FORGOTTEN OR JUST UNSEEN? SEMANTIC ERASURE AND AUDITING FOR FACE UNLEARNING.

Anonymous authors
Paper under double-blind review

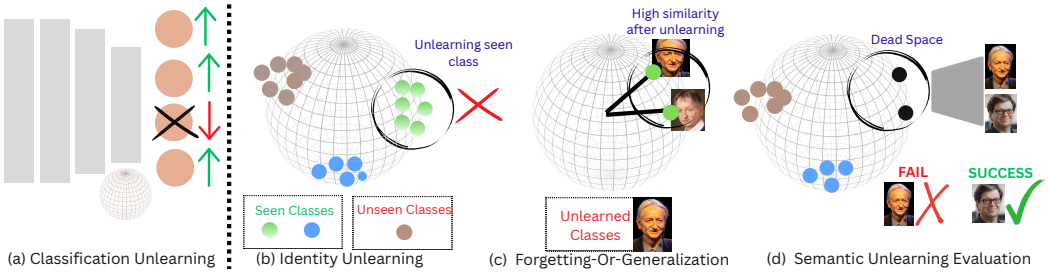


Figure 1: The Forgetting-Or-Generalization (FOG) problem. (a) Classification unlearning measures success by reduced accuracy in a closed set. (b) Face recognition embeddings verify both seen and unseen identities. (c) After unlearning, forgotten samples still exhibit high similarity, creating ambiguity. (d) Semantic unlearning evaluation measures success by the absence of recoverable identity semantics.

ABSTRACT

In open-set face recognition, forgetting an identity and never having seen it in the first place can exhibit indistinguishable behavior, raising a fundamental question of what it truly means to "unlearn" an identity. We term this the Forgetting-Or-Generalization (FOG) ambiguity. Existing unlearning measures rely on accuracy-based metrics, whereas face recognition operates in an open-set, verification-based setting where unseen identities are expected to be recognized correctly. This mismatch makes facial erasure difficult to evaluate and renders verification-accuracy-based unlearning metrics misleading. We formalize this facial unlearning problem and introduce a generative auditing framework for semantic visualization of identity erasure, revealing that existing unlearning methods retain identity information despite appearing successful under conventional metrics. To address this, we propose a feature-space unlearning algorithm that removes structural identity groupings by erasing identity regions using region-aware surrogate mechanisms while preserving recognition of non-target identities. Our framework shifts face unlearning evaluation from accuracy-driven to semantics-based assessment, providing stronger empirical privacy validation. Experiments on CASIA-WebFace, WebFace4M, MS1Mv3, and D-LORD show that FUSE achieves the lowest forget-set semantic residual among nine unlearning methods (as low as 0.27), while preserving retain-set verification and open-set accuracy (99.31% on LFW, against 99.53% for the original model).

1 INTRODUCTION

When a European citizen discovers their face in a commercial recognition system and invokes their "right to be forgotten" under GDPR Article 17, the system must delete their biometric data. This raises a question that current practice cannot answer: how can we verify that the identity has been erased from the model itself, and not from the database? With over 1 billion faces enrolled in recognition systems worldwide (Tiwari et al., 2022; Unique Identification Authority of India; Burt, 2023), this is not a hypothetical concern; rather, it is a daily privacy challenge.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Current machine unlearning methods suffer when applied to face recognition (Nguyen et al., 2025) due to open set scenario. These methods, designed for classification tasks (Li et al., 2025), assume that forgetting success can be measured by accuracy degradation on removed classes. But face recognition operates differently (Deng et al., 2022). These systems work by computing embedding similarities in an open-set scenario, where the model must recognize identities it has never seen during training (Scheirer et al., 2013). This leads to what we call the **FOG problem: Forgetting-Or-Generalization ambiguity**: if a face recognition model maintains its discriminative power, it should successfully verify a "forgotten" identity just as it verifies any unseen person. The very property that makes face recognition useful, generalization to unseen identities, makes forgetting impossible to verify with conventional metrics (Sekhari et al., 2021).

This ambiguity is not just theoretical. When we audit existing unlearning methods with our semantic residual measure, we find forget-set residuals as high as 0.65 (Table 1), indicating that a substantial fraction of identity-specific structure survives in the embedding space despite apparently successful unlearning. Methods like gradient ascent (Thudi et al., 2022), boundary shifting (Chen et al., 2023), and saliency-based unlearning (Fan et al., 2024) all suffer from this fundamental issue. The forgotten faces can still be reconstructed from their supposedly erased embeddings. Privacy regulations require verifiable deletion, but current evaluation protocols using match/non-match or accuracy drop tests (Carlini et al., 2022) fundamentally cannot distinguish between a truly forgotten identity and one that simply behaves like an unseen person, echoing recent findings that output-level unlearning often only suppresses information that remains in the representation (Gao et al., 2026; Jang et al., 2026).

We need a rethinking of how unlearning works in embedding spaces. Rather than trying to degrade verification accuracy, which contradicts the model's purpose, we must ensure that semantic information cannot be recovered from the features. This requires understanding how identities are geometrically encoded (Wang et al., 2018; Liu et al., 2017) and developing methods to surgically remove these representations while preserving the overall discriminative structure (Wang et al., 2025).

In this paper, we formalize the FOG ambiguity problem and demonstrate why conventional evaluation fails for open-set recognition systems (refer to Figure 1 for visual understanding). We introduce semantic residual auditing as a stronger evaluation framework that quantifies identity leakage directly from the embedding space, without requiring image generation. Based on these insights, we propose **FUSE: Forgetting Using Structural Erasure**, a geometry-aware method that models identity distributions as hypercones and performs targeted semantic erasure. Unlike existing approaches that modify classifiers (Tarun et al., 2024) or individual embeddings, FUSE transforms the geometric structure of identity regions, achieving improved privacy protection while maintaining recognition capability. Our contributions are:

- **Problem Formalization:** We formalize the FOG problem, stating that verification-based evaluation cannot distinguish between successful forgetting and natural generalization in open-set systems.
- **Generative Auditing Framework:** We propose a evaluation technique, termed Semantic Residual, that reveals existing methods leak information despite apparent forgetting.
- **FUSE:** We introduce a geometry-aware unlearning approach using region-aware hypercone modeling that minimizes semantic leakage.
- **Comprehensive Evaluation:** We demonstrate across four datasets and three backbones that FUSE most closely matches exact retraining in recoverable identity information (semantic residual) and membership inference while preserving retain-set verification within 0.06 to 0.08 of the original model and open-set accuracy above 99% on LFW.

1.1 Related Work

Classification based unlearning: Most unlearning methods operate on the classifier layer (Fan et al., 2024; Zhou et al., 2025), fine-tuning or resetting weights of specific logits (Tarun et al., 2024; Kodge et al., 2024). Others use gradient ascent on the forget set (Golatkar et al., 2020; Cha et al., 2024; Thudi et al., 2022), boundary shifting (Chen et al., 2023), null-space projection (Chen et al., 2024; Hoang et al., 2024), optimization-based forgetting (Hayase et al., 2020), task arithmetic (Ilharco et al., 2023), adversarial strategies (Di et al., 2025), or differential privacy (Sekhari et al.,

2021). However, these methods modify decision boundaries, whereas face verification relies only on embedding similarity, and they measure success by closed-set accuracy drop, which is ill-defined when unseen identities should still be verified.

Feature level unlearning: Other approaches act on representations, recalibrating contrastive encoders such as SimCLR, MoCo, and CLIP (Chen et al., 2020; He et al., 2020; Radford et al., 2021) via alignment objectives (Wang et al., 2024; 2025), targeting clusters without labels (Shen et al., 2024), or approximating data influence with second-order derivatives (Alexander et al., 2023). Recent analyses further show that output-level success often masks information that remains recoverable from internal representations (Gao et al., 2026; Jang et al., 2026). These methods address sample-level membership rather than identity-level semantics, and none evaluates whether a forgotten identity remains semantically recoverable, which is the central question in open-set face unlearning.

Unlearning in face recognition is limited to classification settings or attribute-level removal (Choi & Na, 2023; Guo et al., 2022). Recent work disperses identity embeddings on the hypersphere (Zakharov, 2026) and concurrent work (Ünsal Öztürk et al., 2026) likewise observes that a retrained model still verifies removed identities, but responds by making them unlinkable through dispersion or near-orthogonal targets, evaluated only at verification operating points. Identity unlearning has also been studied in ID-conditioned generators (Estrada et al., 2026; Tok, 2026). In contrast, we target the geometric structure of the feature space, i.e., the overall class distribution rather than individual logits. Cones on the hypersphere have been used to *train* face embeddings, e.g., CosFace (Zhuang & Lu, 2022) mines hard pairs with cones; we instead use cone geometry to *erase* and *audit* identities after training.

2 Identity Unlearning: The FOG Problem

In face recognition, the goal is to check if a query image x_q matches any enrolled gallery identity $\{x_g^1, x_g^2, \ldots, x_g^n\}$. Unlike closed-set classification, these systems work in an open-set setting. Instead of predicting class labels, the model uses a feature extractor $f_\theta : \mathbb{R}^{H\times W\times C} \to \mathbb{R}^d$ to map both query and gallery images into a d-dimensional embedding space. For two images, verification using feature space is done by comparing embeddings with a similarity metric such as cosine similarity:

$$\mathrm{sim}(f_\theta(x_q), f_\theta(x_g^i)) \geq \tau$$

If the similarity exceeds a threshold τ, the system declares a match; otherwise, a mismatch. This process relies on the geometry of the embedding space and ignores the classifier head from training.

Formally, let the training set be $D_{tr} = \{(x_i, y_i)\}_{i=1}^N$, where x_i are face images and $y_i \in Y$ are identity labels. The model is trained with a discriminative loss $\mathcal{L}_{disc}$ (e.g., ArcFace (Deng et al., 2022), CosFace (Wang et al., 2018)) so that embeddings of identities in Y are well separated. At test time, the system sees a new set $D_{te} = \{x'_j\}_{j=1}^M$ with identities $z_j \in Z$, where Z may be disjoint with Y ($Z \cap Y = \emptyset$). Despite being trained on Y, the model must generalize to unseen identities in Z by encoding features that remain discriminative (Sekhari et al., 2021). Verification at test time therefore depends only on embedding similarity, not on class predictions.

Now consider unlearning. Suppose we want to forget a subset of training identities $Y_u \subset Y$, with corresponding data $D_u = \{(x, y) \mid y \in Y_u\}$. After unlearning, we obtain an updated extractor f_{θ^-}. In principle, the embeddings of Y_u should no longer be tied to their original labels. In practice, however, the system still works in unseen settings. Forgotten samples may behave like unseen identities: they still cluster together in the embedding space and remain verifiable under f_{θ^-}. This means that even if the classifier layer forgets Y_u, the embeddings may still carry strong identity representations in the feature space.

We call this the **FOG: Forgetting-Or-Generalization ambiguity**. Consider two cases:

- **Unseen identity:** If Z was never part of training, the model can still verify it. This reflects generalization of the embedding space.
- **Forgotten identity:** If Y_u was trained but later unlearned, we expect it to behave indistinguishably from an unseen identity, yet in practice, the embeddings of Y_u may still form a cluster, allowing verification just as in the unseen case.

Definition (FOG Ambiguity Criterion). Let Y denote the set of training identities, $Y_u \subseteq Y$ the forget set, and Z a set of held-out identities satisfying $Z \cap Y = \emptyset$. We consider the identity-disjoint open-set face verification setting, where verification is performed on unseen identities.

Let f_θ be the encoder before unlearning and f_{θ^-} after unlearning. For a verification threshold τ, define $\mathrm{ver}_\tau(x, x'; f) = \mathbf{1}[\langle f(x), f(x')\rangle \geq \tau]$, and let $V(f, \tau; S)$ denote the same-identity verification rate computed over identity set S.

We say that the *Forget-or-Generalization (FOG) ambiguity* arises whenever

$$V(f_{\theta^-}, \tau; Y_u) \approx V(f_{\theta^-}, \tau; Z), \tag{1}$$

that is, forgotten identities exhibit verification behavior comparable to held-out identities after unlearning. One could instead force $V(f_{\theta^-}, \tau; Y_u)$ far below $V(f_{\theta^-}, \tau; Z)$ so that forgotten identities become unlinkable (Ünsal Öztürk et al., 2026); however, this departs from the behavior of a model retrained without Y_u and makes the removed identities conspicuous as the only ones that the model fails to verify. We therefore target $V_f \approx V_{uc}$ and resolve the remaining ambiguity semantically.

Consider an auditor whose observations are restricted to verification scores (or their induced match/non-match decisions) produced by the encoder f_{θ^-}. The auditor seeks to distinguish between the following hypotheses: (1) the queried identity belongs to a held-out identity (Z), (2) the queried identity belongs to the forget set (Y_u). If the score distributions induced by Y_u and Z are statistically indistinguishable, then no auditor operating solely on these verification scores can reliably distinguish the two hypotheses.

The FOG criterion therefore highlights a limitation of verification-based evaluation. Similar verification performance for forgotten and unseen identities may arise either because identity information has truly been removed or because the forgotten identity continues to behave as a semantically consistent identity under the learned representation. Consequently, verification accuracy alone cannot certify forgetting, motivating complementary evaluations based on generative reconstruction and representation-level distribution analysis.

3 From FOG to FUSE: Forgetting Using Structural Erasure

Rather than modifying classifier weights, representation-based unlearning removes a class's latent representation from the embedding space. Forgotten test samples may still cluster for verification, but the model should no longer retain the semantic traces that encode the identity. To this end, we propose FUSE (Forgetting Using Structural Erasure), which constructs surrogate representations of the target identity, models its support region on the hypersphere, and erases this region to suppress residual identity information.

3.1 Surrogate Representation Retrieval

To erase an identity's semantic footprint, we first characterize how it is distributed in the embedding space. The embeddings $\{z_i = f_\theta(x_i) \mid x_i \in \mathcal{D}_{Y_u}\}$ of an identity do not collapse to a single point but span a local region on the unit hypersphere $\mathbb{S}^{d-1}$, the geometric support of the class, which unlearning must identify and remove.

Since a face consists of semantically distinct subregions that contribute differently to the embedding, whole-face images alone do not capture the intra-class variation of an identity. We therefore use part-based embeddings $z_r^{\text{reg}} \in \mathbb{R}^d$, $r \in \{\text{upper}, \text{lower}\}$, alongside whole-face embeddings $z^{\text{face}} \in \mathbb{R}^d$; a finer eyes/nose/mouth partition yields comparable results (Appendix E.4.6). For a target identity y_u, we retrieve from a vector database: (i) its sample embeddings $\{z_i^{\text{face}}\}$, (ii) its region embeddings $\{z_{r,i}^{\text{reg}}\}$, (iii) its prototype $\bar{z}_{y_u} = \frac{1}{N}\sum_i z_i$, and (iv) prototypes of its nearest neighbors $\{\bar{z}_{y_n}\}$ under cosine similarity. These vectors serve as surrogate traces of how the identity is encoded, and form the basis for structural erasure (details in Appendix D).

3.2 Structural Trace Identification

A single facial region is not unique enough to reliably represent an identity, since many individuals may share similar local features. However, when different regions are combined, they form a unique

signature. In face recognition models, each regional embedding lies on a hypersphere, where angular distances encode similarity. To ensure that combinations of regions also respect this spherical geometry, we generate fused embeddings using spherical interpolation (SI). Given two normalized regional embeddings $z_{1,i}^{\text{reg}}$ and $z_{2,i}^{\text{reg}}$ for class i, we form a composite embedding as:

$$z_i^{combo} = SI(z_{1,i}^{reg}, z_{2,i}^{reg}, \beta) = \frac{\sin((beta)\theta)}{\sin\theta} z_{1,i}^{reg} + \frac{\sin(\beta\theta)}{\sin\theta} z_{2,i}^{reg} \tag{2}$$

where $\theta = \arccos(z_{1,i}^{\text{reg}\top} z_{2,i}^{\text{reg}})$ is the geodesic angle between them and β is interpolation parameter. To learn which variations are truly identity-specific, we use a contrastive learning objective. A small MLP projection network $p_\sigma : \mathbb{R}^d \to \mathbb{R}^d$ (Appendix D) maps fused embeddings z_i to the structural space s_i of the class. Positive pairs are embeddings from different samples of the same identity, while negatives come from different identities. With normalized features $s_i = p_\sigma(z_i^{\text{combo}})$, the contrastive loss is defined as : $\mathcal{L}_{\text{struct}} = -\sum_{i,j\in Y_u} \log \frac{\exp(s_i^\top s_j/T}{\sum_{k\in Y}\exp(s_i^\top s_k/T)}$, which encourages embeddings of the same person to stay close, while separating them from others. The loss is used to enforce the combined embedding within the class space using context-based dropout to introduce intra-class variation. Thus, we obtain a set of candidate vectors s_i for the target class.

3.3 Target Region Formation

FUSE aims to create a *dead space* in the region of the feature space where the target identity used to reside. In face recognition models (Deng et al., 2022), samples of an identity follow angular relationships and form a hyperconical structure on the hypersphere (Figure 2). Given the candidate features $\{s_i\}_{i=1}^n$ of an identity $k \in Y_u$, we therefore model its embedding distribution as a spherical cone, described by a cone center (the most representative direction) and a cone spread (the variability around it). The cone model itself is not our contribution, since cone-shaped identity regions have been used for training face embeddings (Zhuang & Lu, 2022); our contribution is its use for FOG-aware erasure and auditing, i.e., defining the region to be removed and measuring whether forgotten samples remain within it.

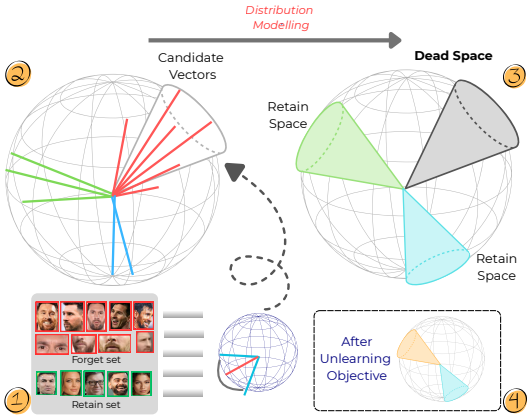


Figure 2: Structural erasure flow of FUSE illustrating 'dead space' (a region of the hypersphere devoid of identity-specific semantic structure) formation using surrogate representation.

Cone Center (Modal Direction): For an identity k, let $\mathcal{S}_k = \{i : y_i = k\}$ denote the set of samples belonging to it, with embeddings $\{s_i\}$. The cone center is simply the mean direction of all embeddings, and its normalized vector μ_k represents the axis of the identity cone, pointing toward the most typical embedding location for class k.

$$\bar{\mu}_k = \frac{1}{|\mathcal{S}_k|}\sum_{i\in\mathcal{S}_k} s_i, \qquad \mu_k = \frac{\bar{\mu}_k}{\|\bar{\mu}_k\|_2} \tag{3}$$

Cone Spread (Angular Deviation): Not all embeddings perfectly align with the center. To measure how much they deviate, we compute the angular distance between each embedding s_i and the cone axis μ_k: $\theta_i = \arccos(\langle s_i, \mu_k\rangle)$. Overall spread is defined as the root-mean-square angular deviation: $\sigma_k = \sqrt{\frac{1}{|\mathcal{S}_k|}\sum_{i\in\mathcal{S}_k}\theta_i^2}$ Here, σ_k captures how tightly or loosely the embeddings of identity k cluster around their cone center. A smaller spread indicates a compact, well-defined identity cluster, whereas a larger spread corresponds to greater intra-class variability.

Probabilistic Cone Model: To capture the variability of embeddings within each identity cone, we model their angular distribution using a wrapped normal distribution on the hypersphere. Intuitively, this distribution treats deviations from the cone center as angular noise, while respecting the spherical geometry of the embedding space. For an embedding s_i associated with identity $k : k \in Y$, the

log-probability of belonging to the cone is given by:

$$\log p(s_i \mid k) = -\frac{\theta_i^2}{2\sigma_k^2} + \text{const}, \qquad \text{LL}_{\text{proxy}}(s_i; k) = \exp\Big(-\frac{\theta_i^2}{2\sigma_k^2}\Big) \tag{4}$$

This probabilistic formulation provides a tractable density model over embeddings, where identity spread depends on σ_k. We approximate the likelihood of an embedding s_i under class k by a Gaussian-like function of its angular deviation from the class center. Formally, we use the unnormalized log-likelihood model $\log p(s_i \mid k)$. Practically, we define a bounded proxy likelihood score $LL_{\text{proxy}} \in (0, 1]$ which provides a tractable similarity measure even though it is not a fully normalized density on the hypersphere.

3.4 Geometry-Aware Erasure

We proposed a region-based loss function for the purpose of trace removal from the feature space.

Cone Repulsion Loss: To erase the representation of a target identity y_u, we introduce a cone repulsion loss that pushes its embeddings out of their original identity cone. Let $\mathcal{S}_{y_u}$ denote the samples of the target identity and $\tilde{s}_i$ their embeddings under the unlearned model. We penalize samples whose angular distance from the original cone axis μ_{y_u} remains below a margin:

$$\mathcal{L}_{\text{cr}} = \frac{1}{|\mathcal{S}_{y_u}|} \sum_{i \in \mathcal{S}_{y_u}} \max\Big(0,\ \alpha\sigma_{y_u} - \arccos(\langle \tilde{s}_i, \mu_{y_u} \rangle)\Big), \tag{5}$$

We use λ_1=1, λ_2=1, λ_3=0.1; sensitivity is analyzed in Appendix E.4.2. where σ_{y_u} is the cone spread and α is a repulsion factor. We set $\alpha = 3$: under a Gaussian approximation of angular deviations, fewer than 0.3% of an identity's samples lie beyond $3\sigma_{y_u}$, so this margin moves forgotten embeddings outside the statistically plausible support of their original cone (derivation in Appendix A; margin analysis in Appendix E.4.5). Geometrically, $\mathcal{L}_{\text{cr}}$ carves a *dead space* around the cone axis, disrupting the semantic coherence of the erased identity, while retain-class geometry is protected by the preservation and global losses described below.

Angular Cone Preservation Loss: While $\mathcal{L}_{\text{cr}}$ disperses the target identity, we preserve the angular structure of all retained identities $k \notin Y_u$. Rather than matching probability distributions, we directly constrain each embedding's angular deviation from its class center before and after unlearning. Let $\theta_i = \arccos(\langle s_i, \mu_k \rangle)$ and $\tilde{\theta}_i = \arccos(\langle \tilde{s}_i, \mu_k \rangle)$ denote the deviations of the original and modified embeddings of identity k. The loss is

$$\mathcal{L}_{\text{cp}} = \frac{1}{|\mathcal{D}_{tr}|} \sum_{k \notin Y_u} \sum_{i \in \mathcal{S}_k} \frac{\Big(\tilde{\theta}_i - \theta_i\Big)^2}{\sigma_k^2}, \tag{6}$$

where σ_k^2 is the angular variance of cone k, which normalizes displacements relative to each identity's natural spread. By penalizing angular displacement rather than enforcing an exact distribution, the loss stays computationally simple while maintaining intra-class compactness and inter-class separability, so the retain set keeps its original geometry and recognition performance.

Global Loss: In addition to preserving individual identity cones, it is important to maintain the global geometric structure of the embedding space. Drastic modifications during unlearning may unintentionally lead to degraded discriminability among the retained identities. To mitigate this, we introduce a global distribution consistency loss that enforces stability of pairwise angular relationships in feature space. We define the Global loss as:

$$\mathcal{L}_g = \frac{1}{|\mathcal{D}_r|^2} \sum_{i,j \in \mathcal{D}_r} \big(\langle s_i, s_j \rangle - \langle \tilde{s}_i, \tilde{s}_j \rangle\big)^2,$$

where $\mathcal{D}_r$ denotes the retain-set samples, $\tilde{s}_i$ the corresponding embeddings after unlearning, and $\langle \cdot, \cdot \rangle$ cosine similarity. This regularizer preserves the pairwise angular structure of the retain set, preventing large-scale distortions while the repulsion and preservation losses act locally on individual identities. The complete objective is

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{cr}} + \lambda_2 \mathcal{L}_{\text{cp}} + \lambda_3 \mathcal{L}_{\text{g}}, \tag{7}$$

with λ_1=1, λ_2=1, λ_3=0.1 (sensitivity in Appendix E.4.2). The projection network p_σ is first trained with $\mathcal{L}_{\text{struct}}$ and then frozen; its candidate vectors define the cones used in $\mathcal{L}_{\text{cr}}$ and $\mathcal{L}_{\text{cp}}$..

Table 1: Quantitative comparison of FUSE with existing unlearning techniques. VA represents Verification Accuracy, and SR represents Semantic Residual. All experiments are performed on the ResNet-50 backbone initially trained with ArcFace loss. OG is the original model; Retrain is trained from scratch without the forget identities and serves as the exact-unlearning reference. $|V_f - V_{uc}|$ is a consistency check; forgetting is assessed by G_f.

Model	CASIA-Webface (Yi et al., 2014)							D-LORD (Manchanda et al., 2024)						
	MIAa	VA				SR		MIAa	VA				SR	
		V_r	V_f	V_{uc}	$V_f - V_{uc}$	G_r	G_f		V_r	V_f	V_{uc}	$V_f - V_{uc}$	G_r	G_f
OG	0.903	0.926	0.836	0.867	0.031	-	-	0.784	0.847	0.795	0.814	0.019	-	-
Retrain	0.508	0.914	0.812	0.821	0.009	0.650	0.335	0.503	0.836	0.802	0.811	0.009	0.529	0.301
FT	0.582	0.863	0.802	0.752	0.050	0.420	0.385	0.539	0.782	0.784	0.797	0.013	0.385	0.372
GA (Thudi et al., 2022)	0.621	0.834	0.518	0.774	0.256	0.488	0.643	0.692	0.749	0.519	0.638	0.119	0.362	0.374
L1-sparse (Jia et al., 2023)	0.572	0.795	0.647	0.721	0.074	0.540	0.648	0.493	0.751	0.558	0.630	0.072	0.442	0.401
RL (Hayase et al., 2020)	0.539	0.815	0.692	0.737	0.045	0.559	0.539	**0.512**	**0.792**	0.539	0.619	0.080	0.410	0.485
BS (Chen et al., 2023)	0.562	0.827	0.684	**0.793**	0.109	0.532	0.572	0.568	0.763	0.603	0.623	0.020	0.386	0.403
BE (Chen et al., 2023)	0.429	0.813	0.659	0.693	0.034	0.561	0.537	0.482	0.683	0.581	0.554	0.027	0.404	0.395
SalUn (Fan et al., 2024)	0.485	0.858	0.599	0.649	0.050	0.498	0.487	0.441	0.774	0.597	**0.672**	0.075	0.412	0.382
SG (Di et al., 2025)	**0.503**	0.849	0.680	0.728	0.048	0.525	0.461	0.381	0.753	0.606	0.642	0.036	0.481	0.342
FUSE (ours)	0.519	**0.851**	**0.739**	0.751	**0.012**	**0.632**	**0.391**	0.461	0.780	**0.635**	0.653	**0.018**	**0.538**	**0.328**

4 RESULTS

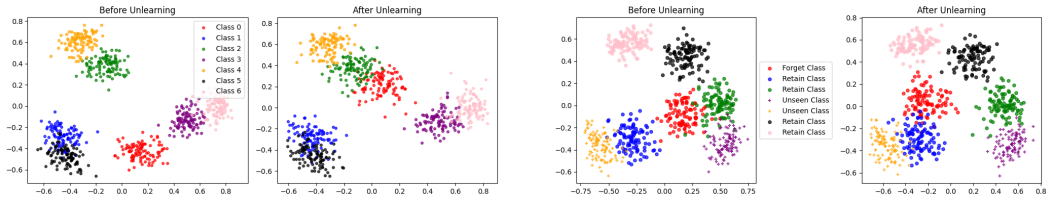
Evaluation Criteria: Datasets, baselines, and training details are provided in Appendix D. We evaluate unlearning effectiveness across three complementary dimensions. First, for membership analysis, we measure membership inference attack (MIA) accuracy (Carlini et al., 2022), where effective unlearning reduces performance to near-random (0.5). Second, for feature-level evaluation, we compute verification accuracy (VA) using feature similarity: V_r for retain classes, V_f for the forget set, and V_{uc} for unseen classes. We use $|V_f - V_{uc}|$ as a necessary but not sufficient condition for unlearning: a large gap reveals failure in either direction ($V_f > V_{uc}$: retained identity structure; $V_f < V_{uc}$: conspicuous over-forgetting), but a small gap cannot distinguish erasure from generalization (Section **??**). We therefore treat it only as a consistency check and rely on the Semantic Residual to separate methods that pass it.

Semantic Evaluation. To measure whether identity-related semantic information remains encoded in the unlearned embeddings, we use a diffusion model pretrained with the original embedding distribution as conditioning. We compare its score-matching error under the original and unlearned conditioning distributions: $L_{\text{score}} = \mathbb{E}_{c\sim p_{\text{orig}}}\left[\|\epsilon_\theta(x_t,t,c) - \epsilon_{\text{true}}\|^2\right]$, $\tilde{L}_{\text{score}} = \mathbb{E}_{c\sim p_{\text{new}}}\left[\|\epsilon_\theta(x_t,t,c) - \epsilon_{\text{true}}\|^2\right]$.

We define the normalized *Semantic Residual* (SR) as $SR = \min\left(1, \frac{L_{\text{score}}}{\tilde{L}_{\text{score}}}\right)$. Thus, $SR \to 1$ indicates that the unlearned embedding retains semantic guidance similar to the original distribution, whereas $SR \to 0$ indicates substantial loss of class-specific semantic information. We report SR for both forget (G_f) and retain (G_r) classes to jointly measure semantic leakage (See Appendix C.1).

Understanding FOG ambiguity: Evaluating face unlearning is hard not because of the algorithm, but because of the ambiguity introduced by the open-set nature of face recognition, so we first validate the FOG problem experimentally. In classification, forgetting should reduce recognition of the removed class; face recognition, however, is trained to generalize to unseen identities (high V_{uc}), so a forgotten identity behaving like an unseen one does not by itself indicate failed unlearning. Indeed, most methods, including FUSE, yield V_f close to V_{uc} (Table 1; a difference of only 0.012 on CASIA-WebFace and 0.018 on D-LORD for FUSE). Figure 3b shows the same ambiguity geometrically: forgotten identities leave their original regions but still form coherent clusters resembling unseen identities, with comparable angular spread (Appendix E.5).(Appendix E.5). Identical verification scores may thus reflect either genuine erasure or preserved identity information. Because verification measures discriminability rather than information removal, it cannot certify unlearning, motivating a semantic evaluation that measures residual identity information.

Quantitative Results: Table 1 compares unlearning methods against Retrain, the exact-unlearning reference (WebFace4M and MS1Mv3 in Appendix E.1). OG shows high MIA and forget-set verification, indicating no forgetting. FT reduces MIA but damages retain semantics, GA and L1-sparse leave large V_f–V_{uc} gaps, and SalUn and SG still leak semantic information (higher G_f). FUSE comes closest to Retrain: near-random MIA, a $|V_f - V_{uc}|$ close to Retrain's (0.012 vs. 0.009 on CASIA-WebFace), and the best-matching semantic residuals on both datasets (G_r/G_f 0.632/0.391



(a) (b)

Figure 3: Feature Representation (a) t-SNE plots for visualization of neighbourhood class change. (b) Demonstrating the behaviour of unseen class structure before and after unlearning. ”Red” - forget set, ”o” - seen test set, ”+” - unseen class (yellow and purple).

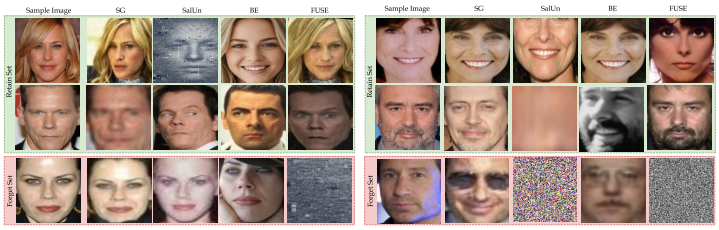


Figure 4: Showcasing reconstructed facial samples generated by IDiff-Face model.

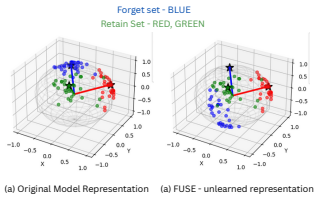


Figure 5: Representation before and after unlearning.

vs. 0.650/0.335 on CASIA-WebFace; 0.538/0.328 vs. 0.529/0.301 on D-LORD). FT matches G_f on CASIA-WebFace only by collapsing retain semantics (G_r = 0.420). FUSE also has the lowest post-inversion residual ($0.391 \rightarrow 0.423$; Appendix E.6) and takes 29 min per identity on a V100, comparable to SalUn (Appendix F.1).

Qualitative Results: Figure 4 illustrates the FOG problem visually. Despite appearing successful under verification, existing methods retain enough semantic information to reconstruct forgotten identities, whereas FUSE yields non-identifiable forget-set reconstructions while faithfully reconstructing the retain set. Semantic evaluation thus reveals leakage that verification metrics miss.

Feature representation: To examine the *dead space* created by FUSE, we visualize forget and retain embeddings with t-SNE (Figure 3a). Before unlearning, the forget class forms a compact cluster; after unlearning, its samples no longer occupy their original identity-specific region, while neighboring retain classes remain intact. FUSE preserves 70.1% of retained k-NN neighborhoods versus 37.8% for SalUn (Appendix E.5). Overlaying forgotten and unseen identities in the same projection (Figure 3b; $\circ$: seen/forgotten, $+$: unseen) shows that forgotten samples move into regions resembling unseen identities rather than persisting as a preserved cluster, consistent with forgotten identities becoming distributionally closer to never-seen identities. A view of forget and retain embeddings is provided in Appendix E.2.

Class Alignment: For forget-class samples, we quantify how strongly they still align with their original class distribution after unlearning. Since the classifier proxy is no longer available, we measure the average likelihood of the samples under the original class cone (Appendix C.2). Class alignment (CA) lies in $[0, 1]$ and decreases as samples move away from the original cone; lower CA on the forget set indicates more effective unlearning. As shown in Figure **??**, FUSE achieves the lowest CA.

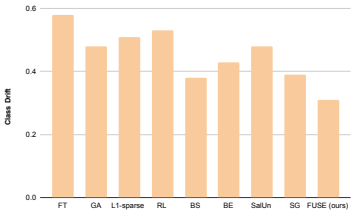
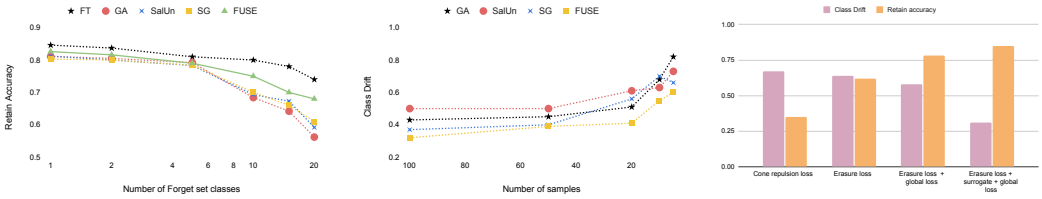


Figure 6: Class alignment of the forget set with its original distribution (lower is better).

Ablation Study: Figure 7 presents an ablation study on the effect of different factors in FUSE. In Figure 7a, as the number of forget classes increases, all methods show reduced retain accuracy; FUSE degrades more gracefully than GA, SalUn and SG, while FT retains higher accuracy only at the cost of collapsed retain-set semantics (Table 1). Figure 7b shows that class alignment



(a) Number of forget set classes (b) Sample size of forget class (c) Effect of various components

Figure 7: Ablation study of FUSE. (a) Retain accuracy vs. number of forget classes (b) Class alignment vs. forget class size (c) Effect of individual components for unlearning.

increases as the forget class shrinks, since fewer samples give a less stable cone estimate, while FUSE maintains the lowest alignment across sample sizes.

Figure 7c examines individual components. Cone repulsion alone yields partial forgetting but hurts retention; adding the preservation loss improves the balance, and the full framework with global and surrogate losses achieves the best trade-off, confirming that each component contributes. Adding an explicit random-target term to $\mathcal{L}_{\text{cr}}$ does not improve erasure (Appendix E.4.3). FUSE also maintains the highest retention on the top-10 nearest (hard) retain classes (Appendix E.3), and the same trends hold for ResNet-101 and ViT backbones (Appendix E.4.1).

Open Set Recognition: We evaluate the unlearned model on three open set datasets: LFW (Huang et al., 2008), CFP-FP (Sengupta et al., 2016) and Age-DB (Moschoglou et al., 2017), containing unseen identities from different domain. The experiment uses face recognition model trained with ArcFace loss on the CASIA dataset with ResNet50 backbone (margin = 0.5). During unlearning, we unlearn 10 classes from the training dataset and report performance on unseen datasets. The results from Table 2 demonstrates that unlearning classes with FUSE does not compromise the model's discriminative capability. On all three datasets, the verfication accuracy remains high.

Table 2: Identity-disjoint verification performance (verification accuracy in (%)) on three datasets.

	LFW	**CFP-FP**	**Age-DB**
Original	99.53	95.54	95.15
GA	83.25	91.36	93.01
L1-sparse	82.44	90.53	92.74
RL	91.12	91.46	94.73
BS	85.47	92.04	94.02
BE	87.97	92.58	93.59
SalUn	93.59	91.99	94.58
SG	98.17	92.16	**95.12**
FUSE (ours)	**99.31**	**92.97**	**95.12**

Reliability of Semantic Residual: To verify that SR reflects residual identity information rather than generator-specific artifacts, we compute it with three generators: IDiff-Face, Arc2Face, and Guided Diffusion (Table 3). SR scores are strongly correlated across generators (0.94 for IDiff-Face vs. Arc2Face, 0.87 for IDiff-Face vs. Guided Diffusion, and above 0.8 for the weakest pair; Table 7), yielding consistent method rankings. SR also correlates with identity recovery rate (r=0.89) and increases monotonically with injected identity strength (Appendix E.4.4), supporting it as a generator-agnostic measure of residual identity.

Table 3: Semantic residual on D-LORD computed with three generative models.

Model	**Semantic Residual**					
	IDiff-Face		**Arc2Face**		**Guided Diffusion**	
	G_r	G_f	G_r	G_f	G_r	G_f
FT	0.385	0.372	0.315	0.482	0.327	0.385
GA	0.362	0.374	0.450	0.505	0.363	0.338
L1-sparse	0.442	0.401	0.385	0.474	0.432	0.431
RL	0.410	0.485	0.483	0.414	0.385	0.453
BS	0.386	0.403	0.373	0.305	0.382	0.385
BE	0.404	0.395	0.528	0.494	0.428	0.385
SalUn	0.412	0.382	0.489	0.392	0.404	0.337
SG	0.481	0.342	0.510	0.292	0.452	0.402
FUSE (ours)	0.538	0.328	0.639	0.274	0.496	0.331

5 Conclusion

We explore a fundamental problem of machine unlearning for discriminative models defined as Forgetting-Or-Generalization (FOG) ambiguity problem. In discriminative models such as face recognition, ideal unlearning method must preserve high discriminability for unseen classes/forgotten classes, thereby maintaining generalization, while simultaneously erasing semantic traces of the target identity. Unlike prior approaches that primarily rely on verification assessment, which inherently reflects generalization rather than forgetting, we propose to evaluate unlearning through the persistence of semantic information of the target class in the feature space. We also introduce FUSE, a method that erases the geometric region of a target class, effectively removing residual semantics while safeguarding the representational integrity of identities. The experimental results demonstrate that FUSE preserves generalization capability while achieving semantic forgetting, establishing a novel and necessary direction for unlearning in discriminative models. Limitations and future directions are discussed in Appendix F.2.

AI USE STATEMENT

We used generative AI tools (large language models) for language editing, including rephrasing, shortening, and improving the clarity of the manuscript text. We did not use generative AI tools to generate research ideas, design the method, write code, run experiments, or produce results, figures, or tables. All AI-assisted text was reviewed and revised by the authors, who verified its technical accuracy against the underlying experiments. We take full responsibility for the final content of this work.

ETHICS STATEMENT

This work aims to strengthen privacy protection in face recognition by enabling verifiable removal of identities, in support of rights such as the GDPR "right to be forgotten." All experiments use publicly available face datasets, obtained and used in accordance with their respective licenses and terms of use; no new facial data were collected. We acknowledge that these datasets contain images of real individuals collected without explicit consent for this purpose, and we display reconstructed faces only to illustrate identity leakage. Our auditing framework and inversion experiments describe how residual identity information can be recovered from unlearned models. We present them to expose weaknesses in current unlearning methods and to provide a stronger evaluation standard, not to enable attacks. We note that Semantic Residual is an empirical measure and does not constitute a formal privacy guarantee, and it should not be used as the sole basis for regulatory compliance claims.

REPRODUCIBILITY STATEMENT

The FUSE formulation, including all loss functions and the complete training objective, is described in paper, with the statistical justification of the repulsion margin in Appendix A. The evaluation metrics, Semantic Residual and Class Drift, are fully specified in Appendix C.1 and Appendix C.2. Datasets, train/unseen identity splits, baseline configurations, network architectures, optimizers, and all hyperparameters (including $\lambda_1, \lambda_2, \lambda_3$ and α) are listed in Appendix D. Sensitivity analyses for loss weights, cone margin, and surrogate construction are provided in Appendices E.4.2, E.4.5, and E.4.6, and computational cost is reported in Appendix F.1. We will release source code upon acceptance, including the FUSE implementation, baseline scripts, and the Semantic Residual evaluation pipeline.

REFERENCES

Warnecke Alexander, Pirch Lukas, Wressnegger) Christian, and Rieck Konrad. Machine unlearning of features and labels. *Proceedings 2023 Network and Distributed System Security Symposium*, 2023. doi: 10.14722/ndss.2023.23087.

Chris Burt. Clearview AI tops 40 billion reference images in facial recognition database. Biometric Update, https://www.biometricupdate.com/202311/clearview-ai-tops-40-billion-reference-images-in-facial-recognition-database, November 2023.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramèr. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (SP)*, pp. 1897–1914, 2022. doi: 10.1109/SP46214.2022.9833649.

Sungmin Cha, Sungjun Cho, Dasol Hwang, Honglak Lee, Taesup Moon, and Moontae Lee. Learning to unlearn: Instance-wise unlearning for pre-trained classifiers. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 11186–11194, 2024.

Huiqiang Chen, Tianqing Zhu, Xin Yu, and Wanlei Zhou. Machine unlearning via null space calibration. IJCAI '24, 2024. ISBN 978-1-956792-04-1. doi: 10.24963/ijcai.2024/40. URL https://doi.org/10.24963/ijcai.2024/40.

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 the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7766–7775, June 2023.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/chen20j.html.

Dasol Choi and Dongbin Na. Towards machine unlearning benchmarks: Forgetting the personal identities in facial recognition systems. *arXiv preprint arXiv:2311.02240*, 2023.

Jiankang Deng, Jia Guo, Jing Yang, Niannan Xue, Irene Kotsia, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):5962–5979, 2022. doi: 10.1109/TPAMI.2021.3087709.

Zonglin Di, Sixie Yu, Yevgeniy Vorobeychik, and Yang Liu. Adversarial machine unlearning. In *International Conference on Learning Representations*, volume 2025, pp. 21612–21633, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/372593bd318ad8b34b3a8da77e20272b-Paper-Conference.pdf.

Jose Edgar Hernandez Cancino Estrada, Mauro Díaz Lupone, Žiga Emeršič, Vitomir Štruc, Peter Peer, and Darian Tomašević. Piu: Proximity-guided identity unlearning in id-conditioned diffusion models. *arXiv preprint arXiv:2605.22311*, 2026.

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 B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 53643–53673, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/ec4d2e436794d1bf55ca83f5ebb31887-Paper-Conference.pdf.

Yichen Gao, Altay Unal, Akshay Rangamani, and Zhihui Zhu. An illusion of unlearning? assessing machine unlearning through internal representations. In Emtiyaz Khan, Yingzhen Li, Arno Solin, and Aaditya Ramdas (eds.), *Proceedings of The 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pp. 4114–4122. PMLR, 02–05 May 2026.

Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9301–9309, 2020. doi: 10.1109/CVPR42600.2020.00932.

Tao Guo, Song Guo, Jiewei Zhang, Wenchao Xu, and Junxiao Wang. Efficient attribute unlearning: Towards selective removal of input attributes from feature representations. *arXiv preprint arXiv:2202.13295*, 2022.

Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *Computer Vision – ECCV 2016*, pp. 87–102, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46487-9.

Tomohiro Hayase, Suguru Yasutomi, and Takashi Katoh. Selective forgetting of deep networks at a finer level than samples. *arXiv preprint arXiv:2012.11849*, 2020.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9726–9735, 2020. doi: 10.1109/CVPR42600.2020.00975.

Tuan Hoang, Santu Rana, Sunil Gupta, and Svetha Venkatesh. Learn to unlearn for deep neural networks: Minimizing unlearning interference with gradient projection. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 4807–4816, 2024. doi: 10.1109/WACV57701.2024.00475.

Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *International Conference on Learning Representations*, 2023.

Yurim Jang, Jaeung Lee, Dohyun Kim, Jaemin Jo, and Simon S Woo. Suppression or deletion: A restoration-based representation-level analysis of machine unlearning. In *Proceedings of the ACM Web Conference 2026*, WWW '26, pp. 8505–8508, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400723070. doi: 10.1145/3774904.3792896. URL https://doi.org/10.1145/3774904.3792896.

Jinghan Jia, Jiancheng Liu, Parikshit Ram, Yuguang Yao, Gaowen Liu, Yang Liu, PRANAY SHARMA, and Sijia Liu. Model sparsity can simplify machine unlearning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 51584–51605. Curran Associates, Inc., 2023. doi: 10.52202/075280-2246.

Sangamesh Kodge, Gobinda Saha, and Kaushik Roy. Deep unlearning: Fast and efficient gradient-free class forgetting. *Transactions on Machine Learning Research*, 3(4):6, 2024.

Na Li, Chunyi Zhou, Yansong Gao, Hui Chen, Zhi Zhang, Boyu Kuang, and Anmin Fu. Machine unlearning: Taxonomy, metrics, applications, challenges, and prospects. *IEEE Transactions on Neural Networks and Learning Systems*, 36(8):13709–13729, 2025.

Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6738–6746, 2017. doi: 10.1109/CVPR.2017.713.

Sunny Manchanda, Kaushik Bhagwatkar, Kavita Balutia, Shivang Agarwal, Jyoti Chaudhary, Muskan Dosi, Chiranjeev Chiranjeev, Mayank Vatsa, and Richa Singh. D-lord: Dysl-ai database for low-resolution disguised face recognition. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 6(2):147–157, 2024. doi: 10.1109/TBIOM.2023.3306703.

Stylianos Moschoglou, Athanasios Papaioannou, Christos Sagonas, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. Agedb: The first manually collected, in-the-wild age database. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1997–2005, 2017. doi: 10.1109/CVPRW.2017.250.

Thanh Tam Nguyen, Thanh Trung Huynh, Zhao Ren, Phi Le Nguyen, Alan Wee-Chung Liew, Hongzhi Yin, and Quoc Viet Hung Nguyen. A survey of machine unlearning. *ACM Transactions on Intelligent Systems and Technology*, 16(5):1–46, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Walter J. Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E. Boult. Toward open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7): 1757–1772, 2013. doi: 10.1109/TPAMI.2012.256.

Ayush Sekhari, Jayadev Acharya, Gautam Kamath, and Ananda Theertha Suresh. Remember what you want to forget: Algorithms for machine unlearning. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 18075–18086. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/9627c45df543c816a3ddf2d8ea686a99-Paper.pdf.

Soumyadip Sengupta, Jun-Cheng Chen, Carlos Castillo, Vishal M. Patel, Rama Chellappa, and David W. Jacobs. Frontal to profile face verification in the wild. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1–9, 2016. doi: 10.1109/WACV.2016.7477558.

Shaofei Shen, Chenhao Zhang, Yawen Zhao, Alina Bialkowski, Weitong Chen, and Miao Xu. Label-agnostic forgetting: A supervision-free unlearning in deep models. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 54217–54235, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/ee1f0da706829d7f198eac0edaacc338-Paper-Conference.pdf.

Ayush K. Tarun, Vikram S. Chundawat, Murari Mandal, and Mohan Kankanhalli. Fast yet effective machine unlearning. *IEEE Transactions on Neural Networks and Learning Systems*, 35(9): 13046–13055, 2024. doi: 10.1109/TNNLS.2023.3266233.

Anvith Thudi, Gabriel Deza, Varun Chandrasekaran, and Nicolas Papernot. Unrolling SGD: Understanding Factors Influencing Machine Unlearning . In *2022 IEEE 7th European Symposium on Security and Privacy (EuroSP)*, pp. 303–319, Los Alamitos, CA, USA, June 2022. IEEE Computer Society. doi: 10.1109/EuroSP53844.2022.00027. URL https://doi.ieeecomputersociety.org/10.1109/EuroSP53844.2022.00027.

Pratyush Ranjan Tiwari, Dhruv Agarwal, Prakhar Jain, Swagam Dasgupta, Preetha Datta, Vineet Reddy, and Debayan Gupta. India's "aadhaar" biometric id: Structure, security, and vulnerabilities. In Ittay Eyal and Juan Garay (eds.), *Financial Cryptography and Data Security*, pp. 672–693, Cham, 2022. Springer International Publishing.

Zeynel Tok. The nearest target is the wrong one: Target separation in arc2face identity unlearning. *arXiv preprint arXiv:2608.30087*, 2026.

Unique Identification Authority of India. Aadhaar. https://uidai.gov.in/en.

Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5265–5274, 2018. doi: 10.1109/CVPR.2018.00552.

Yihan Wang, Yiwei Lu, Guojun Zhang, Franziska Boenisch, Adam Dziedzic, Yaoliang Yu, and Xiao-Shan Gao. Alignment calibration: Machine unlearning for contrastive learning under auditing. In *International Conference on Machine Learning, Next Generation of AI Safety Workshop*, 2024.

Yihan Wang, Yiwei Lu, Guojun Zhang, Franziska Boenisch, Adam Dziedzic, Yaoliang Yu, and Xiao-Shan Gao. MUC: Machine unlearning for contrastive learning with black-box evaluation. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=F9pjSDvuM9.

Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014.

Mikhail Zakharov. Face identity unlearning for retrieval via embedding dispersion. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, pp. 1277–1286, March 2026.

Yu Zhou, Dian Zheng, Qijie Mo, Renjie Lu, Kun-Yu Lin, and Wei-Shi Zheng. Decoupled distillation to erase: A general unlearning method for any class-centric tasks. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20350–20359, 2025. doi: 10.1109/CVPR52734.2025.01895.

Zheng Zhu, Guan Huang, Jiankang Deng, Yun Ye, Junjie Huang, Xinze Chen, Jiagang Zhu, Tian Yang, Dalong Du, Jiwen Lu, and Jie Zhou. Webface260m: A benchmark for million-scale deep face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):2627–2644, 2023. doi: 10.1109/TPAMI.2022.3169734.

Zijun Zhuang and Hongtao Lu. Coneface: Approximate pairwise loss for face recognition. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2949–2953, 2022. doi: 10.1109/ICASSP43922.2022.9746847.

Ünsal Öztürk, Vedrana Krivokuća Hahn, Sushil Bhattacharjee, and Sébastien Marcel. Damnatio memoriae: Adversarially and selectively forgetting identities in the embedding space of face recognition models, 2026. URL https://arxiv.org/abs/2609.27115.

A Statistical Justification of the Repulsion Margin

The repulsion margin $\alpha\sigma_{Y_u}$ in Eq. (5) is motivated by the 3-sigma rule of Gaussian distributions. For $X \sim \mathcal{N}(\mu, \sigma^2)$, approximately 68%, 95%, and 99.7% of the probability mass lies within $\mu \pm \sigma$, $\mu \pm 2\sigma$, and $\mu \pm 3\sigma$, respectively. The probability of observing a sample outside $\mu \pm 3\sigma$ is therefore below 0.3%, which provides a natural threshold for separating inliers from outliers in a Gaussian-like distribution.

In our setting, we model the angular distribution of embeddings around the class center μ_y with a wrapped normal distribution on the hypersphere, where σ_y captures the angular spread of the identity cone. Since the angular deviation $\theta_i = \arccos(\langle s_i, \mu_k \rangle)$ is non-negative, it behaves approximately as a half-normal variable $|X|$ with $X \sim \mathcal{N}(0, \sigma_k^2)$. For this distribution, the root-mean-square value of $|X|$ equals σ_k, which matches our RMS definition of the cone spread, and $P(\theta_i > 3\sigma_k) \approx 0.27\%$.

Enforcing a repulsion margin of at least $3\sigma_k$ thus pushes forgotten embeddings into regions of the hypersphere that lie outside the statistically plausible support of their original distribution. With high probability, the resulting embeddings are no longer consistent with the identity they once represented. At the same time, retain-class embeddings are protected by the cone-preservation and global-consistency losses, so the discriminative geometry of the feature space remains intact.

This justification is conditional on the Gaussian angular assumption. Accordingly, we do not treat $\alpha = 3$ as a guarantee but as a statistically motivated operating point, and empirically, 3σ coincides with the minimum forget-set residual, with stable performance within $[2.5\sigma, 3.5\sigma]$ (Appendix E.4.5).

B Targeted Repulsion Loss

FUSE is intentionally designed to dissolve the identity geometry rather than relocate forgotten samples to a fixed region, since forcing them into a predetermined cluster may inadvertently create a new identity-like structure. Our original goal is to erase identity-specific information by removing the semantic coherence of the original cone while preserving the geometry of all retained classes.

To further understand the geometric behaviour, we conducted an additional experiment where we augment FUSE with a uniform-diffusion loss. This term encourages the forgotten embeddings to spread toward an isotropic, structureless distribution on the hypersphere, that aligns with the motivation of FUSE:

$$L_{\text{targeted}} = \frac{1}{|S_{Y_u}|} \sum_i \theta_i, \quad \theta_i = \arccos\left(\langle \hat{s}_i, u_i \rangle\right)$$

where each u_i is a randomly sampled unit vector drawn uniformly from the hypersphere. This controlled diffusion prevents forgotten embeddings from drifting arbitrarily, ensuring that they migrate toward a stable, non-informative, and identity-free region of the feature space.

Conceptually, this aligns with the purpose of targeted unlearning: instead of forming a new cluster, forgotten samples are dissolved into a uniform noise landscape, their identity structure becomes irrecoverable, and they no longer interfere with retained identities. Empirically, we observe that adding this loss produces a structured form of forgetting while maintaining the stability of the retained classes.

C Detailed Evaluation Protocol

We use $|V_f - V_{uc}|$ as a necessary but not sufficient condition for unlearning: a large gap reveals failure in either direction, since $V_f > V_{uc}$ indicates retained identity-specific structure, while $V_f < V_{uc}$ indicates over-forgetting that makes the removed identity conspicuous, but a small gap cannot distinguish genuine erasure from generalization. We therefore treat it only as a consistency check and rely on the Semantic Residual to separate methods that pass it.

C.1 SEMANTIC RESIDUAL

The Semantic Residual (SR) measures how much class-specific semantic information remains usable in an embedding after unlearning. The key intuition is that a diffusion model conditioned on an informative embedding should predict the noise trajectory accurately. When the conditioning embedding no longer captures the corresponding class semantics, the quality of conditional noise prediction should degrade.

We use a denoising diffusion model with a UNet backbone, pretrained using embeddings from the original recognition model. For a timestep t, a clean sample x_0 is corrupted through the standard forward diffusion process to obtain x_t with ground-truth noise ϵ_{true}. Given a conditioning embedding c, the diffusion model predicts $\epsilon_\theta(x_t, t, c)$.

For embeddings sampled from the original distribution $c \sim p_{\text{orig}}$, the score-matching error is

$$L_{\text{score}} = \mathbb{E}_{c\sim p_{\text{orig}}}\left[\|\epsilon_\theta(x_t, t, c) - \epsilon_{\text{true}}\|_2^2\right]. \tag{8}$$

For embeddings produced by the unlearned recognition model, $c \sim p_{\text{new}}$, we compute

$$\tilde{L}_{\text{score}} = \mathbb{E}_{c\sim p_{\text{new}}}\left[\|\epsilon_\theta(x_t, t, c) - \epsilon_{\text{true}}\|_2^2\right]. \tag{9}$$

We then define

$$SR = \min\left(1, \frac{L_{\text{score}}}{\tilde{L}_{\text{score}}}\right). \tag{10}$$

The normalization makes $SR \in (0, 1]$ and expresses the effectiveness of the unlearned conditioning relative to the original conditioning.

An SR value close to 1 means that the unlearned embedding remains nearly as useful for conditional noise prediction as the original embedding, indicating strong semantic retention and, for the forget class, potential semantic leakage. Conversely, an SR value close to 0 indicates a large increase in score-matching error and therefore a substantial loss of useful class-specific conditioning information. Intermediate values represent partial retention.

This interpretation follows from the connection between diffusion noise prediction and conditional score estimation:

$$\epsilon_\theta(x_t, t, c) \approx -\sqrt{1-\bar{\alpha}_t}\,\nabla_x \log p(x_t \mid c). \tag{11}$$

Therefore, degradation in noise prediction reflects a reduced ability of the conditioning embedding to specify the corresponding conditional data distribution. Importantly, SR is computed for both the forget set (G_f) and retain set (G_r). A reduction in SR for G_f indicates removal of class-specific semantic information, while stable SR for G_r indicates preservation of non-target semantics.

C.2 CLASS ALIGNMENT

We additionally measure the geometric drift of forgotten embeddings relative to their original class distribution. Let μ_A denote the modal direction of class A on the hypersphere and σ_A its angular spread. For an embedding z, the angular distance from the original class center is

$$\theta(z, A) = \arccos\left(\langle z, \mu_A\rangle\right). \tag{12}$$

Assuming a cone-shaped class distribution, we define the normalized likelihood of z under the original class distribution as

$$\text{LL}(z; A) = \exp\left(-\frac{\theta(z, A)^2}{2\sigma_A^2}\right). \tag{13}$$

A high likelihood indicates that the embedding remains close to the original class direction and angular spread, whereas a low likelihood indicates that the embedding has drifted away from the original class cone.

For a set of N_A samples from class A, we aggregate the per-sample likelihoods as

$$\overline{\text{LL}}(A) = \frac{1}{N_A}\sum_z \text{LL}(z; A). \tag{14}$$

Values close to 1 indicate that the embeddings retain a strong geometric trace of the original class, whereas values close to 0 indicate substantial angular drift. For the forget set, this provides a complementary measure to MIA, VA, and SR: rather than asking whether membership or semantic information remains detectable, class drift directly measures how far the representation has moved from the original identity geometry. Lower CA is desirable on the forget set, and higher CA on the retain set.

D Implementation Details

Datasets: We evaluate FUSE on four large-scale publicly available datasets: CASIA-WebFace (Yi et al., 2014), WebFace4M (Zhu et al., 2023), MS1Mv3 (Guo et al., 2016) and D-LORD (Manchanda et al., 2024).

Baselines: We use existing unlearning methods trained on same protocol for fair comparison. OG denotes the original model before unlearning, trained on the full training set including the forget identities. Retrain denotes a model trained from scratch on the retain identities only, with the same architecture and settings as OG. We leverage Fine Tuning (Golatkar et al., 2020), Gradient Ascent (GA) (Thudi et al., 2022), l_1-sparse (Jia et al., 2023), Random Labeling (RL) (Hayase et al., 2020), Boundary Shrinking (BS), Boundary Expansion (BE) (Chen et al., 2023), SalUn (Fan et al., 2024) and SG-unlearn (Di et al., 2025) for comparison. For each classification based unlearning, classifier level changes are made while the evaluation is performed using feature level to compare efficiently with FUSE. For fine-tuning (FT), we use representative normal images from non-target set and finetune model for 20 epochs.

Training details: Unless stated otherwise, we use a ResNet-50 backbone trained with ArcFace loss (ResNet-101 and ViT results in Appendix E.4.1). Models are trained separately on each of the four datasets. For CASIA-WebFace, 10,000 classes are used for training, and 575 are kept for unseen class testing. For D-LORD, 1600 are used for training, while 600 are used for unseen testing. The FR model is trained using a 128 batch size, a learning rate of $1e-4$, and using SGD as optimizer.

For training FUSE, training is performed for 20 epochs, and the forget set size varies for different experiments, as shown in the ablation study. For surrogate formation, the face image is divided into two parts, the upper face and the lower face. Repulsion Factor (α). We recommend values in the range $\alpha \in [2.5, 3.5]$ to enforce meaningful separation while avoiding over-displacement that could distort nearby cones. Baselines are also trained with the same parameters for fair comparison. A balanced configuration is $\lambda_1 = 1.0$ (repulsion as primary driver), $\lambda_2 = 1.0$ (equal importance to preservation), and $\lambda_3 = 0.1$ (light regularization for global consistency).

For semantic residual (SR) evaluation, we use IDiff-Face, taking class embeddings as conditions. We compute G_f on held-out samples of the forget identities and G_r on held-out samples of the retain identities, take a random t for denoising and compute error. The training of this diffusion is done using an initial ArcFace model without any unlearning, the training parameters are the same as in the main paper.

Surrogate representation retrieval: We build a vector database by extracting ℓ_2-normalized embeddings from a frozen face recognition backbone, including full-face embeddings and part-based embeddings obtained via fixed spatial masking of upper and lower facial regions followed by global average pooling. All embeddings are indexed using an approximate nearest neighbor (ANN) structure with cosine similarity as the distance metric. This indexing enables efficient retrieval of identity-specific samples, region embeddings, and semantically close class prototypes for surrogate construction.

Structural trace Identification: We divide face image into two parts: upper and lower (cropping and 50% masking). Its embeddings are extracted from frozen arcface model, which is used to form candidate vectors. We use a 3-layer MLP (512 dimension output) as p_σ to extract meaningful fused embedding from part based images.

Table 4: Quantitative comparison of FUSE on WebFace4m and MS1Mv3 datasets.

Model	**WebFace4M** (Zhu et al., 2023)							**MS1Mv3** (Guo et al., 2016)						
		VA				SR			VA				SR	
	MIA_a	V_r	V_f	V_{uc}	$V_f - V_{uc}$	G_r	G_f	MIA_a	V_r	V_f	V_{uc}	$V_f - V_{uc}$	G_r	G_f
OG	0.88	0.89	0.81	0.82	0.01	-	-	0.89	0.91	0.92	0.88	0.04	-	-
Retrain	0.51	0.87	0.80	0.81	0.01	0.45	0.40	0.50	0.89	0.86	0.87	0.01	0.48	0.39
FT	0.53	0.79	0.61	0.76	0.15	0.21	0.57	0.46	0.72	0.80	0.78	0.02	0.42	0.37
GA (Thudi et al., 2022)	0.64	0.75	0.65	0.78	0.13	0.25	0.44	**0.51**	0.77	0.64	0.79	0.15	0.45	0.51
L1-sparse (Jia et al., 2023)	0.60	0.78	0.69	0.74	0.05	0.31	0.48	0.55	0.78	0.72	**0.81**	0.09	0.49	0.47
RL (Hayase et al., 2020)	0.70	0.77	0.67	0.69	**0.02**	0.35	0.51	0.49	0.78	0.61	0.72	0.11	0.48	0.48
BS (Chen et al., 2023)	0.61	0.81	0.70	0.76	0.06	0.28	0.55	0.67	0.81	0.76	0.76	0.00	0.39	0.40
BE (Chen et al., 2023)	0.44	0.80	0.71	0.69	**0.02**	0.38	0.49	0.52	0.75	0.78	0.75	0.03	0.50	0.41
SalUn (Fan et al., 2024)	0.48	0.77	0.68	0.71	0.03	0.33	0.51	0.48	0.78	**0.81**	0.72	0.09	0.51	0.36
SG (Di et al., 2025)	**0.52**	0.80	0.67	0.73	0.06	0.38	0.43	0.41	0.76	0.73	0.78	0.05	0.47	0.33
FUSE (ours)	0.53	**0.83**	**0.75**	**0.79**	0.04	**0.43**	**0.42**	0.47	**0.85**	0.80	0.80	**0.00**	**0.58**	**0.29**

E RESULTS

E.1 RESULTS ON LARGE-SCALE DATASETS

Table 4 reports results on WebFace4M and MS1Mv3, which contain substantially more identities than CASIA-WebFace and D-LORD and therefore test whether FUSE scales to denser embedding spaces.

Retention and generalization. Among approximate unlearning methods, FUSE comes closest to Retrain in retain-set verification on both datasets (V_r: 0.83 vs. 0.87 on WebFace4M and 0.85 vs. 0.89 on MS1Mv3). It also yields the unseen-identity verification closest to Retrain on WebFace4M (V_{uc}: 0.79 vs. 0.81), indicating that erasing the target identity does not compromise open-set generalization. In contrast, FT and GA incur the largest retain-set degradation (V_r as low as 0.72 and 0.75, respectively).

Forgetting-or-generalization gap. On MS1Mv3, FUSE produces identical verification on forgotten and unseen identities ($V_f = V_{uc} = 0.80$), matched only by BS, whose forget and retain accuracies are lower. On WebFace4M, the FUSE gap (0.04) is comparable to RL and BE (0.02), while GA and FT exhibit the largest gaps (0.13–0.15). Notably, SalUn on MS1Mv3 verifies forgotten identities *better* than unseen ones (V_f = 0.81 vs. V_{uc} = 0.72), which indicates that identity-specific structure of the forget set persists after unlearning. As discussed in Section **??**, small gaps alone do not certify forgetting, which motivates the semantic analysis below.

Semantic residual. On WebFace4M, FUSE is the closest approximate method to Retrain in semantic residual (G_r/G_f: 0.43/0.42 vs. 0.45/0.40). On MS1Mv3, FUSE's forget-set residual falls below Retrain's (0.29 vs. 0.39) while retaining more retain-set semantics (0.58 vs. 0.48). On WebFace4M, the forget-set margin over SG is small (0.42 vs. 0.43), but FUSE retains considerably more retain-set semantics (0.43 vs. 0.38). Several baselines show the opposite of the desired behavior. GA and RL on MS1Mv3 leave forget-set residuals equal to or higher than their retain-set residuals ($G_f \geq G_r$). FT on WebFace4M collapses retain semantics (G_r = 0.21) while leaving the highest forget-set residual (G_f = 0.57), so its apparent forgetting is driven by global representation damage rather than targeted erasure.

Membership inference. FUSE keeps membership inference close to random on both datasets (0.53 and 0.47), compared with 0.88–0.89 for the original model and 0.50–0.51 for Retrain. Some baselines obtain lower MIA (e.g., BE with 0.44 and SG with 0.41), but values below 0.5 do not indicate stronger privacy. Values substantially below chance can themselves reveal a systematic difference between forgotten and non-member samples.

Overall, the trends on WebFace4M and MS1Mv3 are consistent with those on CASIA-WebFace and D-LORD (Table 1): FUSE provides the best balance between semantic erasure of the forget set and preservation of retain-set and open-set performance, and this advantage becomes more pronounced as dataset scale increases.

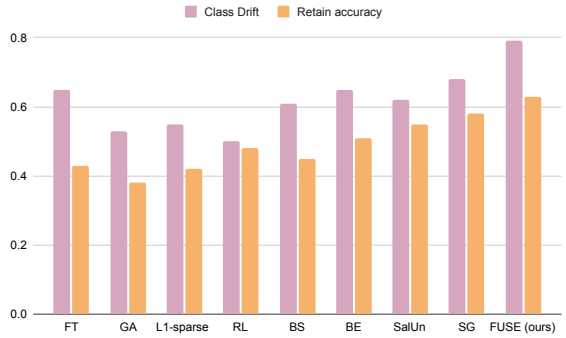
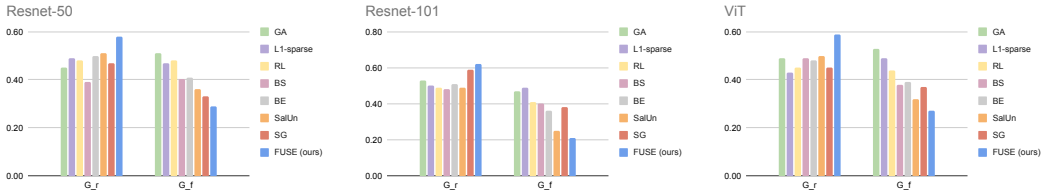


Figure 8: Class alignment and retain accuracy on the top-10 hard retain classes (higher is better for both).



(a) (b) (c)

Figure 9: Comparison of unlearning performance across three backbone architectures: ResNet-50 (a), ResNet-101 (b), and ViT (c). Each bar plot reports the semantic residual for the retained (G_r) and forgotten (G_f) classes for seven unlearning baselines and FUSE.

E.2 Feature Representation Before and After Unlearning

We visualize the embedding distributions of forget and retain samples before and after unlearning to assess FUSE's impact on class representation (Figure 5). Before unlearning, samples of each identity form compact clusters around their class centers, reflecting strong discriminability. After FUSE, retain-set samples remain tightly clustered around their original means, showing high retention. Forget-set samples still form coherent groups, as expected of any identity under an open-set encoder, but these groups are no longer aligned with their original class centers. FUSE therefore achieves unlearning at the representation level: it removes the identity-specific alignment of the forgotten class while preserving the discriminative capacity of the embedding space.

E.3 Hard Set Retention

Owing to the fine-grained nature of face recognition, unlearning can disproportionately affect closely related classes. To study this, we construct a hard set by retrieving the top-k ($k = 10$) nearest classes from the training set using cosine similarity by computed using a ArcFace trained feature extractor. We evaluate both the average retain accuracy and the class drift (higher values are better). Results (Figure 8) show that retaining performance on the hard set is significantly more challenging compared to the others. Unlearning one class in feature space, distort the feature geometry of closely retained class. However, due to FUSE's geometry-aware preservation during unlearning, it consistently retains higher performance, even on the hard set, demonstrating robustness in fine-grained and compact class distribution scenarios.

E.4 Ablation Studies

E.4.1 Ablation on various backbones

As shown in Figure 9 Across all three backbone architectures—ResNet-50 (a), ResNet-101 (b), and ViT (c)—a consistent trend emerges in terms of semantic residuals for both retained and forgotten

Table 5: Ablation study on loss weights (MS1Mv3).

Hyperparameter Configuration	**semantic residuals**	
	G_r	G_f
$\lambda_1 = 1, \lambda_2 = 1, \lambda_3 = 0.1$	0.58	0.29
$\lambda_1 = 0.5, \lambda_2 = 1, \lambda_3 = 0.2$	0.56	0.48
$\lambda_1 = 1, \lambda_2 = 0.5, \lambda_3 = 0.1$	0.46	0.33
$\lambda_1 = 0.8, \lambda_2 = 1, \lambda_3 = 0.1$	0.56	0.44
$\lambda_1 = 1, \lambda_2 = 0.8, \lambda_3 = 0.1$	0.50	0.32
$\lambda_1 = 1, \lambda_2 = 1, \lambda_3 = 0.2$	0.54	0.29

classes. For the retained set (G_r), FUSE achieves the highest or near-highest values among all baselines, indicating that identity information for non-forgotten classes is well preserved. In contrast, for the forgotten set (G_f), FUSE produces the lowest residuals across all architectures, demonstrating that it removes class-specific information more effectively than competing methods. Other baselines show varying degrees of unlearning, but none simultaneously achieve high G_r and low G_f as reliably as FUSE. This consistent separation highlights FUSE's strong ability to maintain utility while maximally erasing information from the forgotten identity classes.

E.4.2 Hyperparameter sensitivity

Hyperparameter Sensitivity: To analyze the model's behavior and understand the contribution of each loss component in FUSE, we conduct experiments by varying the hyperparameters, which weight the different loss terms. Using the MS1MV3 dataset, we evaluate the model across multiple configurations. In the paper, we adopt the setting $\lambda_1 = 1$, $\lambda_2 = 1$, $\lambda_3 = 0.1$, and the experimental results support this choice. As shown in Table 5, when λ_1 is reduced, the repulsion applied to forgotten identities weakens. As a result, the retain-set geometry remains mostly unaffected, but forgotten identities remain close to their original cone, yielding a high G_f. Conversely, when λ_2 is reduced, the retain set drifts away from its original geometric structure, indicating lower preservation. These observations show that an appropriate balance between the loss weights is necessary to achieve effective forgetting while preserving the retain set structure.

E.4.3 Ablation: Random-Target Repulsion

$\mathcal{L}_{\text{cr}}$ only requires forgotten embeddings to leave their original identity cone and does not prescribe where they should go. To test whether an explicit destination helps, we compare it with a variant that additionally pulls each forgotten embedding toward a random target:

$$\mathcal{L}_{\text{rt}} = \frac{1}{|\mathcal{S}_{Y_u}|} \sum_{i \in \mathcal{S}_{Y_u}} \arccos\left(\langle \tilde{s}_i, u_i \rangle\right), \tag{15}$$

where each u_i is drawn once, uniformly from $\mathbb{S}^{d-1}$, and kept fixed during training. This relocates forgotten samples toward independent random directions instead of only excluding them from their original cone.

Table 6: Cone repulsion vs. random-target repulsion (CASIA-WebFace, ResNet-50, ArcFace).

Objective	V_r	V_f	G_r	$G_f \downarrow$
$\mathcal{L}_{\text{cr}}$ (FUSE)	0.85	0.74	0.63	**0.39**
$\mathcal{L}_{\text{cr}} + \mathcal{L}_{\text{rt}}$	0.84	0.75	0.64	0.41

As shown in Table 6, the two objectives perform comparably, with all metrics within 0.02, but the random-target variant leaves a slightly higher forget-set residual (G_f: 0.41 vs. 0.39). Prescribing a destination adds a constraint without improving erasure, consistent with observations that target choice strongly affects identity unlearning in generative models (Tok, 2026), which supports the design choice in FUSE of removing forgotten embeddings from their identity support without fixing where they relocate. Near-orthogonal per-sample targets are effective when the goal is unlinkability

Table 7: Correlation of SR scores computed using different generative models. High correlations indicate that SR is stable across generators and not tied to a specific reconstruction model.

	Arc2Face	**Guided Diffusion**
IDiff-Face	0.94	0.87
Arc2Face	1.00	0.84
Guided Diffusion	0.84	1.00

Table 8: Effect of the cone margin $\alpha\sigma$ on semantic residual (D-LORD). Retain-set residual G_r is stable across margins, whereas forget-set residual G_f drops sharply from 2σ to 2.5σ, reaches its minimum at 3σ, and rises moderately beyond 3.5σ.

Metric	2σ	2.5σ	3σ	3.5σ	4σ
G_r	0.52	0.55	0.53	0.54	0.54
G_f	0.48	0.35	0.32	0.33	0.38

(Ünsal Öztürk et al., 2026), but under our goal of unseen-like behavior they offer no benefit over cone repulsion.

E.4.4 Semantic Residual Validity

To further validate SR, we measure whether higher SR values correspond to stronger recoverable identity information. We define the Identity Recovery Rate (IRR) as:

$$IRR = \frac{\text{correct matches}}{\text{total reconstructed}}$$

where correct matches denote reconstructed samples successfully matched to their original identities using a face recognition model.

Figure 10a shows the relationship between SR and IRR across multiple unlearning methods. We observe a strong positive Pearson correlation of $r = 0.89$, indicating that higher SR values consistently correspond to greater identity recoverability. Importantly, methods with lower SR, such as FUSE and FT, exhibit substantially lower recovery rates, whereas methods with higher SR retain significantly more identifiable information. This result supports SR as a meaningful proxy for residual biometric leakage.

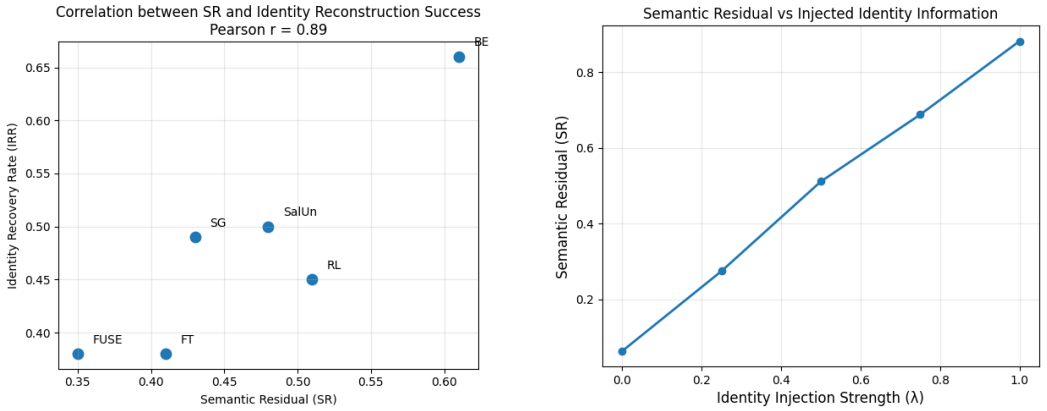
Finally, we perform a controlled experiment to determine whether SR responds predictably to varying amounts of injected identity information. Starting from identity-neutral generations, we progressively inject identity embeddings with increasing strength λ. Figure 10b shows that SR increases monotonically as the injected identity strength increases.

Specifically, SR rises from approximately 0.06 in the absence of injected identity information to nearly 0.88 at full injection strength. The smooth monotonic trend demonstrates that SR scales proportionally with the amount of identity information present in generated samples. This behavior confirms that SR measures genuine identity leakage rather than random generator variance or reconstruction artifacts.

Overall, these experiments demonstrate that SR is: (i) stable across generative models, (ii) strongly correlated with identity recoverability, and (iii) sensitive to controlled variations in injected identity information. Together, these findings establish SR as a reliable metric for quantifying residual biometric identity leakage after unlearning.

E.4.5 Sensitivity to Cone Margin Selection

Our method adopts a 3σ cone margin as a statistically grounded approximation of the identity support region under a Gaussian assumption in angular embedding space. To assess how this choice affects unlearning, we vary the margin from 2σ to 4σ (Table 8).



(a) Correlation between SR and Identity Recovery Rate (IRR).

(b) SR under controlled identity injection strength λ.

Figure 10: Reliability analysis of Semantic Residual (SR). Left: Higher SR strongly correlates with increased identity reconstruction success ($r = 0.89$). Right: SR increases monotonically with injected identity strength, indicating that SR reliably tracks residual identity information.

Table 9: Comparison between the proposed upper–lower surrogate construction and fine-grained face parsing regions.

Region Strategy	V_r	V_f	V_{uc}	G_r	G_f
Upper–Lower (FUSE)	0.83	0.73	0.75	0.63	0.38
Fine Regions (eyes/nose/mouth)	0.82	0.75	0.78	0.67	0.36

The retain-set residual is stable across all margins (G_r between 0.52 and 0.55), confirming that the preservation and global losses protect retained identities independently of the repulsion margin. The forget-set residual, in contrast, depends on the margin. At 2σ, G_f remains high (0.48), a 50% relative increase over 3σ. This is expected from the half-normal model in Appendix A: roughly 95% of an identity's samples lie within 2σ of its center, so a 2σ margin moves forgotten embeddings only to the edge of the identity's high-density region, where they remain semantically consistent with it. Increasing the margin to 2.5σ removes most of this residual ($G_f = 0.35$), and G_f reaches its minimum at 3σ (0.32). Beyond 3.5σ, G_f rises moderately (0.38 at 4σ). We attribute this to the larger displacement required, which conflicts more strongly with the preservation losses on neighboring identities, so fewer forget samples reach the margin during optimization.

The margin therefore acts as a lower bound on effective forgetting rather than a freely tunable parameter. Margins below 2.5σ are insufficient, while performance within $[2.5\sigma, 3.5\sigma]$ varies by at most 0.03 in G_f and degrades gracefully beyond it. The statistically motivated choice of 3σ lies at the center of this range and coincides with the empirical optimum.

E.4.6 Face parsing and surrogate representation strategy.

We additionally evaluate a finer-grained surrogate construction strategy using facial regions obtained through face parsing (e.g., eyes, nose, and mouth). While this approach introduces more localized feature combinations, we observe only marginal differences in both verification and unlearning metrics compared to the simpler upper–lower partition used in FUSE, as shown in Table 9. The fine-grained strategy provides slightly improved forgetting performance (G_f: 0.36 vs. 0.38), whereas the upper–lower partition achieves marginally better retain verification (V_r: 0.83 vs. 0.82). Overall, both approaches remain highly comparable, with differences within 0.01–0.02 across metrics.

We view this consistency across different surrogate constructions as evidence that FUSE's core mechanism—geometry-aware erasure of the identity support region—is not tightly dependent on a specific partitioning strategy. We initially adopted the upper–lower design because it requires no

Table 10: Embedding-space consistency metrics before and after unlearning.

Metric	Before	After (FUSE)	After (SalUn)
Avg intra-class cosine distance (retain)	0.584	0.635	0.707
Avg nearest non-matching class-centroid distance (retain)	1.289	1.294	1.336
Retained k-NN overlap@5	–	70.10%	37.76%

Table 11: Average angular deviation for unseen and forgotten classes. Similar deviation patterns indicate that forgotten identities occupy embedding regions comparable to naturally unseen identities.

Class	Unseen	Forget
Class 1	0.31	0.38
Class 2	0.28	0.27
Class 3	0.58	0.47

external face parsing, avoids segmentation noise, and remains more reliable under occlusion or non-frontal poses where fine-grained parsing can become unstable. The fine-grained experiment further suggests that FUSE can also leverage richer regional variation when available, albeit with a slightly different retention–forgetting trade-off.

E.5 Feature Space Deviation Analysis

Class Boundary To test whether unlearning subtly distorts neighboring retained classes, we compute embedding statistics for retained classes before and after unlearning (Table 10), including a retained k-NN overlap metric that measures how much each retained sample's local neighborhood is preserved. FUSE retains 70.1% of each retained sample's k-NN neighborhood at k=5, compared to 37.8% for SalUn (32.3% higher preservation), indicating substantially better preservation of local neighborhood structure. At k=10, the same trend follows and FUSE outperforms SalUn by 16.5% in maintaining the consistent neighbourhood after unlearning. A similar pattern is observed at the class level: intra-class deviation increases only modestly under FUSE (+8.7%) relative to SalUn (+21.0%), while inter-class separation remains nearly unchanged (+0.4% vs. +3.7%). These results are consistent with our cone-likelihood analysis, where FUSE reduces the fraction of forgotten samples that remain within the statistical support of their original class to 27.69%, compared to 62–85% for the baselines. Overall, the visual and geometric evidence supports that FUSE removes the forgotten identity's support region while preserving retained-class geometry.

Angular Deviation Similarity Between Forgotten and Unseen Identities A desirable property of biometric unlearning is that forgotten identities should become statistically indistinguishable from identities never observed during training. To evaluate this behavior, we analyze both verification characteristics and angular embedding structure after unlearning.

After unlearning, the angular variance and verification behavior of the forgotten identity closely resemble those of unseen identities. In particular, the verification accuracy difference between unseen and forgotten identities remains very small (0.012–0.018 for FUSE - refer Table 1 and 4), indicating that the forgotten class behaves similarly to identities absent from the original training distribution. In contrast, retained identities preserve substantially higher verification accuracy, remaining approximately 0.1 higher than unseen identities. This separation suggests that the unlearning process selectively suppresses identity-discriminative information for the target class while preserving recognition structure for retained identities.

To further examine embedding-space behavior, we compute the angular deviation for both unseen and forgotten classes. Specifically, for each class, we measure the average cosine-distance deviation θ between individual samples and the corresponding class mean embedding. Formally, for a class centroid μ_c and sample embeddings x_i, the angular deviation is computed as the average cosine distance from μ_c to x_i. We evaluate this metric on three randomly sampled unseen classes and three target forgotten classes from the same model. Table 11 reports the resulting deviations.

The observed deviations show comparable angular spread patterns between unseen and forgotten identities across all sampled classes. While minor variations exist due to class-specific embedding

Table 12: Forget set semantic residual before and after attack.

Model	Semantic Residual	
	Before Attack	After Attack
FT	0.385	0.524
GA	0.643	0.672
L1-sparse	0.648	0.732
RL	0.539	0.592
BS	0.572	0.619
BE	0.537	0.543
SalUn	0.487	0.582
SG	0.461	0.482
FUSE (ours)	0.391	0.423

geometry, the overall magnitude and distribution of deviations remain similar. This consistency suggests that the forgotten identities no longer exhibit a distinctive embedding structure separating them from naturally unseen identities.

Together, these findings support the hypothesis that successful biometric unlearning should not merely reduce recognition performance, but should instead transform forgotten identities into representations statistically similar to identities never encountered during training.

E.6 Robustness of FUSE:

We evaluate the effectiveness of FUSE under three attack: adversarial robustness, inversion attacks, and membership inference. For adversarial reconstruction, the attacker attempts to extract remaining information for the target class. SR metric serves as a proxy for this attack by quantifying recovery of class-specific structure, without requiring reconstruction. Using three generative models, we compute semantic residuals for both retain and forget sets. Across all settings, retain-set residuals indicates preserved utility, while FUSE consistently produces the lowest residuals for the forget set. MIA results are shown in table 1 and 4.

E.6.1 Inversion Attack Results :

Given an unlearned model, inversion attack produce images that reveal the identity/appearance of the forgotten class. In this attack, we applied optimization-based attack, where we assume white-box setting, and we have access to the mean embedding of forgotten class. If the attacker knows the target embedding (e.g., the forgotten class centroid), they can optimize the image so that: $f(x_{\text{generated}}) \approx$ target_embedding. This is essentially "searching" for an image whose features match the forgotten identity. We initialize with the random noise, and we apply the loss: $\mathcal{L}_{\text{feat}} = \|f(x) - \text{target_embedding}\|^2$. This forces the reconstruction to match the identity features of the forgotten class. After getting such pool of embeddings, we compute Semantic residual G_f for the forgotten class to see if it reaches any close to its original class distribution - if the attack is successful.

The results (Table 12) show that all baseline methods experience a substantial increase in semantic residual after the inversion attack, suggesting that some degree of forgotten-class information can still be recovered. Methods such as GA, L1-sparse, and SalUn show particularly large increases, indicating higher vulnerability. In contrast, FUSE exhibits the lowest semantic residual both before and after the attack, with only a modest increase from 0.391 to 0.423. This demonstrates that FUSE leaves behind minimal recoverable identity information and is significantly more robust to inversion attacks compared to other unlearning methods. The behavior confirms FUSE strong resistance against feature-space reconstruction attempts targeting the forgotten class.

F Discussion

F.1 Computational Efficiency

Our proposed FUSE training is computationally efficient, leveraging a small CNN-based feature extractor to obtain class-level representations. Our loss objectives consist of simple repulsion and

preservance terms. To reduce the computational cost of the global loss, we compute it on a randomly selected subset of sample pairs in each training epoch. The surrogate data representation similarly employs a lightweight 3-layer MLP to generate face-related embeddings. For evaluating Semantic Residual, we avoid full denoising, which is costly, and instead compute the reconstruction error at random time steps per sample, significantly reducing computation while retaining meaningful evaluation. During unlearning, optimization operates on surrogate embeddings with mini-batches, and the global loss is approximated using randomly sampled embedding pairs to avoid quadratic complexity. In practice, forgetting one identity using FUSE takes 29 minutes per identity on a V100, comparable to SalUn (28 min) and faster than SG (48 min), while achieving better forgetting performance.

F.2 Limitations

Our work with FUSE establishes a first step toward principled representation unlearning in face recognition, but several directions remain open for future research. The cone-based parameterization provides a tractable way to model identity distributions, yet future approaches could extend this to richer distributions that disentangle pose, age, and illumination, enabling more precise forgetting while preserving retention. Similarly, while FUSE leverages fixed hyperparameters for repulsion, preservation, and consistency, adaptive strategies for tuning these weights dynamically may improve stability across datasets and varying unlearning scales.

Beyond the algorithm itself, our evaluation framework, particularly Semantic Residual (SR), offers a measure of semantic leakage. While we evaluate SR under inversion and membership inference attacks, stronger black-box and adaptive adversaries, as well as sequential unlearning requests arriving over time, remain open directions for establishing more rigorous privacy assurances. Extending the principles of distribution erasure and semantic auditing beyond faces to other fine-grained tasks like text, voice, or multimodal models may uncover broader foundations for representation-level unlearning.