Disentangled Knowledge Forgetting in Machine Unlearning

Yuhang Xia[1,2], **Cheng Zhen**[1,2], **Yirui Wu**[1,2], **Lixin Yuan**[1,2,3], **Wenxiao Zhang**[1,2] and **Jun Liu**[4]
[1]College of Computer Science and Software Engineering, Hohai University
[2]Key Laboratory of Water Big Data Technology of Ministry of Water Resources, Hohai University
[3]State Key Lab. for Novel Software Technology, Nanjing University, P.R. China
[4]School of Computing and Communication, Lancaster University
{xiayuhang, zhencheng,wuyirui,yuanlixin}@hhu.edu.cn, wenxxiao.zhang@gmail.com,
j.liu81@lancaster.ac.uk

Abstract

With the increasing demand of privacy protection, Machine Unlearning (MU) appears to remove private data from an already trained model without retraining from scratch. Most current works suffer from overly unlearning (low fidelity) or incomplete unlearning (low effectiveness). To identify the issues behind, we conduct causal analysis to obtain a resolvable route, i.e., disentangling shared knowledge into attribute-level semantics to remove it as the confounder. We further perform MU loss analysis to reformulate it as balanced form of constraints, thus guaranteeing high fidelity and effectiveness. Based on theoretical analysis, we propose disentangled knowledge forgetting constrained by the reformulated MU loss, which disentangles knowledge with variational auto-encoder and refines knowledge with counterfactual inference. Extensive experimental results demonstrate that our method achieves state-of-the-art performance.

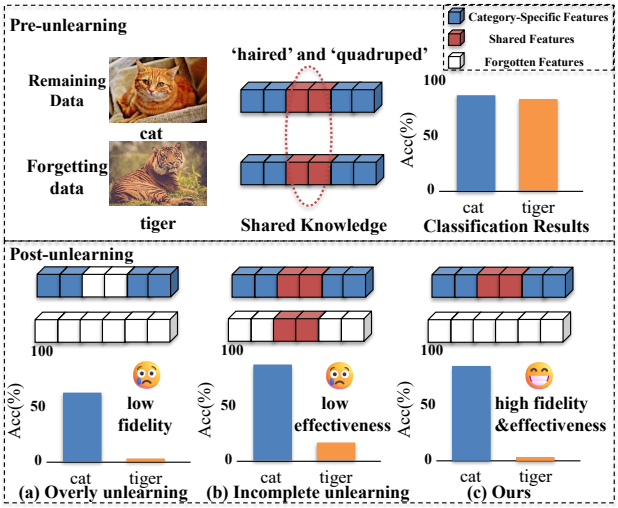


Figure 1: After forgetting, current MU methods predict with overly unlearning (degraded forgetting on cat) and incomplete unlearning (redundant remembering on tiger). Our method obtains balance between forgetting and remembering via knowledge disentangling.

1 Introduction

A growing number of regulations have been introduced to grant users the right to be forgotten. A straightforward yet effective approach is to retrain from scratch, nevertheless suffering from computation overhead. Machine unlearning (MU) thus appears to remove the forgetting data (effectiveness), while preserving model performance on the remaining data (fidelity).

Existing methods address MU with either exact or approximate unlearning. The former [Bourtoule *et al.*, 2021; Graves *et al.*, 2021] retrains with optimized efficiency, still being costly in computation. The latter approximates the retrained model to relieve burden from retraining, where several works [Mehta *et al.*, 2022; Foster *et al.*, 2024] estimate the influence of forgetting data to update parameters, and others adopt knowledge distillation (KD) [Chundawat *et al.*, 2023; Kurmanji *et al.*, 2023] to relearn mis-forgotten knowledge. Despite achieving notable advances, they struggle to simultaneously avoid overly unlearning with low fidelity and incomplete unlearning with low effectiveness (see Fig. 1). After forgetting, model may inadvertently forget knowledge of the remaining category, or revive part of should-forgotten knowledge, that is the shared knowledge between forgetting and remembering categories. For example, the shared knowledge between tiger and cat like 'haired' and 'quadruped' lead to incomplete unlearning on tiger, since the resident feature-to-category mappings could be coincidentally activated by utilizing shared knowledge features.

In Fig.2(a) and (b), we look into causal relationship of overly and incomplete unlearning using Structural Causal Model(SCM) [Pearl *et al.*, 2016] respectively, including shared knowledge K, forgetting data $\mathcal{D}_f$, remaining data $\mathcal{D}_r$, feature representations E, and prediction labels Y. Machine unlearning expects to forget $\mathcal{D}_f$ for risk-free prediction $Y_r = Y - Y_f$. Specifically, $K \to \mathcal{D}_f$ and $K \to \mathcal{D}_r$ represent that $\mathcal{D}_r$ and $\mathcal{D}_f$ share knowledge, $\mathcal{D}_f \to E$ and $\mathcal{D}_r \to E$ represent that E is extracted from $\mathcal{D}_r$ and $\mathcal{D}_f$, $\mathcal{D}_f \to Y$ and $\mathcal{D}_r \to Y$ represent causal effects from images to labels, and $E \to Y$ represents prediction. In overly unlearning, MU removes $K \to \mathcal{D}_r$, inadvertently forgetting shared knowledge of the remaining data to induce model collapse on $\mathcal{D}_r$. In

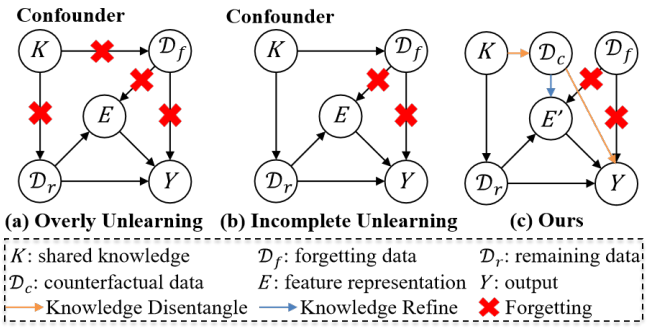


Figure 2: Structural Causal Models of overly unlearning (a), incomplete unlearning (b) and our method (c).

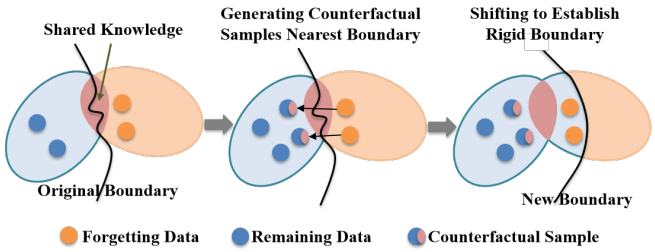


Figure 3: Disentanglement of the shared knowledge to remove it as the confounder, with generating counterfactual samples and shifting to increase rigidness of the decision boundary.

incomplete unlearning, MU maintains $K \rightarrow \mathcal{D}_f$, hindering accurately forgetting knowledge.

Since SCM implies the confusing knowledge in K as the confounder, we delve into approximate unlearning with KD to identify key path for a balanced MU, regarding MU as a two-level classification problem: (1)multi-category prediction for each remembering class; (2)binary prediction for the forgetting class and all the remembering classes. Proved by Sec 3.2, MU should be further regulated with the separated remembering loss (RL) and forgetting loss (FL), which respectively constricts knowledge of fidelity and effectiveness to aid in removing K for balanced MU. RL roughly form boundary with the remembering classes for high fidelity, whereas FL explicitly and rigidly reforms the boundary with the forgetting for high effectiveness. By contrast, traditional MU loss is coupled to emphasize knowledge about fidelity and effectiveness, failing in alleviating overly and incomplete unlearning.

Guided by the separated constraints of RL and FL, K prevents the independent optimization of these losses, requiring attribute-level semantic disentanglement for K to make it fit with their precise optimization. As identifiable semantic units, we argue the disentangled knowledge can be carefully refined via counterfactual data to enforce rigid enough boundary across semantically overlapped samples, thus separately and precisely minimizing RL and FL to guarantee high fidelity and effectiveness. Take Fig.3 to illustrate the disentanglement of the shared knowledge with counterfactual inference, where we first establish decision boundary between the remaining and forgetting data by first generating counterfactual samples near the decision boundary, and then shifting to improve rigidness of boundary by containing all forgetting data. Representing D_c as counterfactual data and E' as the disentangled feature representation, we thus build $K \rightarrow D_c \rightarrow E'\&Y$ to replace $K \rightarrow D_f \rightarrow Y$ in Fig.2(c), effectively revealing causal effects from the disentangled knowledge to prediction in a resolvable form.

Specifically, we first establish $K \rightarrow D_c \rightarrow Y$ via β-VAE to disentangle knowledge as attribute-level semantics. Since counterfactual inference provides the imagination ability to analyze confounds [Niu *et al.*, 2021], we bravely imagine non-existent samples to form D_c based on disentangled attribute-level knowledge, which is represented as $D_c \rightarrow E'$. For example, based on the shared attribute knowledge 'haired' extracted from tiger and cat, we add new concept 'winged' from another known class like eagle to imagine virtual samples for learning, which could locate near the original decision boundary between tiger and cat to promote for the balanced MU. Regarding counterfactual and remembering samples as key anchors, contrastive learning is further adopted to pull positive pairs close and push negative pairs apart, thus gradually shifting boundary with high rigidness. The contributions are:

- **Theoretically:** We offer an insight view in MU, where causal and loss analysis guide route of disentangled knowledge to guarantee high fidelity and effectiveness.
- **Methodologically:** We propose disentangled knowledge forgetting, which benefits from variational autoencoder to disentangle knowledge towards independent optimization of RL and FL, and counterfactual inference to refine knowledge constrained by RL and FL.
- **Empirically:** Extensive experimental results demonstrate our method achieves state-of-the-art performance.

2 Related Work

Exact unlearning accelerates retraining by training submodels [Bourtoule *et al.*, 2021; Yan *et al.*, 2022] or subtracting gradients in training phases [Graves *et al.*, 2021]. They require extensive computational cost, when facing large-scale AI systems or frequent data elimination requests.

Approximate unlearning has emerged as a research focus, focusing on reducing the influence of forgetting data. Early works directly modify model weights via theoretically verified deletion [Golatkar *et al.*, 2020], noisy Newton update [Golatkar *et al.*, 2020], or Fisher information matrix [Foster *et al.*, 2024]. However, they generally suffer from locally slow convergency. Coupled with LLMs and generative model, several optimization-based methods alleviate convergency problem and address advanced challenges like hallucinations [Yuan *et al.*, 2025], correlated knowledge persistence [Wei *et al.*, 2025], and concept revival [George *et al.*, 2025]. With further interest on MU deployment constraints, recent works introduce source-free Hessian estimation [Ahmed *et al.*, 2025], certified noisy fine-tuning [Koloskova *et al.*, 2025], and logits tempering [Spartalis *et al.*, 2025] to efficiently approximate the gold standard of retraining.

3 Preliminaries

3.1 Task Definition

Let $\mathcal{D} = \{x_i, y_i\}_{i=1}^{N}$ be training data containing N examples, where x_i is the ith sample, $y_i \subseteq \{1, ..., K\}$ is the corresponding class label, and K is the total number of classes. We define $\mathcal{D}_f \subseteq \mathcal{D}$ as subset of the training data to forget, and $\mathcal{D}_r = \mathcal{D} \setminus \mathcal{D}_f$ as the remaining data to retain. Defining the original model trained on $\mathcal{D}$ as $h_{\mathbf{w}_0}$ with parameters $\mathbf{w}_0$, we optimize model with cross-entropy loss $\mathcal{L}_{ce}$:

$$\mathbf{w}_0 = \arg\min_{\mathbf{w}} \sum_{(x_i, y_i) \in \mathcal{D}} \mathcal{L}_{ce}(x_i, y_i, \mathbf{w}) \tag{1}$$

Regarded as the optimal unlearning model, $h_{\mathbf{w}^*}$ is the model retrained with $\mathcal{D}_r$. MU aims to derive a new model $h_{\mathbf{w}_u}$ to forget the information in $\mathcal{D}_f$ by updating parameters from $\mathbf{w}_0$ to $\mathbf{w}_u$, ensuring $h_{\mathbf{w}_u}$ to be similar with $h_{\mathbf{w}^*}$.

3.2 Rethinking Machine Unlearning Loss

Inspired by DKD [Zhao *et al.*, 2022], we reformulate unlearning loss with remembering loss (RL) and forgetting loss (FL). For a forgetting sample $x \in \mathcal{D}_f$, denoting classification probability as $p = [p_1, p_2, \ldots, p_f, \ldots, p_K] \in \mathbb{R}^{1\times K}$, where p_i is probability to be classified as ith class and K is number of classes. Each element in p is obtained via softmax function:

$$p_i = \frac{\exp(z_i)}{\sum_{j=1}^{K} \exp(z_j)}, \tag{2}$$

where z_i represents logit of the ith class. To separate the predictions as relevant and irrelevant to the forgetting class, we define binary probabilities $b = [p_f, p_{\setminus f}] \in \mathbb{R}^{1\times 2}$ as

$$p_f = \frac{\exp(z_f)}{\sum_{j=1}^{K} \exp(z_j)}, \quad p_{\setminus f} = \frac{\sum_{k=1, k\neq f}^{K} \exp(z_k)}{\sum_{j=1}^{K} \exp(z_j)}. \tag{3}$$

Meanwhile, we define $\hat{p} = [\hat{p}_1, \ldots, \hat{p}_{f-1}, \hat{p}_{f+1}, \ldots, \hat{p}_K] \in \mathbb{R}^{1\times(K-1)}$ to independently model the probabilities among $K-1$ classes as

$$\hat{p}_i = \frac{\exp(z_i)}{\sum_{j=1, j\neq f}^{K} \exp(z_j)}. \tag{4}$$

Unlearning loss $\mathcal{L}$ is defined as the KL-divergence between target probability distribution p and the grounded unlearning distribution q, which can be written as

$$\mathcal{L} = \mathrm{KL}(p\|q) = p_f \log(\frac{p_f}{q_f}) + \sum_{i=1, i\neq f}^{K} p_i \log(\frac{p_i}{q_i}). \tag{5}$$

According to Eq. 3, we define normalized probabilities $\hat{p}_i = p_i / p_{\setminus f}$, which changes Eq. 5 as

$$\begin{aligned}\mathcal{L} =& p_f \log\left(\frac{p_f}{q_f}\right) + p_{\setminus f} \sum_{i=1, i\neq f}^{K} \hat{p}_i \left[\log\left(\frac{\hat{p}_i}{\hat{q}_i}\right) + \log\left(\frac{p_{\setminus f}}{p_{\setminus f}}\right)\right] \\ =& p_f \log\left(\frac{p_f}{q_f}\right) + p_{\setminus f} \log\left(\frac{p_{\setminus f}}{q_{\setminus f}}\right) + p_{\setminus f} \sum_{i=1, i\neq f}^{K} \hat{p}_i \log\left(\frac{\hat{p}_i}{\hat{q}_i}\right).\end{aligned} \tag{6}$$

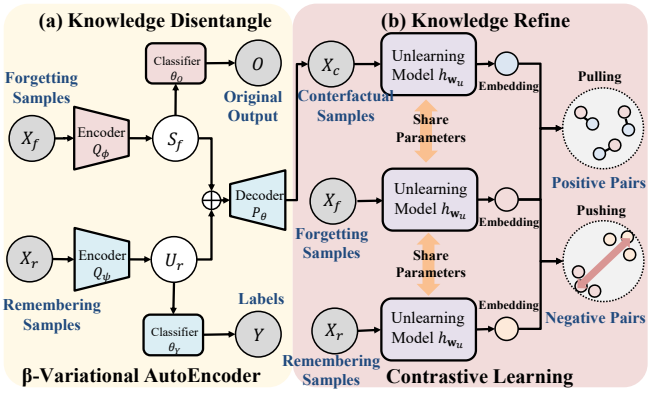


Figure 4: Structure design of (a) Knowledge Disentangle with β-VAE and (b) Knowledge Refine with counterfactual generating and contrastive learning.

Hence, $\mathcal{L}$ can be reformulated as :

$$\mathcal{L} = \underbrace{\mathrm{KL}(p^f \| q^f)}_{\text{FL}} + (1 - p_f) \underbrace{\mathrm{KL}(\hat{p} \| \hat{q})}_{\text{RL}}, \tag{7}$$

$\mathrm{KL}(p^f \,\|\, q^f)$ defines the discrepancy to minimize p_f near 0 in forgetting class distribution, and $\mathrm{KL}(\hat{p} \,\|\, \hat{q})$ computes the normalized distribution gap of remaining classes weighted by $1 - p_f$, thus preserving irrelevant knowledge to the forgetting. Essentially, MU can be regarded as multi-task learning of forgetting and remembering [Lin *et al.*, 2024], where loss of each task jointly constrains to learn the specific knowledge.

However, due to the coupling term $(1-p_f)$, the updates for these two tasks during training become entangled and biased, often causing one task to dominate the model update at the expense of the other. Specifically, an over-emphasis on the remaining loss may compromise the forgetting efficacy (incomplete unlearning), whereas a dominant forgetting loss can degrade the model's classification performance on remaining classes (low fidelity). To resolve such conflict, it's essential to decouple the optimization of these two objectives, where the forgetting and remembering goals should be treated as **independent constraints**, allowing the optimization to explicitly balance their gradients and achieve dual optimality in both effectiveness and fidelity.

4 Proposed Method

4.1 Overview

Built on knowledge distillation framework for unlearning, we illustrate in Fig. 4 with knowledge disentangle and refine. Regarding the original model $h_{\mathbf{w}_0}$ as teacher network, we train a student network $h_{\mathbf{w}_u}$ to remember $\mathcal{D}_r$ and forget $\mathcal{D}_f$.

During disentanglement, we adopt β-Variational AutoEncoder(β-VAE) to extract the shared attribute knowledge of forgetting data S_f and unique attribute knowledge of remaining data U_r. Specifically, we feed forgetting samples X_f and remembering samples X_r into two learnable encoders Q_ϕ and Q_ψ to extract S_f and U_r, respectively. Afterwards, S_f is supervised via a classifier θ_o to predict the original output O, and U_r is used to predict the label Y via classifier θ_y, where

both designs ensure θ_o and U_r capture correct semantics for knowledge disentanglement.

During refinement, we adopt schemes of counterfactual inference and contrastive learning. First, we generate counterfactual samples X_c with $[S_f, U_r]$. Then, we minimize the discrepancy between X_f and X_c by optimizing the counterfactual loss. Regarding counterfactual and remembering samples as key anchors, contrastive learning is finally performed constrained by RL and FL, simultaneously pulling the embeddings of positive pairs close and pushing the embeddings of negative pairs apart.

4.2 Knowledge Disentangle Towards Independent Optimization with β-VAE

Rethinking the MU loss reveals that while FL and RL aim at distinct objectives, the shared knowledge K acts as a confounder that prevents their independent optimization. Such entanglement necessitates a shift from logit-level manipulation to attribute-level semantic disentanglement for precise optimization. With constraints of independent MU loss, we further remove confounder K by disentangling knowledge with aid of the generated counterfactual data $\mathcal{D}_c$ following the guidance in Fig.2(c). However, generating $\mathcal{D}_c$ encounters two difficulties. First, sampling $\mathcal{D}_c$ from real-world images is impractical, since $\mathcal{D}_c$ is defined as a mixture of multiple attributes with rare relations. Secondly, generating $\mathcal{D}_c$ with individual attributes is unreachable as well, since both class-unique attributes U and shared attributes S are inherently unobservable and entangled in sample-level semantic space. Therefore, we disentangle knowledge with Variational Auto-encoders(VAEs), which separates the unobservable and entangled semantic attributes into two groups of distinct latent variables, thus being observable as individual samples with no confounders for further counterfactual generation.

In Fig. 4, we design an encoder-decoder structure to infer the unobservable S and U with two observable constraints, i.e., the observation context O and class label Y. Specifically, we first decompose each sample X as 1) unique attributes U corresponding to the intrinsic and class-related information to predict label Y, and 2) shared attributes S with attributes that shared across different classes. Such decomposing is achieved by extract S and U from samples with encoders ϕ and ψ respectively, where ϕ encodes shared attributes as $Q_\phi(S|X)$, and ψ encodes class-unique attributes as $Q_\psi(U|X)$. Then, we disentangle the joint distribution of entangled attributes as $P(S,U|X) = Q_\phi(S|X)Q_\psi(U|X)$ based on the Markov condition between these factors:

$$P(S,U|X) = Q_\phi(S|X)Q_\psi(U|X) = \frac{P(X,S,U,O,Y)}{P(X|S,U)P(O|S)P(Y|U)} \tag{8}$$

where $P(X,S,U,O,Y)$ denotes the joint distributions of all factors, $P(X|S,U)$ denotes the reconstruction output by decoder P_θ, $P(O|S)$ is predicted by classifier θ_O, and $P(Y|U)$ is predicted by classifier θ_S.

To ensure disentangle is constrained by $P(X,S,U,O,Y)$ with $\mathcal{D}$, we optimize by maximizing a weighted Evidence Lower Bound(ELBO)[Yue *et al.*, 2021]:

$$\begin{aligned}\mathcal{L}_U = & -\mathbb{E}_{Q_\phi}[\log P_\theta(O \mid S)] - \mathbb{E}_{Q_\psi}[\log P_\theta(Y \mid U)] \\ & - \mathbb{E}_{Q_\phi,Q_\psi}[\log P_\theta(X \mid S,U)] \\ & + \beta D_{\mathrm{KL}}(Q_\phi(S \mid X) \,\|\, P(S)) \\ & + \beta D_{\mathrm{KL}}(Q_\psi(U \mid X) \,\|\, P(U))\end{aligned} \tag{9}$$

where the first two terms supervise the classification accuracy of θ_O and θ_S, the third term improves the reconstruction quality by generating the original samples, and β is the weighting parameter. Essentially, ELBO optimizes the disentangling in a self-supervised manner, where ϕ is supervised by θ_O to focus on observation context of sample X, regardless of its belonging class. Meanwhile, ψ is supervised by θ_S to focus on class-unique semantics, which excludes the confounders from the shared knowledge for further counterfactual generation.

4.3 Knowledge Refine Constrained by FL and RL with Counterfactual Inference

Based on the disentanglement results $Q_\phi(S|X)$ and $Q_\psi(U|X)$, we employ them to generate $\mathcal{D}_c$ by counterfactual inference, where we further use $\mathcal{D}_c$ to refine knowledge.

Counterfactual Generation. For each forgetting sample $x_f \in \mathcal{D}_f$, we generate counterfactual sample $x_c \in \mathcal{D}_f$ by imagining a counterfactual reasoning case like "Given a tiger with attribute *haired*, what would it like with *wings*? " Such case is conducted with three typical steps, i.e., Abduction, Action, and Prediction. First, we sample the shared attributes $s_f \in S_f$ from disentangled $S \sim Q_\phi(S|X = x_f)$ as Abduction, which defines a *haired* tiger. By adding a unique attribute $u_r \in U_r$ from remaining data $\mathcal{D}_r$ which is sampled from $U \sim Q_\psi(U|X = x_r)$, we then utilize Action to borrow the unique *winged* attribute from bird. With both s_f and u_r, we finally imagine a counterfactual sample x_c by decoding $P_\theta(X|S = s_f, U = U_r)$ as Prediction.

With the generated $\mathcal{D}_c$, we refine knowledge with dual objectives of remembering and forgetting. Remembering establishes a new causal route $K \to \mathcal{D}_c \to Y$ to resolve the confounders for $\mathcal{D}_r$. Since multiple counterfactual samples $x_c \in \mathcal{D}_c$ and forgetting sample $x_f \in \mathcal{D}_f$ shares attributes S, we design counterfactual loss to keep S in forgetting sample:

$$\mathcal{L}_f = \sum_{(x_f,y_f)\in\mathcal{D}_f} KL\big(h_{\mathbf{w}_0}(X = (S_f, U_r)), h_{\mathbf{w}_u}(X = (S_f, U_f)\big) \tag{10}$$

where $h_{\mathbf{w}_0}(\cdot)$ and $h_{\mathbf{w}_u}(\cdot)$ denotes the output of original and unlearned model, respectively.

By leveraging the decoupled S_f and U_r in counterfactual samples, minimizing $\mathcal{L}_f$ aligns U_f with the manifold of U_r while keeping S_f invariant. Such mechanism implicitly substitutes $\mathcal{D}_f$ with $\mathcal{D}_c$, thereby effectively eliminating the confounder. The optimization objective of $\mathcal{L}_f$ implies:

$$\begin{aligned}\min \mathcal{L}_f &\Longrightarrow h_{w_u}(x_f) \approx h_{w_0}(S_f, u_r) \\ &\Longrightarrow p_f \to 0 \Longleftrightarrow \min \underbrace{\mathrm{KL}(p^f \| q^f)}_{\text{FL}},\end{aligned} \tag{11}$$

Optimizing $\mathcal{L}_f$ is equivalent to minimizing (FL). This drives the forgetting class probability (p_f) to zero, which enforces FL constraint to guarantee the effectiveness.

Dataset	Method	$\text{Acc}_{\mathcal{D}_r}$ ($\Delta\downarrow$)	$\text{Acc}_{\mathcal{D}_f}$ ($\Delta\downarrow$)	Acc_{val} ($\Delta\downarrow$)	MIA ($\Delta\downarrow$)	Avg. Gap $\downarrow$	RTE $\downarrow$
CIFAR-10	Retrain	100.00 (0.00)	0.00 (0.00)	77.09 (0.00)	9.64 (0.00)	0.00	14.11
	Finetune	99.63 (0.37)	0.22 (0.22)	76.82 (0.27)	25.48 (15.84)	4.18	2.37
	Negative Gradient	97.16 (2.84)	7.84 (7.84)	72.86 (4.23)	24.31 (14.67)	7.40	2.31
	Random Labels	98.49 (1.51)	10.40 (10.40)	71.69 (5.40)	0.60 (9.04)	6.59	0.73
	BadT	94.92 (5.08)	4.59 (4.59)	78.46 (1.37)	1.48 (8.16)	4.80	1.12
	Boundary	99.24 (0.76)	5.94 (5.94)	73.83 (3.26)	0.96 (8.68)	4.66	1.67
	SSD	92.35 (7.65)	0.00 (0.00)	71.25 (5.84)	1.85 (7.79)	5.32	3.10
	L1-Sparse	91.28 (8.72)	3.48 (3.48)	76.23 (0.86)	14.76 (5.12)	4.54	2.25
	ERM-KTP	94.50 (5.50)	1.20 (1.20)	73.80 (3.29)	11.50 (1.86)	2.96	2.50
	SCRUB	98.20 (1.80)	0.50 (0.50)	74.10 (2.99)	10.90 (1.26)	1.64	2.64
	GDR-GMA	97.45 (2.55)	2.55 (2.55)	74.50 (2.59)	12.10 (2.46)	2.54	1.40
	SalUn	96.10 (3.90)	3.90 (3.90)	72.90 (4.19)	8.50 (1.14)	3.28	2.77
	Ours	99.93 (0.07)	2.84 (2.84)	76.54 (0.55)	7.82 (1.82)	**1.32**	2.53
CIFAR-100	Retrain	99.98 (0.00)	0.00 (0.00)	73.08 (0.00)	2.54 (0.00)	0.00	35.28
	Finetune	94.25 (5.73)	0.00 (0.00)	70.12 (2.96)	18.23 (15.69)	6.10	2.45
	Negative Gradient	97.89 (2.09)	18.53 (18.53)	72.69 (0.39)	19.28 (16.74)	9.44	5.76
	Random Labels	90.07 (9.91)	4.89 (4.89)	68.21 (4.87)	0.24 (2.30)	5.49	0.81
	BadT	94.89 (5.09)	2.61 (2.61)	74.28 (1.20)	0.87 (1.67)	2.64	1.29
	Boundary	96.89 (3.09)	3.29 (3.29)	69.45 (3.63)	1.97 (0.57)	2.65	3.87
	SSD	82.19 (17.79)	0.14 (0.14)	65.27 (7.81)	1.80 (0.74)	6.62	5.48
	L1-Sparse	88.45 (11.53)	1.12 (1.12)	68.30 (4.78)	4.12 (1.58)	4.25	4.26
	ERM-KTP	92.15 (7.83)	0.85 (0.85)	69.40 (3.68)	3.80 (1.26)	3.41	6.25
	SCRUB	96.50 (3.48)	0.40 (0.40)	70.20 (2.88)	3.10 (0.56)	1.83	4.31
	GDR-GMA	95.80 (4.18)	0.95 (0.95)	69.80 (3.28)	3.45 (0.91)	2.33	3.38
	SalUn	93.40 (6.58)	0.65 (0.65)	68.90 (4.18)	2.90 (0.36)	2.94	2.94
	Ours	97.69 (2.29)	1.69 (1.69)	71.59 (1.49)	2.18 (0.36)	**1.46**	4.07

Table 1: Performance comparison on CIFAR-10 (ResNet-18) and CIFAR-100 (ResNet-34). The values in blue color represent the absolute gap between the experimental method and Retrain (lower is better).

Contrastive Learning. By establishing $K \rightarrow \mathcal{D}_c \rightarrow E'$, we refine knowledge to form its disentangled feature representation E'. Inspired by the impressive performance of contrastive learning [Chen *et al.*, 2020] to boost representation capability, we utilize $\mathcal{D}_c$ to refine E' by comparing both forgetting and remembering samples. Specifically, we firstly take (x_c, x_f) as positive pairs, where $P_x(x_f) = \{x_p | x_p \in \mathcal{D}_c\}$ and $N_x(x_f) = \{x_n | x_n \in \mathcal{D}_r, y_n \neq y_f\}$ are formed as negative pairs. We then pull close the positives to encourage U_f being varying to the random anchors U_r, meanwhile we push away negatives to enforce U_f shifting from its original representation through contrastive loss:

$$\mathcal{L}_c = \sum_{x_f \in \mathcal{D}_f} \frac{-1}{|N_x|} \log \frac{\exp(z_f \cdot z_p/\tau)}{\exp(z_f \cdot z_p/\tau) + \sum_{x_n \in N_x} \exp(z_f \cdot z_n/\tau)}, \tag{12}$$

where τ is a scalar temperature parameter, and $z = h_{\mathbf{w}_u}(x)$ denotes the extracted feature representation.

Minimizing the contrastive loss $\mathcal{L}_c$ preserves the intrinsic topology of the representations for the remaining data. Such structural stability implies that the normalized prediction distribution over the remaining classes is maintained. This alignment effectively minimizes the KL divergence between the unlearning model and the original model on the remaining data, thereby explicitly satisfying RL constraint and guaranteeing high fidelity on the remaining data.

Considering the causal route of $K \rightarrow \mathcal{D}_c \rightarrow E' \& Y$, the overall loss for knowledge refinement is $\mathcal{L} = \mathcal{L}_f + \lambda \mathcal{L}_c$, where λ is designed to balance remembering and forgetting.

5 Experiments

5.1 Experimental Setup

Unlearning Settings. We compare performance with several baseline MU methods in three scenarios, i.e., single-class forgetting, subclass forgetting and multi-class forgetting.

Datasets and Models. Following [Chundawat *et al.*, 2023; Chen *et al.*, 2023], we evaluate with three image classification datasets: CIFAR-10, CIFAR-20, CIFAR-100 [Krizhevsky *et al.*, 2009]. ResNet-18 [He *et al.*, 2016] and ResNet-34 [He *et al.*, 2016] are used as backbone network.

Baseline. The adopted baselines are: **Retrain**, **Finetune**, **Negative Gradient** [Golatkar *et al.*, 2020], **Random Labels** [Golatkar *et al.*, 2020], **BadT** [Chundawat *et al.*, 2023], **Boundary** [Chen *et al.*, 2023], **SSD** [Foster *et al.*, 2024], **L1-Sparse** [Jia *et al.*, 2023], **ERM-KTP** [Lin *et al.*, 2023], **SCRUB** [Kurmanji *et al.*, 2023], **GDR-GMA** [Lin *et al.*, 2024], **SalUn** [Fan *et al.*, 2024].

Metrics. Following [Chundawat *et al.*, 2023; Chen *et al.*, 2023], we use 6 metrics to evaluate unlearning: (1) $\textbf{Acc}_{\mathcal{D}_r}$, the classification accuracy on remaining data; (2) $\textbf{Acc}_{\mathcal{D}_f}$, the classification accuracy on forgetting data; (3) $\textbf{Acc}_{val}$, the classification accuracy on validation dataset; (4) Membership inference attack (**MIA**), examining whether the forgetting samples are in the training set; (5) Average Gap (**Avg.Gap**), performance gap comparing with Retrain; (6) **RTE** offers the consuming time of unlearning. Note that the better performance of an MU method corresponds to the smaller performance gap with Retrain.

Dataset	Method	$\text{Acc}_{\mathcal{D}_r}$ ($\Delta\downarrow$)	$\text{Acc}_{\mathcal{D}_f}$ ($\Delta\downarrow$)	Acc_{val} ($\Delta\downarrow$)	MIA ($\Delta\downarrow$)	Avg.Gap $\downarrow$	RTE $\downarrow$
CIFAR-20	Retrain	99.98 (0.00)	1.63 (0.00)	77.60 (0.00)	3.85 (0.00)	0.00	33.24
	Finetune	91.26 (8.72)	6.84 (5.21)	77.38 (0.22)	16.64 (12.79)	6.73	2.43
	Negative Gradient	97.57 (2.41)	16.38 (14.75)	78.04 (0.44)	4.52 (0.67)	4.57	3.29
	Random Labels	86.29 (13.69)	4.80 (3.17)	72.84 (4.76)	0.27 (3.58)	6.30	0.78
	BadT	95.31 (4.67)	6.84 (5.21)	79.25 (1.65)	1.47 (2.38)	3.48	1.15
	Boundary	96.31 (3.67)	5.98 (4.35)	73.59 (4.01)	1.28 (2.57)	3.65	2.43
	SSD	88.23 (11.75)	2.14 (0.51)	70.26 (7.34)	5.40 (1.55)	5.29	4.92
	L1-Sparse	90.15 (9.83)	3.10 (1.47)	75.80 (1.80)	9.45 (5.60)	4.67	2.86
	ERM-KTP	93.40 (6.58)	2.55 (0.92)	74.10 (3.50)	8.20 (4.35)	3.84	4.87
	SCRUB	96.26 (3.72)	3.24 (2.21)	75.85 (1.75)	5.48 (1.63)	2.33	2.05
	GDR-GMA	96.80 (3.18)	4.50 (2.87)	75.16 (2.44)	6.50 (2.65)	2.78	2.87
	SalUn	94.50 (5.48)	4.15 (2.52)	75.20 (2.40)	4.88 (1.03)	2.86	2.74
	Ours	96.58 (3.40)	1.84 (0.21)	74.43 (3.17)	1.89 (1.96)	**2.19**	3.04

Table 2: Performance comparisons for subclass forgetting on CIFAR-20.

Method	1 Class			5 Classes			10 Classes			20 Classes		
	$\text{Acc}_{\mathcal{D}_r}$	$\text{Acc}_{\mathcal{D}_f}$	Avg.Gap	$\text{Acc}_{\mathcal{D}_r}$	$\text{Acc}_{\mathcal{D}_f}$	Avg.Gap	$\text{Acc}_{\mathcal{D}_r}$	$\text{Acc}_{\mathcal{D}_f}$	Avg.Gap	$\text{Acc}_{\mathcal{D}_r}$	$\text{Acc}_{\mathcal{D}_f}$	Avg.Gap
Retrain	99.98	0.00	0.00	99.85	0.00	0.00	99.50	0.00	0.00	99.10	0.00	0.00
Finetune	94.25	0.00	6.10	90.15	0.15	9.85	85.40	0.80	14.15	78.20	1.50	21.05
Negative Gradient	97.89	18.53	9.44	95.50	20.10	11.20	92.20	22.50	14.85	88.50	25.90	17.60
Random Labels	90.07	4.89	5.49	88.40	6.50	7.25	85.50	8.80	10.35	80.10	11.20	16.50
BadT	94.89	2.61	2.64	93.40	3.50	3.45	91.20	4.80	4.90	88.50	6.10	7.80
Boundary	96.89	3.29	2.65	95.20	4.80	3.55	93.50	5.90	5.10	90.80	7.50	8.15
SSD	82.19	0.14	6.62	80.10	0.30	9.50	78.50	0.50	14.10	75.40	0.90	18.55
L1-Sparse	88.45	1.12	4.25	86.80	1.90	6.10	84.20	2.50	8.95	80.10	3.80	13.80
ERM-KTP	92.15	0.85	3.41	90.50	1.55	4.85	88.80	2.10	7.15	85.20	3.90	10.90
SCRUB	96.50	0.40	1.83	95.50	0.60	2.55	94.80	1.10	3.45	92.50	1.80	5.15
GDR-GMA	95.80	0.95	2.33	94.80	1.80	3.10	93.40	2.40	4.25	91.10	3.50	6.35
SalUn	93.40	0.65	2.94	92.10	0.90	4.05	90.20	1.20	5.85	88.80	1.95	8.70
Ours	97.69	1.69	**1.46**	97.10	1.80	**1.95**	96.50	2.10	**2.65**	95.80	2.50	**3.80**

Table 3: Performance comparisons with varying numbers of forgetting classes on CIFAR-100.

5.2 Experimental Evaluation

Comparison experiments for single-class forgetting. Following settings of [Chundawat *et al.*, 2023; Chen *et al.*, 2023], Table 1 presents the comparison results for single-class forgetting on CIFAR-10 and CIFAR-100 dataset. As the most vital metrics, our method achieves the lowest performance gaps against Retrain, reaching 1.32 on CIFAR-10 and 1.46 on CIFAR-100 respectively. Some methods may be the strongest when considering only a single metric, but this comes at the cost of sacrificing the other metrics. For example, Salun suffers from low fidelity by reaching 3.90% lower in $Acc_{\mathcal{D}_r}$ than Retrain, while Finetune fails in privacy scenario proved by gaining 15.84% MIA gap against Retrain. These results demonstrate that our idea of attribute-level semantic disentanglement enables precise unlearning, maintaining high fidelity without compromising effectiveness across most application scenarios.

Comparison experiments for subclass forgetting. Following settings of [Chundawat *et al.*, 2023; Foster *et al.*, 2024], we compare performance of subclass forgetting in Table 2, where the experiment of forgetting one sub-class per hyper-class on CIFAR-20 dataset is challenging, due to significant semantic overlap in the embedding space. While most methods suffer from severe performance degradation, our method achieves close performance with Retrain, achieving the smallest average gap of 2.19 against Retrain. Such phenomenon ensures our method is capable to balance forgetting and remembering, even with high semantic interference.

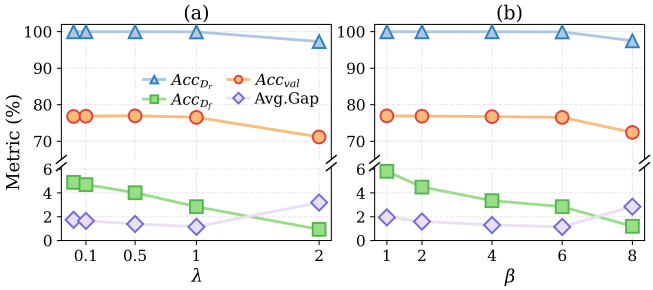
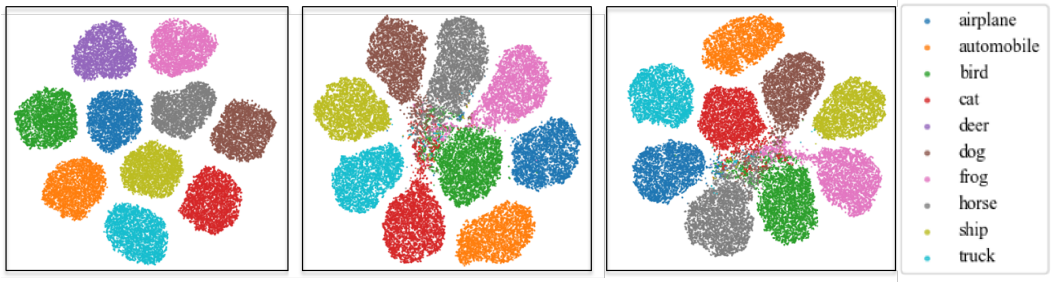


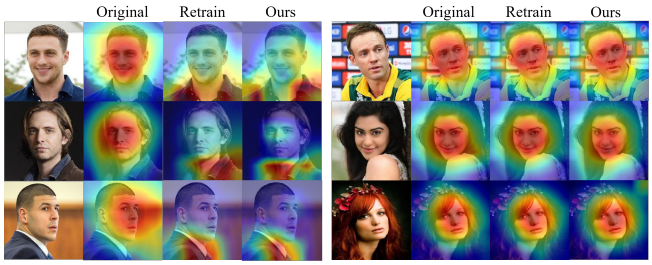
Figure 5: Ablation experiments for different hyper-parameters: (a) λ in overall loss and (b) β in VAE loss represented as Eq. 9.

Comparison experiments for multi-class forgetting. Table 3 presents the results of comparison experiments by forgetting multiple classes on CIFAR-100 dataset, i.e., 1, 5, 10 and 20 classes. while baseline performance degrades significantly as the number of forgetting classes gets larger, our method demonstrates superior scalability with the lowest Avg.Gap. In the highly challenging 20-class scenario, our method gains 95.80% $Acc_{\mathcal{D}_r}$, surpassing 92.50% achieved by SCRUB. Such phenomenon proves that our method with attribute-level knowledge disentanglement effectively resolves the confound of the shared knowledge, especially facing complex scenarios of forgetting multiple classes.



(a) Original (b) Retrain (c) Ours

Figure 6: Visualization of decision space tested on CIFAR-10 dataset, where purple points in (a) represent the forgetting class of "deer", and they are removed in (b) and (c).



(a) Forgetting Visualization (b) Remembering Visualization

Figure 7: Grad-CAM visualization of forgetting and remembering in face recognition task.

Class	$\text{Acc}_{\mathcal{D}_r}$	$\text{Acc}_{\mathcal{D}_f}$	Acc_{val}
0	99.98	2.88	76.11
1	99.93	4.34	74.66
2	99.96	3.82	77.41
3	99.93	2.84	76.54
4	99.93	2.84	76.54
5	100.00	2.22	77.90
6	99.94	0.70	75.22
7	99.89	4.10	75.19
8	99.96	3.54	75.41
9	99.99	1.00	75.26

Table 4: Results with different forgetting classes on CIFAR-10.

5.3 Ablation Study

Different classes to be forgotten. In Table 4, we present ablation results on CIFAR-10, where the forgetting class $\mathcal{D}_f$ is varied across 10 independent runs. Our method performs consistently well across all runs, regardless of the semantic overlap of the chosen class. Such insensitivity experiment highlights the robustness and applicability of our method.

Impact of λ. Fig. 5(a) shows the ablation results with different λ values, where λ is the weight of the overall loss. We can observe that a larger λ reduces the Avg.Gap, meanwhile removing it ($\lambda = 0$) increases the Avg.Gap by 0.59, demonstrating its effectiveness in establishing a rigid decision boundary. However, an excessive value of $\lambda = 2$ results in a significant performance drop with the rising Avg. Gap of 3.19, due to the over-fitting with the contrastive constraints. The optimal configuration is empirically determined to be $\lambda = 1$, which achieves the minimum Avg. Gap comparing with Retrain.

Impact of β. In Fig. 5(b), we compare different values of β, where β is the penalty factor in β-VAE loss. A lower β fails to fully remove the confounding knowledge, resulting in an incomplete unlearning with a high $Acc_{\mathcal{D}_f}$ of 5.82%. Conversely, an excessive value of defining $\beta = 8$ over-compresses features for discriminative representation, causing $Acc_{\mathcal{D}_r}$ to greatly drop. Therefore, we defined $\beta = 6$ to achieve the minimum Avg.Gap.

5.4 Qualitative Analysis

Visualization of Decision Space. As shown in Figure 6, we visualize the decision space by projecting with t-SNE, where each dot denotes a sample and the corresponding color indicating its belonging classes. Before unlearning in (a), all classes are learned as clusters in decision space, with clear boundary separating each other. When removing "deer" during training, its corresponding samples are misclassified in (b) achieved by Retrain. In Fig.6(c), the forgetting data has been predicted as the nearest classes after boundary shifting, even the cluster does not entirely spread out like Retrain.

Visualization of Forgetting and Remembering. Following [Chen *et al.*, 2023] to test capability of privacy preserving, we apply machine unlearning on face recognition task [Cao *et al.*, 2018]. As shown in the supplementary material, our method outperforms all other methods and achieves the smallest average performance gap against Retrain model. To further illustrate in Fig.7, we employ Grad-CAM [Selvaraju *et al.*, 2017] to visualize model's attention areas. The original model focuses on facial regions in (a), while the Retrain model and our method effectively adjust the recognition focus on areas out of face, thus successfully forgetting individual's privacy. Meanwhile, both methods maintain strong attention on relevant facial features of the remembering cases in (b), demonstrating high fidelity achieved by our method comparable to the Retrain model.

6 Conclusion

In this paper, we propose disentangled knowledge forgetting to address overly and incomplete unlearning. Extensive experimental results demonstrate that our method achieves state-of-the-art performance.

Acknowledgments

This work was supported in part by the National Key R&D Program of China under Grant 2023YFC3006501, the Natural Science Foundation of Jiangsu Province of China under Grants BK20242050 and BK20251504, the National Natural Science Foundation of China under Grant 6250074646, the High Performance Computing Platform, Hohai University, the Fundamental Research Funds for the Central Universities under Grant B250201042, and the Open Research Project of State Key Laboratory for Novel Software Technology, Nanjing University, under Grant KFKT2025B04.

Contribution Statement

Yuhang Xia and Cheng Zhen contributed equally to this work. Yirui Wu is the corresponding author.

References

[Ahmed *et al.*, 2025] Sk Miraj Ahmed, Umit Yigit Basaran, Dripta S. Raychaudhuri, Arindam Dutta, Rohit Kundu, Fahim Faisal Niloy, Basak Guler, and Amit K. Roy-Chowdhury. Towards source-free machine unlearning. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4948–4957, 2025.

[Bourtoule *et al.*, 2021] Lucas Bourtoule, Varun Chandrasekaran, Christopher A. Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *Proceedings of Symposium on Security and Privacy, SP*, pages 141–159, 2021.

[Cao *et al.*, 2018] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *Proceedings of International Conference on Automatic Face & Gesture Recognition, FG*, pages 67–74, 2018.

[Chen *et al.*, 2020] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML*, volume 119, pages 1597–1607, 2020.

[Chen *et al.*, 2023] Min Chen, Weizhuo Gao, Gaoyang Liu, Kai Peng, and Chen Wang. Boundary unlearning: Rapid forgetting of deep networks via shifting the decision boundary. In *Proceedings of Conference on Computer Vision and Pattern Recognition, CVPR*, pages 7766–7775, 2023.

[Chundawat *et al.*, 2023] Vikram S. Chundawat, Ayush K. Tarun, Murari Mandal, and Mohan S. Kankanhalli. Can bad teaching induce forgetting? unlearning in deep networks using an incompetent teacher. In *Proceedings of Conference on Artificial Intelligence, AAAI*, pages 7210–7217, 2023.

[Fan *et al.*, 2024] Chongyu Fan, Jiancheng Liu, Yihua Zhang, Eric Wong, Dennis Wei, and Sijia Liu. Salun: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. In *Proceedings of International Conference on Learning Representations, ICLR*, 2024.

[Foster *et al.*, 2024] Jack Foster, Stefan Schoepf, and Alexandra Brintrup. Fast machine unlearning without retraining through selective synaptic dampening. In *Proceedings of Conference on Artificial Intelligence, AAAI*, pages 12043–12051, 2024.

[George *et al.*, 2025] Naveen George, Karthik Nandan Dasaraju, Rutheesh Reddy Chittepu, and Konda Reddy Mopuri. The illusion of unlearning: The unstable nature of machine unlearning in text-to-image diffusion models. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13393–13402, 2025.

[Golatkar *et al.*, 2020] Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *Proceedings of Conference on Computer Vision and Pattern Recognition, CVPR*, pages 9301–9309, 2020.

[Graves *et al.*, 2021] Laura Graves, Vineel Nagisetty, and Vijay Ganesh. Amnesiac machine learning. In *Proceedings of Conference on Artificial Intelligence, AAAI*, pages 11516–11524, 2021.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of Conference on Computer Vision and Pattern Recognition, CVPR*, pages 770–778, 2016.

[Jia *et al.*, 2023] Jinghan Jia, Jiancheng Liu, Parikshit Ram, Yuguang Yao, Gaowen Liu, Yang Liu, Pranay Sharma, and Sijia Liu. Model sparsity can simplify machine unlearning. In *Proceedings of Annual Conference on Neural Information Processing Systems, NeurIPS*, 2023.

[Koloskova *et al.*, 2025] Anastasia Koloskova, Youssef Allouah, Animesh Jha, Rachid Guerraoui, and Sanmi Koyejo. Certified unlearning for neural networks. In *Proceedings of International Conference on Machine Learning*, 2025.

[Krizhevsky *et al.*, 2009] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[Kurmanji *et al.*, 2023] Meghdad Kurmanji, Peter Triantafillou, Jamie Hayes, and Eleni Triantafillou. Towards unbounded machine unlearning. In *Proceedings of Annual Conference on Neural Information Processing Systems , NeurIPS*, 2023.

[Lin *et al.*, 2023] Shen Lin, Xiaoyu Zhang, Chenyang Chen, Xiaofeng Chen, and Willy Susilo. ERM-KTP: knowledge-level machine unlearning via knowledge transfer. In *Proceedings of Conference on Computer Vision and Pattern Recognition, CVPR*, pages 20147–20155, 2023.

[Lin *et al.*, 2024] Shen Lin, Xiaoyu Zhang, Willy Susilo, Xiaofeng Chen, and Jun Liu. GDR-GMA: machine unlearning via direction-rectified and magnitude-adjusted gradients. In *Proceedings of International Conference on Multimedia, MM*, pages 9087–9095, 2024.

[Mehta *et al.*, 2022] Ronak Mehta, Sourav Pal, Vikas Singh, and Sathya N. Ravi. Deep unlearning via randomized conditionally independent hessians. In *Proceedings of Conference on Computer Vision and Pattern Recognition, CVPR*, pages 10412–10421, 2022.

[Niu *et al.*, 2021] Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. Counterfactual VQA: A cause-effect look at language bias. In *Proceedings of Conference on Computer Vision and Pattern Recognition, CVPR*, pages 12700–12710, 2021.

[Pearl *et al.*, 2016] Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. Causal inference in statistics: A primer. 2016. *Internet resource*, 2016.

[Selvaraju *et al.*, 2017] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of International Conference on Computer Vision, ICCV*, pages 618–626, 2017.

[Spartalis *et al.*, 2025] Christoforos N. Spartalis, Theodoros Semertzidis, Efstratios Gavves, and Petros Daras. Lotus: Large-scale machine unlearning with a taste of uncertainty. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10046–10055, 2025.

[Wei *et al.*, 2025] Rongzhe Wei, Peizhi Niu, Hans Hao-Hsun Hsu, Ruihan Wu, Haoteng Yin, Mohsen Ghassemi, Yifan Li, Vamsi K. Potluru, Eli Chien, Kamalika Chaudhuri, Olgica Milenkovic, and Pan Li. Do llms really forget? evaluating unlearning with knowledge correlation and confidence awareness. *CoRR*, abs/2506.05735, 2025.

[Yan *et al.*, 2022] Haonan Yan, Xiaoguang Li, Ziyao Guo, Hui Li, Fenghua Li, and Xiaodong Lin. ARCANE: an efficient architecture for exact machine unlearning. In *Proceedings of International Joint Conference on Artificial Intelligence, IJCAI*, pages 4006–4013, 2022.

[Yuan *et al.*, 2025] Xiaojian Yuan, Tianyu Pang, Chao Du, Kejiang Chen, Weiming Zhang, and Min Lin. A closer look at machine unlearning for large language models. In *Proceedings of International Conference on Learning Representations*, 2025.

[Yue *et al.*, 2021] Zhongqi Yue, Tan Wang, Qianru Sun, Xian-Sheng Hua, and Hanwang Zhang. Counterfactual zero-shot and open-set visual recognition. In *Proceedings of Conference on Computer Vision and Pattern Recognition, CVPR*, pages 15404–15414, 2021.

[Zhao *et al.*, 2022] Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. In *Proceedings of Conference on Computer Vision and Pattern Recognition, CVPR*, pages 11943–11952, 2022.