Unlearning's Blind Spots

Over-Unlearning and Prototypical Relearning Attack

SeungBum Ha **Saerom Park**† **Sung Whan Yoon**†

Ulsan National Institute of Science and Technology (UNIST) • ICML 2026

What is Machine Unlearning?

Machine Unlearning (MU) removes the influence of training data from a model.

1 The Goal

Erase the influences of "forget set" from a trained model, so it behaves as if that data was never used.

2 Why It Matters

Enforces the "Right to be Forgotten" (GDPR, CCPA) and removes harmful, biased, or copyrighted data.

3 The Challenge

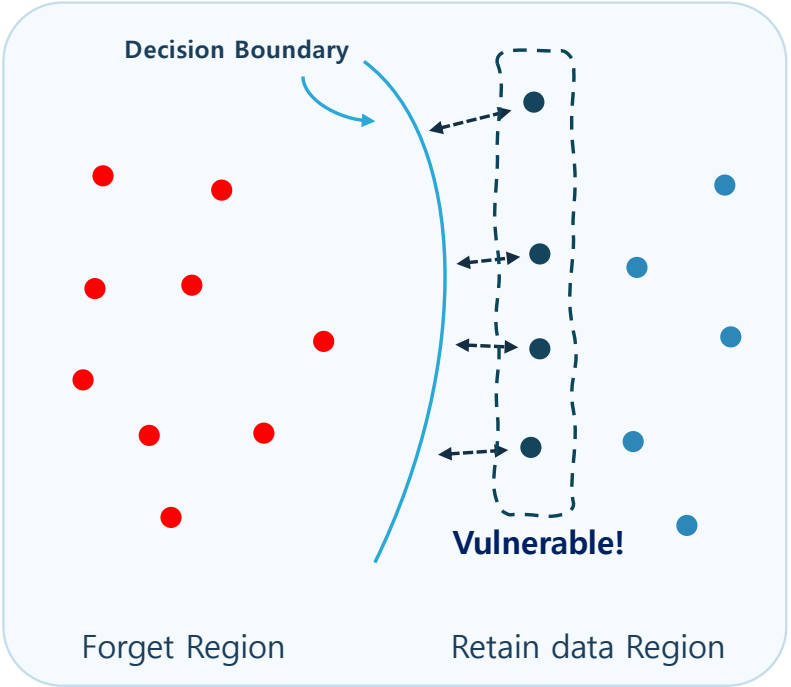
Must forget the target data while preserving accuracy on everything else, all without full retraining.

Blind Spot #1: Over-Unlearning

THE PROBLEM

Over-unlearning is boundary-local.

- Average retained accuracy is a **global metric that hides where damage occurs.**
- Suppressing a class reshapes the decision boundary around the forget region.
- The most vulnerable retained samples are those **closest to the forget class.**
- Methods keep high overall accuracy but suffer large drops on the top-2/5/10% forget-adjacent subset.



Quantifying It: The OU@ε Metric

A retain-data-free metric for boundary-proximal collateral damage.

1 Perturbed set Generation

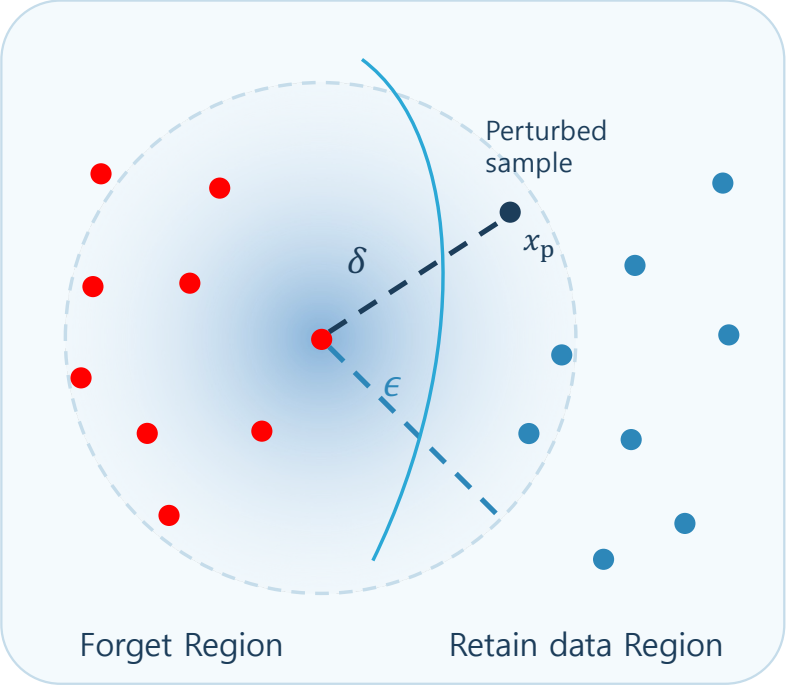
Apply ℓ_∞ PGD attack (ε = 0.03) to forget samples, pushing them toward the decision boundary

2 Masking

Measure each sample's masked predictive distribution before vs. after unlearning.

3 Compare

OU@ε, the JS divergence between the two distributions. Higher = more collateral damage.



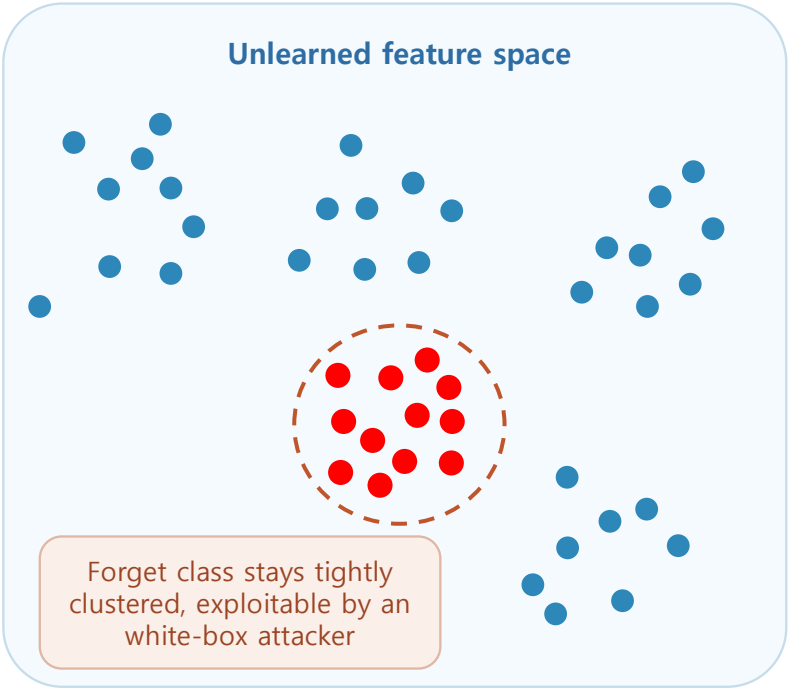
$$\mathrm{OU}@\varepsilon := \mathbb{E}_{\boldsymbol{x}_\mathrm{p}\sim\mathcal{A}_\varepsilon(\mathcal{D}_f)}\left[D\left(\tilde{\sigma}(z(\boldsymbol{x}_\mathrm{p};\theta)) \,\middle\|\, \sigma(z(\boldsymbol{x}_\mathrm{p};\theta_u))\right)\right]$$

Blind Spot #2: The Relearning Threat

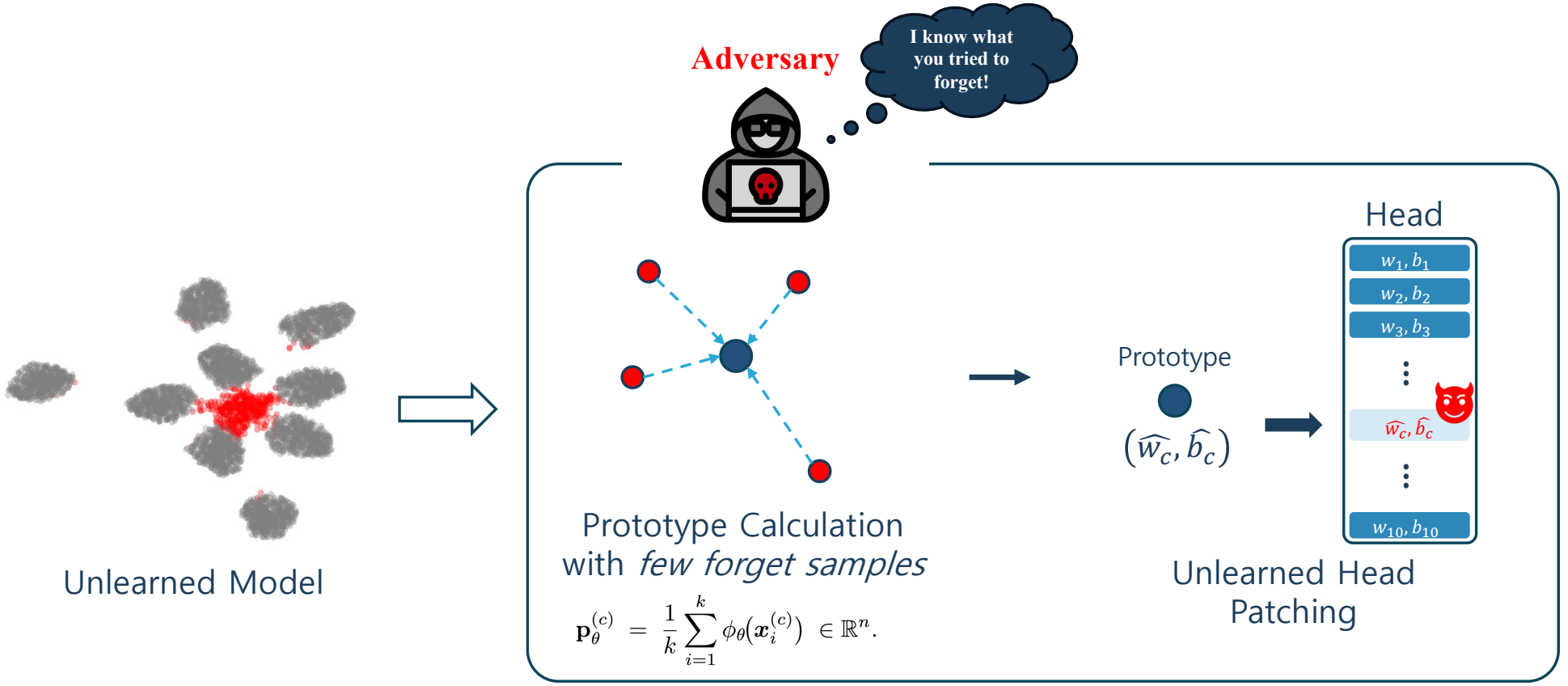
THE KEY OBSERVATION

Forgotten classes stay clustered.

- Even SotA methods leaves forget-class samples **tightly clustered in feature space**, despite their absence from the decision boundary.
- The model still retains enough capacity to **infer the "erased" identity.**
- Prior fine-tuning relearning attacks (from LLMs) **fail to exploit this in vision.**
- **Real risk: an adversary can harvest a few face images from social media to seed the attack.**



Prototypical Relearning Attack (PRA)



Spotter: Novel Unlearning Framework

One objective that simultaneously closes both blind spots, with no retained data needed.

$\mathcal{L}_u$ Masked KD Unlearning

Distills the original model's masked distribution, forget-class probabilities driven to zero.

Plug-and-play: swappable with any existing MU loss.

→ **Forgets the class**

$\mathcal{L}_o$ Over-Unlearning Penalty

Applies the same masked KD to PGD-perturbed forget samples, regularizing the local decision boundary.

→ **Suppresses OU@ε**

$\mathcal{L}_{sim}$ Intra-Class Dispersion

Minimizes pairwise cosine similarity of forget embeddings, scatters the cluster so no prototype can be built.

→ **Neutralizes PRA**

Spotter: Novel Unlearning Framework

$$\mathcal{L} = \lambda_1 \mathcal{L}_u + (1 - \lambda_1) \mathcal{L}_o + \lambda_2 \mathcal{L}_{sim}$$

Unlearning

Masked-softmax KD on the forget set, erases the class.

Over-Unlearning

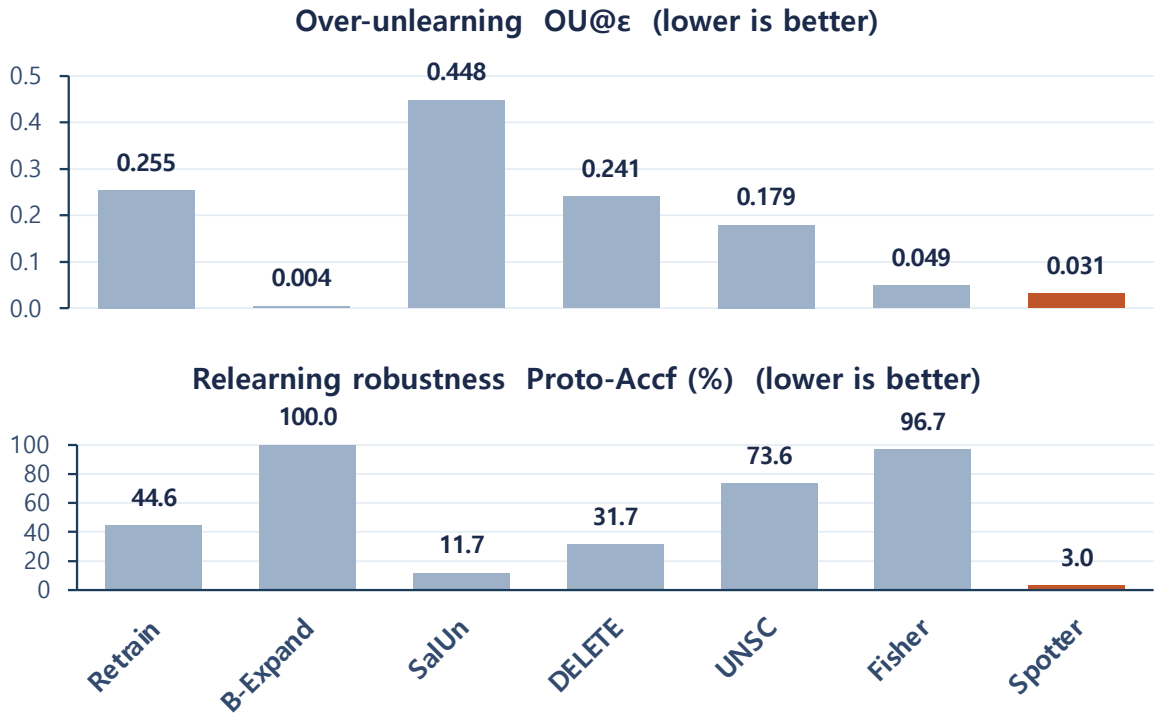
Same KD on perturbed set, protects the boundary region.

Dispersion

Pairwise cosine similarity of forget embeddings, breaks the forget cluster.

- **Only two hyperparameters (λ_1, λ_2) added**
- **Plug-and-play Compatibility: can be attached to the other unlearning pipelines.**

Results: State-of-the-Art on CIFAR-100



Why it wins

8×

lower OU@ε than DELETE (0.031 vs 0.241)

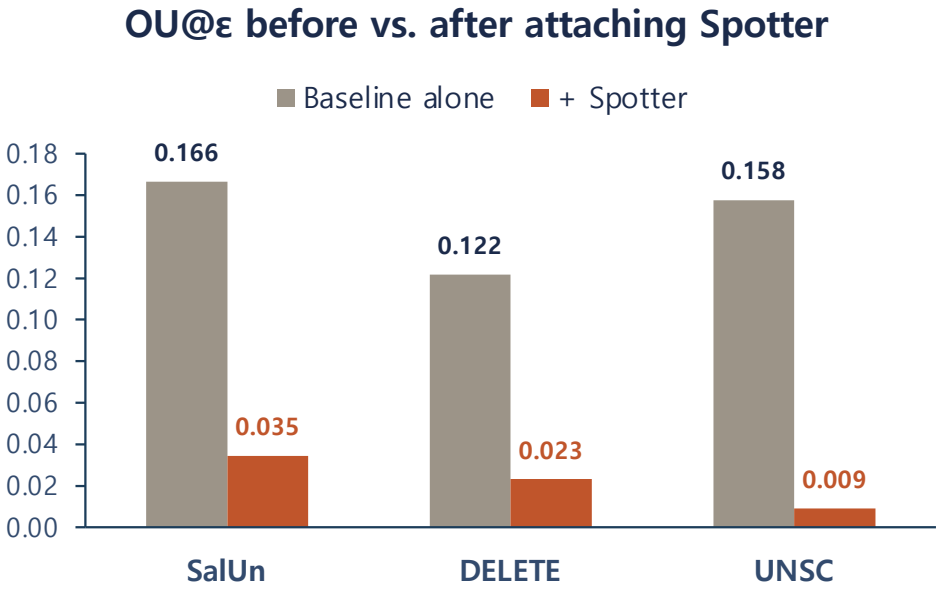
~3%

Proto-Acc$_f$, PRA suppressed at $\lambda_2 = 1$

100%

forget accuracy removed, retained utility intact

Plug-and-Play Compatibility



Drop-in upgrade

UNSC

OU@ε 0.158 → 0.009 (× 0.05 reduction)

DELETE

0.122 → 0.023 with PRA neutralized

SalUn

0.166 → 0.035, even recovers retention

Generalizing Beyond CIFAR

Spotter holds up along two axes: it transfers cleanly to new datasets, and it works across different model backbones.

Across Datasets

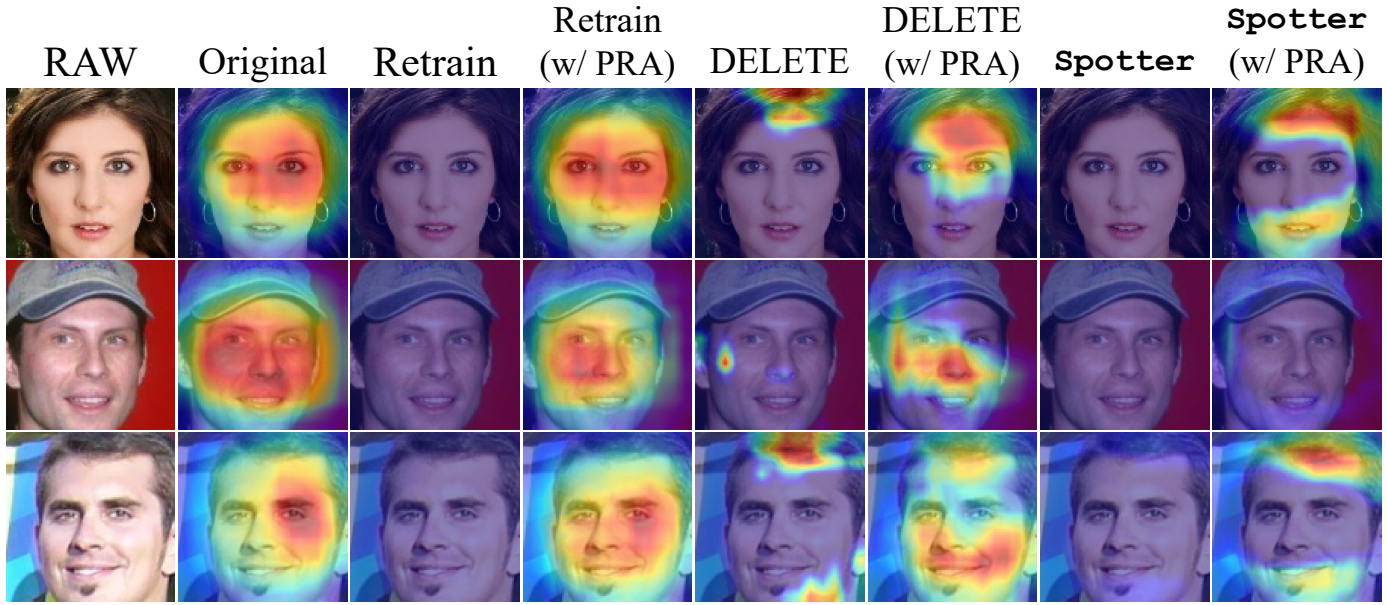
- Forgets cleanly on a larger, harder benchmark (Tiny-ImageNet) without eroding retained classes.
- Scales to real-world face recognition (CASIA-WebFace), unlearning hundreds of identities while keeping the rest intact.
- Stays stable even when a large fraction of classes is removed at once on CIFAR-100.

Across Models & Backbones

- Extends to modern vision transformers (ViT and DeiT) with the same behavior.
- Applies to a pretrained vision-language model (CLIP) that classifies via fixed text prompts.

Grad-CAM Under the Prototypical Relearning Attack

Class activation maps for a forgotten identity, clean and under PRA



DELETE lets PRA re-concentrate attention on identity cues (eyes, nose, mouth); Spotter keeps activations weak and off-target.

Toward Robust, Safety-Driven Unlearning

Spotter reframes unlearning as a safety problem, not just only forgetting, but forgetting without leaving exploitable traces.

Audit, don't assume

A certified deletion is not proof of safety. OU@ε and PRA give model owners concrete tools to audit residual leakage after unlearning.

Defense by design

Dispersion and boundary-masked distillation bake robustness into the unlearning step itself, with no retained data and no separate hardening pass.

Realistic regimes

Holds under sequential deletion requests and severe (30–50%) class removal, matching "unlearning-as-a-service" deployments.

Conclusion

We illuminate two overlooked blind spots in class-level unlearning and deliver a practical remedy.

1

Over-Unlearning Metric OU@ε

Retain-data-free measure of boundary-local over-unlearning.

2

Prototypical Relearning Attack

A few forget samples resurrect erased knowledge via residual structure.

3

Spotter

Masked distillation + dispersion neutralizes both, plug-and-play, no retained data.

Toward reliable, regulation-conscious, and attack-resilient machine unlearning

Thank You.

Questions and discussion are welcome.
Let's connect!

PROJECT PAGE
github.com/Seung-B/Spotter-Unlearning

EMAIL (First Author)
ethereal0507@unist.ac.kr

ICML International Conference On Machine Learning
UNIST ULSAN NATIONAL INSTITUTE OF SCIENCE AND TECHNOLOGY
MIIT Machine Intelligence and Information Theory Laboratory
Safe AI Safe and Trustworthy AI Lab