Machine Unlearning. In *NeurIPS*.

Brophy, J.; and Lowd, D. 2021. Machine unlearning for random forests. In *International Conference on Machine Learning*, 1092–1104. PMLR.

Calzada, I. 2022. Citizens' data privacy in china: The state of the art of the personal information protection law (pipl). *Smart Cities*, 5(3): 1129–1150.

Cao, Y.; and Yang, J. 2015. Towards making systems forget with machine unlearning. In *2015 IEEE symposium on security and privacy*, 463–480. IEEE.

Chen, M.; Zhang, Z.; Wang, T.; Backes, M.; Humbert, M.; and Zhang, Y. 2022. Graph unlearning. In *Proceedings of the 2022 ACM SIGSAC conference on computer and communications security*, 499–513.

Chundawat, V. S.; Tarun, A. K.; Mandal, M.; and Kankanhalli, M. 2023a. Can bad teaching induce forgetting? Unlearning in deep networks using an incompetent teacher. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 7210–7217.

Chundawat, V. S.; Tarun, A. K.; Mandal, M.; and Kankanhalli, M. 2023b. Zero-shot machine unlearning. *IEEE Transactions on Information Forensics and Security*, 18: 2345–2354.

Fan, C.; Liu, J.; Zhang, Y.; Wong, E.; Wei, D.; and Liu, S. 2024. SalUn: Empowering Machine Unlearning via Gradient-based Weight Saliency in Both Image Classification and Generation. In *International Conference on Learning Representations (ICLR)*.

Golatkar, A.; Achille, A.; Ravichandran, A.; Polito, M.; and Soatto, S. 2021. Mixed-privacy forgetting in deep networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 792–801.

Golatkar, A.; Achille, A.; and Soatto, S. 2020. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9304–9312.

Goldman, E. 2020. An introduction to the california consumer privacy act (ccpa). *Santa Clara Univ. Legal Studies Research Paper*.

Gu, H.; Tae, H. X.; Chan, C. S.; and Fan, L. 2024. A few-shot Label Unlearning in Vertical Federated Learning. *arXiv preprint arXiv:2410.10922*.

Guo, C.; Goldstein, T.; Hannun, A. Y.; and van der Maaten, L. 2020. Certified Data Removal from Machine Learning Models. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, 3832–3842.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Izzo, Z.; Smart, M. A.; Chaudhuri, K.; and Zou, J. 2021. Approximate data deletion from machine learning models. In *International Conference on Artificial Intelligence and Statistics*, 2008–2016. PMLR.

Kim, Y.; Cha, S.; and Kim, D. 2025. Are we truly forgetting? a critical re-examination of machine unlearning evaluation protocols. *arXiv preprint arXiv:2503.06991*.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.

Liu, Y.; Fan, M.; Chen, C.; Liu, X.; Ma, Z.; Wang, L.; and Ma, J. 2022a. Backdoor defense with machine unlearning. In *IEEE INFOCOM 2022-IEEE conference on computer communications*, 280–289. IEEE.

Liu, Y.; Xu, L.; Yuan, X.; Wang, C.; and Li, B. 2022b. The right to be forgotten in federated learning: An efficient realization with rapid retraining. In *IEEE INFOCOM 2022-IEEE conference on computer communications*, 1749–1758. IEEE.

Parisi, G. I.; Kemker, R.; Part, J. L.; Kanan, C.; and Wermter, S. 2019. Continual lifelong learning with neural networks: A review. *Neural networks*, 113: 54–71.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, 618–626.

Shaik, T.; Tao, X.; Xie, H.; Li, L.; Zhu, X.; and Li, Q. 2023. Exploring the landscape of machine unlearning: A comprehensive survey and taxonomy. *arXiv preprint arXiv:2305.06360*.

Springenberg, J. T.; Dosovitskiy, A.; Brox, T.; and Riedmiller, M. 2014. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*.

Su, N.; and Li, B. 2023. Asynchronous federated unlearning. In *IEEE INFOCOM 2023-IEEE conference on computer communications*, 1–10. IEEE.

Suriyakumar, V.; and Wilson, A. C. 2022. Algorithms that approximate data removal: New results and limitations. *Advances in Neural Information Processing Systems*, 35: 18892–18903.

Tanno, R.; F Pradier, M.; Nori, A.; and Li, Y. 2022. Repairing neural networks by leaving the right past behind. *Advances in Neural Information Processing Systems*, 35: 13132–13145.

Tarun, A. K.; Chundawat, V. S.; Mandal, M.; and Kankanhalli, M. 2023. Fast yet effective machine unlearning. *IEEE Transactions on Neural Networks and Learning Systems*.

Thudi, A.; Deza, G.; Chandrasekaran, V.; and Papernot, N. 2022. Unrolling sgd: Understanding factors influencing machine unlearning. In *2022 IEEE 7th European Symposium on Security and Privacy (EuroS&P)*, 303–319. IEEE.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Villaronga, E. F.; Kieseberg, P.; and Li, T. 2018. Humans forget, machines remember: Artificial intelligence and the right to be forgotten. *Computer Law & Security Review*, 34(2): 304–313.

Voigt, P.; and Von dem Bussche, A. 2017. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 10(3152676): 10–5555.

Wu, Y.; Dobriban, E.; and Davidson, S. 2020. Deltagrad: Rapid retraining of machine learning models. In *International Conference on Machine Learning*, 10355–10366. PMLR.

Xiao, H.; Rasul, K.; and Vollgraf, R. 2017. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*.

Yan, H.; Li, X.; Guo, Z.; Li, H.; Li, F.; and Lin, X. 2022. ARCANE: An Efficient Architecture for Exact Machine Unlearning. In *IJCAI*, volume 6, 19.

Yoon, Y.; Nam, J.; Yun, H.; Lee, J.; Kim, D.; and Ok, J. 2024. Few-Shot Unlearnings. In *Proceedings of 2024 IEEE Symposium on Security and Privacy (SP)*, 3276–3292.

Zhang, C.; Shen, S.; Chen, W.; and Xu, M. 2025. Toward Efficient Data-Free Unlearning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 21, 22372–22379.