Novelty Assessment Report

Paper: Decoupling the Class Label and the Target Concept in Machine Unlearning
PDF URL: https://openreview.net/pdf?id=Xpj0yeMhpz
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-27

Abstract

Machine unlearning as an emerging research topic for data regulations, aims to adjust a trained model to approximate a retrained one that excludes a portion of training data. Previous studies showed that class-wise unlearning is effective in forgetting the knowledge of a training class, either through gradient ascent on the forgetting data or fine-tuning with the remaining data. However, while these methods are useful, they are insufficient as the class label and the target concept are often considered to coincide. In this work, we expand the scope by considering the label domain mismatch and investigate three problems beyond the conventional all matched forgetting, e.g., target mismatch, model mismatch, and data mismatch forgetting. We systematically analyze the new challenges in restrictively forgetting the target concept and also reveal crucial forgetting dynamics in the representation level to realize these tasks. Based on that, we propose a general framework, namely, TARget-aware Forgetting (TARF). It enables the additional tasks to actively forget the target concept while maintaining the rest part, by simultaneously conducting annealed gradient ascent on the forgetting data and selected gradient descent on the hard-to-affect remaining data. Various experiments under our new settings are conducted to demonstrate the effectiveness of our TARF.

Core Task Landscape

This paper addresses: **Decoupling Class Labels and Target Concepts in Machine Unlearning**

A total of **27 papers** were analyzed and organized into a taxonomy with **16 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Concept-Level Unlearning Methods**
- **Class-Level Unlearning Methods**
- **Label-Concept Mismatch Unlearning**
- **Domain and Modality-Specific Unlearning**
- **Specialized Unlearning Contexts**
- **Unlearning Robustness and Security**
- **Unlearning Surveys and Frameworks**

Complete Taxonomy Tree

- Decoupling Class Labels and Target Concepts in Machine Unlearning Survey Taxonomy
- Concept-Level Unlearning Methods
  - Representation Disentanglement Approaches
  - Feature-Based Disentanglement (2 papers)
    - [23] MLLM Machine Unlearning via Visual Knowledge Distillation (Yuhang Wang, 2025) View paper
    - [27] MaGA: Machine-Guided Amnesiac Unlearning through Target Feature Disentanglement (DISENTANGLEMENT, n.d.) View paper
  - Causal and Geometric Disentanglement (2 papers)
    - [2] CaMU: disentangling causal effects in deep model unlearning (Shaofei Shen, 2024) View paper
    - [16] Geometric-Disentangelment Unlearning (Duo Zhou, 2025) View paper
  - Direct Concept Erasure
  - Alignment and Distribution Matching (3 papers)
    - [5] Score Forgetting Distillation: A Swift, Data-Free Method for Machine Unlearning in Diffusion Models (Chen, 2024) View paper
    - [9] Machine Unlearning in Hyperbolic vs. Euclidean Multimodal Contrastive Learning: Adapting Alignment Calibration to MERU (Ã￿lex Pujol Vidal, 2025) View paper
    - [26] Concept Siever: Towards Controllable Erasure of Concepts from Diffusion Models without Side-effect (AK Singh, n.d.) View paper
  - Sparse Feature Interventions (1 papers)
    - [22] CRISP: Persistent Concept Unlearning via Sparse Autoencoders (Arad, 2025) View paper
- Class-Level Unlearning Methods
  - Distillation-Based Class Unlearning (2 papers)
  - [7] Decoupled distillation to erase: A general unlearning method for any class-centric tasks (Zhou Yu, 2025) View paper
  - [13] Machine unlearning with affine hyperplane shifting and maintaining for image classification (Mengda Liu, 2023) View paper
  - Gradient and Generative Class Unlearning

    - Gradient-Based Weight Manipulation (1 papers)
        - [4] SalUn: Empowering Machine Unlearning via Gradient-based Weight Saliency in Both Image Classification and Generation (Liu Jiancheng, 2023) View paper
    - Generative Model Class Unlearning (2 papers)
        - [14] Machine Unlearning using a Multi-GAN based Model (Amartya Hatua, 2024) View paper
        - [17] Class Machine Unlearning for Complex Data via Concepts Inference and Data Poisoning (Chang Wenhan, 2024) View paper
- Label-Concept Mismatch Unlearning ★ (2 papers)
    - [0] Decoupling the Class Label and the Target Concept in Machine Unlearning (Anon et al., 2026) View paper
    - [18] A Targeted Machine Unlearning Method for Sensitive Data in Military Helicopter Models (Hyun KWON, 2025) View paper
- Domain and Modality-Specific Unlearning
    - Text and Document Unlearning (2 papers)
    - [1] Disentangling biased knowledge from reasoning in large language models via machine unlearning (Bayar, 2025) View paper
    - [12] Machine unlearning for document classification (Lei Kang, 2024) View paper
    - Image Generation Unlearning (3 papers)
    - [11] Erasing Concepts, Steering Generations: A Comprehensive Survey of Concept Suppression (Xie Yi-wei, 2025) View paper
    - [20] Realistic Image-to-Image Machine Unlearning via Decoupling and Knowledge Retention (Varshney, 2025) View paper
    - [24] Forget-Me-Not: Learning to Forget in Text-to-Image Diffusion Models (Zhang Gong, 2023) View paper
    - Multimodal and Graph Unlearning (2 papers)
    - [6] Forgetting and remembering are both you need: Balanced graph structure unlearning (Chenhan Zhang, 2024) View paper
    - [19] Approximate Domain Unlearning for Vision-Language Models (Kawamura Kodai, 2025) View paper
- Specialized Unlearning Contexts
    - Federated and Distributed Unlearning (1 papers)
    - [8] Learning to unlearn in federated learning (Yi-xiong Wang, 2024) View paper
    - Noisy and Imbalanced Data Unlearning (1 papers)
    - [3] Classifying Long-tailed and Label-noise Data via Disentangling and Unlearning (Shu Chen, 2025) View paper
    - Distributional Shift and Independence (1 papers)
    - [21] Towards Independence Criterion in Machine Unlearning of Features and Labels (Han Ling, 2024) View paper
- Unlearning Robustness and Security (2 papers)
    - [15] Boundary-Anchored Functional Duplication Attacks on Machine Unlearning (Chengzhi Shangguan, 2025) View paper
    - [25] Unlearnable Clusters: Towards Label-Agnostic Unlearnable Examples (Jiaming Zhang, 2023) View paper
- Unlearning Surveys and Frameworks (1 papers)
    - [10] A Survey of Machine Unlearning in Generative AI Models: Methods, Applications, Security, and Challenges (An Huang, 2025) View paper

Narrative

Core task: Decoupling class labels and target concepts in machine unlearning. The field of machine unlearning has evolved to address the challenge that class labels and the underlying concepts a model learns do not always align perfectly. The taxonomy reflects this complexity through several main branches: Concept-Level Unlearning Methods focus on removing specific learned features or attributes rather than entire classes, while Class-Level Unlearning Methods target traditional label-based removal. Label-Concept Mismatch Unlearning explicitly tackles scenarios where the two diverge, such as when a class contains multiple distinct concepts or when spurious correlations exist. Domain and Modality-Specific Unlearning adapts techniques to particular data types like images or graphs, and Specialized Unlearning Contexts address settings such as federated learning. Unlearning Robustness and Security examines adversarial challenges, while Unlearning Surveys and Frameworks provide overarching perspectives. Works like SalUn Gradient Saliency[4] and Score Forgetting Distillation[5] illustrate gradient-based and distillation-based approaches, respectively, while Federated Unlearning[8] and Hyperbolic Multimodal Unlearning[9] show domain-specific adaptations.

A particularly active line of work explores how to disentangle biased or spurious knowledge from legitimate class information, as seen in Disentangling Biased Knowledge[1] and CaMU Causal Unlearning[2], which leverage causal reasoning to isolate unwanted associations. Another thread addresses noisy or imbalanced data scenarios, exemplified by Longtailed Label Noise[3], where label quality itself complicates the unlearning target. The original paper, Decoupling Class Label[0], sits squarely within the Label-Concept Mismatch Unlearning branch alongside Military Helicopter Unlearning[18], emphasizing the need to separate what a class name denotes from what the model has actually encoded. Compared to works like Disentangling Biased Knowledge[1], which focus on bias removal, Decoupling Class Label[0] more directly interrogates the structural misalignment between labels and learned representations, offering a complementary perspective on ensuring precise and interpretable unlearning.

Related Works in Same Category

The following **1 sibling papers** share the same taxonomy leaf node with the original paper:

1. A Targeted Machine Unlearning Method for Sensitive Data in Military Helicopter Models

Authors: Hyun KWON, Jang-Woon Baek | **Year/Venue**: 2025 | **URL**: View paper

Abstract

â¦ In response, this study introduces a novel unlearning method â¦ a structured machine unlearning technique for military data, â¦ the class label and the target concept in machine unlearning,ââ¦

Relationship Analysis

Both papers belong to the Label-Concept Mismatch Unlearning category, addressing scenarios where class labels and target concepts do not align in machine unlearning. While the original paper proposes a general framework (TARF) for handling various label domain mismatches (target mismatch, model mismatch, data mismatch) through annealed forgetting and target-aware retaining, the candidate paper focuses specifically on applying targeted unlearning methods to sensitive military helicopter data. The key difference is that the original paper provides a comprehensive theoretical and empirical analysis of the label-concept decoupling problem with a general-purpose solution, whereas the candidate paper appears to be a domain-specific application targeting military data protection.

Contributions Analysis

Overall novelty summary. The paper proposes TARF, a framework for machine unlearning that decouples class labels from target concepts, addressing scenarios where the two do not align. It sits in the 'Label-Concept Mismatch Unlearning' leaf of the taxonomy, which contains only two papers including this one. This is a notably sparse research direction compared to more crowded branches like 'Concept-Level Unlearning Methods' or 'Class-Level Unlearning Methods', suggesting the paper explores a relatively underexplored problem space within the broader unlearning literature.

The taxonomy reveals that most prior work assumes label-concept alignment, with neighboring branches focusing on either concept-level removal (e.g., disentangling biased knowledge, causal unlearning) or class-level forgetting (e.g., gradient-based weight manipulation, distillation methods). The paper's position bridges these areas by explicitly addressing mismatch scenarios—target mismatch, model mismatch, and data mismatch—that fall outside the scope of traditional concept-level or class-level methods. Its sibling paper in the same leaf examines military helicopter unlearning, indicating shared interest in label-concept divergence but different application contexts.

Among the 22 candidates examined through limited semantic search, none were found to clearly refute any of the three main contributions. The first contribution (decoupling labels and concepts) examined 2 candidates with no refutations; the second (representation-level forgetting dynamics) and third (TARF framework) each examined 10 candidates with no refutations. This suggests that within the search scope, the specific combination of addressing label-concept mismatch through representation-level analysis and annealed gradient ascent appears relatively novel, though the limited search scale means potentially relevant work outside the top-22 semantic matches may exist.

Based on the available signals from 22 examined candidates and the sparse taxonomy leaf, the work appears to occupy a distinct position in the unlearning landscape. The explicit focus on label-concept decoupling and the systematic treatment of three mismatch scenarios differentiate it from neighboring concept-level and class-level methods. However, the limited search scope and small sibling set mean this assessment reflects novelty within the examined literature rather than an exhaustive field-wide comparison.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: Decoupling class label and target concept in machine unlearning

Description: The authors introduce new unlearning settings that decouple the class label from the target concept, modeling scenarios where the forgetting data, model output, and target concept have mismatched label domains. This expands beyond the conventional assumption that the target concept coincides with the class label.

This contribution was assessed against **2 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Intelligent Continuous Monitoring to Handle Data Distributional Changes for IoT Systems

URL: View paper

Brief Assessment

Continuous Monitoring IoT[48] addresses catastrophic forgetting in IoT systems with continuous learning for operational changes and drift detection. This is fundamentally different from machine unlearning scenarios involving label domain mismatches and decoupled target concepts.

---

2. Uncertainty-Calibrated Test-Time Model Adaptation Without Forgetting

URL: View paper

Brief Assessment

Uncertainty Calibrated Adaptation[47] addresses test-time model adaptation for distribution shifts, not machine unlearning with label domain mismatch. The candidate focuses on adapting models during testing without forgetting in-distribution performance, while the original contribution concerns unlearning scenarios where forgetting data, model output, and target concept have mismatched label domains.

Contribution 2: Systematic analysis of forgetting dynamics at the representation level

Description: The authors provide a systematic empirical and theoretical analysis of how representation-level dynamics affect unlearning under label domain mismatch. They identify challenges such as insufficient representation and decomposition lacking, and derive formal results connecting representation similarity to forgetting dynamics.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. CRFU: Compressive Representation Forgetting Against Privacy Leakage on Machine Unlearning

URL: View paper

Brief Assessment

CRFU Compressive Representation[38] focuses on privacy leakage defense through compressive representation and information bottleneck theory, not on analyzing representation-level dynamics under label domain mismatch or identifying challenges like insufficient representation and decomposition lacking.

---

2. Unlearning Isn't Deletion: Investigating Reversibility of Machine Unlearning in LLMs

URL: View paper

Brief Assessment

Reversibility Investigation[40] focuses on representation-level analysis to assess whether unlearning is reversible (i.e., easily restored), not on how representation-level dynamics affect unlearning under label domain mismatch or the challenges of insufficient representation and decomposition lacking.

---

3. Ferrari: federated feature unlearning via optimizing feature sensitivity

URL: View paper

Brief Assessment

Ferrari Feature Sensitivity[44] focuses on federated feature unlearning via optimizing feature sensitivity in federated learning settings, not on representation-level dynamics in centralized machine unlearning with label domain mismatch.

---

4. Understanding the behavior of representation forgetting in continual learning

URL: View paper

Brief Assessment

Representation Forgetting Behavior[39] focuses on measuring representational shifts in continual learning through linear transformation perspectives, not on machine unlearning scenarios with label domain mismatch. The candidate addresses a fundamentally different problem domain.

---

5. Feature-based machine unlearning for vertical federated learning in iot networks

URL: View paper

Brief Assessment

Vertical Federated Feature[45] focuses on vertical federated learning in IoT networks with feature-based machine unlearning, not on representation-level dynamics and forgetting challenges in centralized machine unlearning settings.

---

6. Towards Reliable Forgetting: A Survey on Machine Unlearning Verification, Challenges, and Future Directions

URL: View paper

Brief Assessment

Reliable Forgetting Survey[41] focuses on verification methods for machine unlearning rather than analyzing representation-level dynamics during the unlearning process itself. The survey categorizes verification approaches but does not provide systematic analysis of how representation similarity affects forgetting dynamics or derive formal results connecting these concepts.

---

7. Feature-Selective Representation Misdirection for Machine Unlearning

URL: View paper

Brief Assessment

Feature Selective Misdirection[46] focuses on feature-selective perturbations in representation space for unlearning harmful knowledge in LLMs, not on analyzing representation-level dynamics under label domain mismatch in classification tasks.

---

8. How Secure is Forgetting? Linking Machine Unlearning to Machine Learning Attacks

URL: View paper

Brief Assessment

Secure Forgetting[43] focuses on security threats in machine unlearning (backdoor attacks, membership inference, adversarial attacks) rather than representation-level dynamics and forgetting challenges under label domain mismatch.

---

9. An information theoretic evaluation metric for strong unlearning

URL: View paper

Brief Assessment

Information Theoretic Metric[42] focuses on evaluating unlearning through mutual information in intermediate features, not on analyzing representation-level dynamics under label domain mismatch or identifying challenges like insufficient representation and decomposition lacking in mismatched scenarios.

---

10. Representation space maintenance: Against forgetting in continual learning

URL: View paper

Brief Assessment

Representation Space Maintenance[37] focuses on preventing catastrophic forgetting in continual learning by maintaining representation space, not on machine unlearning dynamics under label domain mismatch.

Contribution 3: TARF framework for target-aware forgetting

Description: The authors propose TARF, a unified framework that addresses mismatched unlearning scenarios through annealed forgetting and target-aware retaining. The method dynamically identifies target data and separates entangled representations to approximate retraining on the retaining data.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. â Ï: Gradient-based and Task-Agnostic machine Unlearning

URL: View paper

Brief Assessment

Gradient Task Agnostic[28] focuses on adaptive gradient ascent for general subset forgetting without addressing label domain mismatch scenarios (target/model/data mismatch) that TARF specifically tackles through annealed forgetting and target-aware retaining mechanisms.

---

2. A Survey of Machine Unlearning in Generative AI Models: Methods, Applications, Security, and Challenges

URL: View paper

Brief Assessment

Generative AI Survey[10] is a survey paper on machine unlearning in generative AI models. It does not propose a specific unlearning framework or method, but rather reviews existing approaches. The minimal context provided does not demonstrate that it presents a target-aware forgetting framework predating TARF.

---

3. CUFG: Curriculum Unlearning Guided by the Forgetting Gradient

URL: View paper

Brief Assessment

CUFG Curriculum[36] focuses on curriculum-based unlearning with forgetting gradients for stability, not on target-aware forgetting with annealed gradient ascent for mismatched unlearning scenarios.

---

4. Zero-Shot Class Unlearning in CLIP with Synthetic Samples

URL: View paper

Brief Assessment

Zero-Shot CLIP[35] focuses on class unlearning in CLIP models using synthetic samples and Lipschitz regularization, not on general target-aware forgetting frameworks with annealed gradient ascent for mismatched unlearning scenarios.

---

5. Federated Unlearning for Samples Based on Adaptive Gradient Ascent of Angles

URL: View paper

Brief Assessment

Adaptive Gradient Angles[33] focuses on federated unlearning of random samples using gradient ascent with adaptive angle adjustment, not on target-aware forgetting frameworks that address mismatched unlearning scenarios through annealed forgetting and target-aware retaining.

---

6. Forget the Token and Pixel: Rethinking Gradient Ascent for Concept Unlearning in Multimodal Generative Models

URL: View paper

Brief Assessment

The candidate paper (Token Pixel Rethinking[31]) focuses on concept unlearning in multimodal generative models using gradient ascent approaches, while TARF addresses machine unlearning in classification tasks with mismatched label domains through annealed forgetting and target-aware retaining mechanisms.

---

7. Knowledge Unlearning for Mitigating Privacy Risks in Language Models

URL: View paper

Brief Assessment

Knowledge Unlearning Privacy[29] focuses on privacy-preserving unlearning in language models through gradient ascent on target sequences, not on the general RL framework with annealed forgetting and target-aware retaining for mismatched unlearning scenarios in classification tasks.

---

8. Unified Gradient-Based Machine Unlearning with Remain Geometry Enhancement

URL: View paper

Brief Assessment

Remain Geometry Enhancement[34] focuses on gradient decomposition with Hessian-based manifold geometry for efficient unlearning, while TARF addresses mismatched label domains through annealed forgetting and dynamic target identification—fundamentally different problem formulations and technical approaches.

---

9. Fine-grained Pluggable Gradient Ascent for Knowledge Unlearning in Language Models

URL: View paper

Brief Assessment

Pluggable Gradient Ascent[30] focuses on fine-grained token-level gradient control for knowledge unlearning in language models, while TARF addresses mismatched unlearning scenarios in classification tasks through annealed forgetting and target-aware retaining with representation-level analysis.

---

10. Multi-Objective Large Language Model Unlearning

URL: View paper

Brief Assessment

Multi-Objective LLM[32] focuses on gradient ascent methods for LLM unlearning with multi-objective optimization to address gradient explosion and catastrophic forgetting. TARF addresses different challenges (mismatched label domains, target identification, representation entanglement) in class-wise unlearning for classification tasks, not LLM unlearning.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] Decoupling the Class Label and the Target Concept in Machine Unlearning View paper
- [1] Disentangling biased knowledge from reasoning in large language models via machine unlearning View paper
- [2] CaMU: disentangling causal effects in deep model unlearning View paper
- [3] Classifying Long-tailed and Label-noise Data via Disentangling and Unlearning View paper
- [4] SalUn: Empowering Machine Unlearning via Gradient-based Weight Saliency in Both Image Classification and Generation View paper
- [5] Score Forgetting Distillation: A Swift, Data-Free Method for Machine Unlearning in Diffusion Models View paper
- [6] Forgetting and remembering are both you need: Balanced graph structure unlearning View paper
- [7] Decoupled distillation to erase: A general unlearning method for any class-centric tasks View paper
- [8] Learning to unlearn in federated learning View paper
- [9] Machine Unlearning in Hyperbolic vs. Euclidean Multimodal Contrastive Learning: Adapting Alignment Calibration to MERU View paper
- [10] A Survey of Machine Unlearning in Generative AI Models: Methods, Applications, Security, and Challenges View paper
- [11] Erasing Concepts, Steering Generations: A Comprehensive Survey of Concept Suppression View paper
- [12] Machine unlearning for document classification View paper
- [13] Machine unlearning with affine hyperplane shifting and maintaining for image classification View paper
- [14] Machine Unlearning using a Multi-GAN based Model View paper
- [15] Boundary-Anchored Functional Duplication Attacks on Machine Unlearning View paper
- [16] Geometric-Disentangelment Unlearning View paper
- [17] Class Machine Unlearning for Complex Data via Concepts Inference and Data Poisoning View paper
- [18] A Targeted Machine Unlearning Method for Sensitive Data in Military Helicopter Models View paper
- [19] Approximate Domain Unlearning for Vision-Language Models View paper
- [20] Realistic Image-to-Image Machine Unlearning via Decoupling and Knowledge Retention View paper
- [21] Towards Independence Criterion in Machine Unlearning of Features and Labels View paper
- [22] CRISP: Persistent Concept Unlearning via Sparse Autoencoders View paper
- [23] MLLM Machine Unlearning via Visual Knowledge Distillation View paper
- [24] Forget-Me-Not: Learning to Forget in Text-to-Image Diffusion Models View paper
- [25] Unlearnable Clusters: Towards Label-Agnostic Unlearnable Examples View paper
- [26] Concept Siever: Towards Controllable Erasure of Concepts from Diffusion Models without Side-effect View paper
- [27] MaGA: Machine-Guided Amnesiac Unlearning through Target Feature Disentanglement View paper
- [28] â Ï: Gradient-based and Task-Agnostic machine Unlearning View paper
- [29] Knowledge Unlearning for Mitigating Privacy Risks in Language Models View paper
- [30] Fine-grained Pluggable Gradient Ascent for Knowledge Unlearning in Language Models View paper
- [31] Forget the Token and Pixel: Rethinking Gradient Ascent for Concept Unlearning in Multimodal Generative Models View paper

- [32] Multi-Objective Large Language Model Unlearning View paper
- [33] Federated Unlearning for Samples Based on Adaptive Gradient Ascent of Angles View paper
- [34] Unified Gradient-Based Machine Unlearning with Remain Geometry Enhancement View paper
- [35] Zero-Shot Class Unlearning in CLIP with Synthetic Samples View paper
- [36] CUFG: Curriculum Unlearning Guided by the Forgetting Gradient View paper
- [37] Representation space maintenance: Against forgetting in continual learning View paper
- [38] CRFU: Compressive Representation Forgetting Against Privacy Leakage on Machine Unlearning View paper
- [39] Understanding the behavior of representation forgetting in continual learning View paper
- [40] Unlearning Isn't Deletion: Investigating Reversibility of Machine Unlearning in LLMs View paper
- [41] Towards Reliable Forgetting: A Survey on Machine Unlearning Verification, Challenges, and Future Directions View paper
- [42] An information theoretic evaluation metric for strong unlearning View paper
- [43] How Secure is Forgetting? Linking Machine Unlearning to Machine Learning Attacks View paper
- [44] Ferrari: federated feature unlearning via optimizing feature sensitivity View paper
- [45] Feature-based machine unlearning for vertical federated learning in iot networks View paper
- [46] Feature-Selective Representation Misdirection for Machine Unlearning View paper
- [47] Uncertainty-Calibrated Test-Time Model Adaptation Without Forgetting View paper
- [48] Intelligent Continuous Monitoring to Handle Data Distributional Changes for IoT Systems View paper