# Selective Amnesia using Contrastive Subnet Erasure for Class Level Unlearning in Vision Models

Vishal Pramanik Maisha Maliha Susmit Jha Alvaro Velasquez Olivera Kotevska
Sumit Kumar Jha

# Supplementary Materials

**Algorithm 1** Contrastive Subnet Erasure (CSE)

**Require:** encoder $\phi$ with layers $\mathcal{L}$; target set $D_t$; non-target set $D_b$; hyperparameters $(\alpha, k_{\max}, \beta, \tau_{\text{cov}}, \tau_0, \lambda_0, \varepsilon)$
**Ensure:** edited encoder $\phi'$

**Stage 1: Standardization**
**for** $\ell \in \mathcal{L}$ **do**
  extract $h^{(\ell)}(x)$ for all $x \in D_t \cup D_b$
  $\mu^{(\ell)} \leftarrow \text{mean}_x\, h^{(\ell)}(x)$
  $\sigma^{(\ell)} \leftarrow \big(\text{var}_x(h^{(\ell)}(x)) + \varepsilon\big)^{1/2}$
  $\tilde{h}^{(\ell)}(x) \leftarrow \big(h^{(\ell)}(x) - \mu^{(\ell)}\big) \oslash \sigma^{(\ell)}$ ▷ channelwise
**end for**
**Stage 2: Subnet discovery**
**for** $\ell \in \mathcal{L}$ **do**
  $\Sigma_t^{(\ell)} \leftarrow \text{EmpCov}\big(\tilde{h}^{(\ell)}(x), x \in D_t\big)$
  $\Sigma_b^{(\ell)} \leftarrow \text{EmpCov}\big(\tilde{h}^{(\ell)}(x), x \in D_b\big)$
  $\tilde{\Sigma}_b^{(\ell)} \leftarrow \Sigma_b^{(\ell)} + \alpha \frac{\text{tr}(\Sigma_b^{(\ell)})}{d_\ell} I$
  solve $\Sigma_t^{(\ell)} v_j^{(\ell)} = \rho_j^{(\ell)} \tilde{\Sigma}_b^{(\ell)} v_j^{(\ell)}$
  $k_\ell \leftarrow \min\big(k_{\max}, \lfloor \beta d_\ell \rfloor\big)$
  $s_c^{(\ell)} \leftarrow \sum_{j=1}^{k_\ell} \rho_j^{(\ell)} \big(v_j^{(\ell)}[c]\big)^2 \;\; \forall c$
  sort $c$ by $s_c^{(\ell)}$; choose smallest $C^{(\ell)}$ with $\sum_{c \in C^{(\ell)}} s_c^{(\ell)} \geq \tau_{\text{cov}} \sum_c s_c^{(\ell)}$
**end for**
**Stage 3: Attenuation and runtime form**
**for** $\ell \in \mathcal{L}$ **do**
  $\beta_c^{(\ell)} \leftarrow \text{clip}_{[0,1]}\Big(\frac{s_c^{(\ell)} - \tau_0}{s_c^{(\ell)} + \lambda_0}\Big)$ for all $c$
  $\text{scale}_c^{(\ell)} \leftarrow 1 - \beta_c^{(\ell)}$
  $\text{bias}_c^{(\ell)} \leftarrow \beta_c^{(\ell)} \mu_c^{(\ell)}$
  runtime: $h_{\text{att}}^{(\ell)}(x) \leftarrow \text{scale}^{(\ell)} \odot h^{(\ell)}(x) + \text{bias}^{(\ell)}$
**end for**
fold per-channel scales/biases into following linear/conv layers to obtain $\phi'$
**return** $\phi'$

## A. Algorithm

The algorithm of our method is in 1.

## B. Theoretical Properties and Error Bounds

This appendix summarizes several basic theoretical properties of contrastive subnet erasure (CSE) and gives simple error bounds that clarify when it is well-behaved. All statements are made for a single layer, so we drop the layer index whenever it is clear from context.

### B.1. Setup and Notation

Let $h(x) \in \mathbb{R}^d$ denote the pooled feature vector at a fixed encoder layer for an input $x$. The standardized feature $\tilde{h}(x)$ is defined as

$$\begin{aligned} \mu &= \frac{1}{n_t + n_b} \sum_{x \in D_t \cup D_b} h(x), \\ \sigma_c &= \sqrt{\frac{1}{n_t + n_b} \sum_x \big(h_c(x) - \mu_c\big)^2 + \varepsilon}, \end{aligned} \tag{15}$$

and

$$\begin{aligned} S &= \text{diag}\big(1/\sigma_1, \ldots, 1/\sigma_d\big), \\ \tilde{h}(x) &= S\big(h(x) - \mu\big), \end{aligned} \tag{16}$$

where $D_t$ and $D_b$ are the target and non-target sets, $\varepsilon > 0$ is a small constant, and $c \in \{1, \ldots, d\}$ indexes channels.

The empirical target and non-target covariances in standardized space are

$$\begin{aligned} \Sigma_t &= \frac{1}{n_t} \sum_{x \in D_t} \tilde{h}(x)\tilde{h}(x)^\top, \\ \Sigma_b &= \frac{1}{n_b} \sum_{x \in D_b} \tilde{h}(x)\tilde{h}(x)^\top, \end{aligned} \tag{17}$$

and the regularized background covariance is

$$\begin{aligned} \tilde{\Sigma}_b &= \Sigma_b + \delta I_d, \\ \delta &= \alpha \frac{\text{tr}(\Sigma_b)}{d}, \end{aligned} \tag{18}$$

with regularization factor $\alpha > 0$.

The contrastive Rayleigh quotient of a nonzero vector $v$ is

$$\rho(v) = \frac{v^\top \Sigma_t v}{v^\top \tilde{\Sigma}_b v}. \tag{19}$$

The generalized eigenproblem

$$\Sigma_t v_j = \rho_j \tilde{\Sigma}_b v_j \tag{20}$$

has real eigenvalues $\rho_j$ and eigenvectors $v_j \in \mathbb{R}^d$, ordered so that $\rho_1 \geq \rho_2 \geq \cdots \geq \rho_d$.

CSE uses the top

$$k \;=\; \min\bigl(k_{\max}, \lfloor \beta d \rfloor\bigr) \tag{21}$$

eigenpairs to define channel salience scores

$$s_c \;=\; \sum_{j=1}^{k} \rho_j \, v_j[c]^2, \qquad c \in \{1, \ldots, d\}, \tag{22}$$

and selects the smallest index set $C \subset \{1, \ldots, d\}$ such that

$$\sum_{c \in C} s_c \;\geq\; \tau_{\mathrm{cov}} \sum_{c=1}^{d} s_c, \tag{23}$$

for a coverage threshold $\tau_{\mathrm{cov}} \in (0, 1)$.

Given scores $s_c$, the attenuation factors $\beta_c \in [0, 1]$ and $a_c = 1 - \beta_c$ are

$$\begin{aligned} \beta_c \;&=\; \mathrm{clip}_{[0,1]}\Bigl(\frac{s_c - \tau_0}{s_c + \lambda_0}\Bigr), \\ a_c \;&=\; 1 - \beta_c, \end{aligned} \tag{24}$$

with parameters $\tau_0 > 0$, $\lambda_0 > 0$. The diagonal attenuation matrix in standardized coordinates is

$$\begin{aligned} A \;&=\; \mathrm{diag}(a_1, \ldots, a_d), \\ M \;&=\; S^{-1} A S, \end{aligned} \tag{25}$$

and the attenuated feature in the original coordinate system is

$$h_{\mathrm{att}}(x) \;=\; M h(x) + (I_d - M)\mu. \tag{26}$$

This expression is exactly equivalent to applying $A$ in standardized coordinates, then un-standardizing.

## B.2. Properties of the Contrastive Eigenproblem

Assume $\tilde{\Sigma}_b \succ 0$, which holds by construction since $\delta > 0$. Then the generalized eigenproblem $\Sigma_t v = \rho \, \tilde{\Sigma}_b v$ has the following standard properties:

**Optimality and ordering.** There exists a basis of generalized eigenpairs $\{(\rho_j, v_j)\}_{j=1}^{d}$ such that:

- $v_i^\top \tilde{\Sigma}_b v_j = \delta_{ij}$ (orthonormality in the $\tilde{\Sigma}_b$–inner product);
- $\rho_1 = \max_{\|v\|>0} \rho(v)$ and $\rho_d = \min_{\|v\|>0} \rho(v)$;
- for any nonzero $v$, $\rho_d \leq \rho(v) \leq \rho_1$.

These follow by whitening $\tilde{\Sigma}_b$ and reducing to an ordinary eigenproblem for the symmetric matrix $\tilde{\Sigma}_b^{-1/2} \Sigma_t \tilde{\Sigma}_b^{-1/2}$.

**Salience conservation.** For defining salience, we normalize the eigenvectors in Euclidean norm, $\|v_j\|_2 = 1$. Under this normalization,

$$\begin{aligned} \sum_{c=1}^{d} s_c \;&=\; \sum_{c=1}^{d} \sum_{j=1}^{k} \rho_j v_j[c]^2 \\ &=\; \sum_{j=1}^{k} \rho_j \sum_{c=1}^{d} v_j[c]^2 \\ &=\; \sum_{j=1}^{k} \rho_j. \end{aligned} \tag{27}$$

Thus $\{s_c\}_{c=1}^{d}$ form a nonnegative decomposition of the total contrastive signal carried by the top $k$ generalized eigendirections. The coverage constraint

$$\sum_{c \in C} s_c \;\geq\; \tau_{\mathrm{cov}} \sum_{c=1}^{d} s_c \tag{28}$$

guarantees that the selected subnet captures at least a fraction $\tau_{\mathrm{cov}}$ of this signal.

## B.3. Effect of Attenuation on Covariances

Consider random standardized feature vectors $\tilde{H}_t$ and $\tilde{H}_b$ with population covariances

$$\begin{aligned} \Sigma_t^\star \;&=\; \mathbb{E}[\tilde{H}_t \tilde{H}_t^\top], \\ \Sigma_b^\star \;&=\; \mathbb{E}[\tilde{H}_b \tilde{H}_b^\top], \end{aligned} \tag{29}$$

and let $A$ be the diagonal attenuation matrix defined above. In standardized coordinates,

$$\begin{aligned} \tilde{H}_t' \;&=\; A \tilde{H}_t, \\ \tilde{H}_b' \;&=\; A \tilde{H}_b, \end{aligned} \tag{30}$$

so the post-attenuation covariances are

$$\begin{aligned} \Sigma_t' \;&=\; \mathbb{E}[\tilde{H}_t' \tilde{H}_t'^\top] \;=\; A \Sigma_t^\star A, \\ \Sigma_b' \;&=\; \mathbb{E}[\tilde{H}_b' \tilde{H}_b'^\top] \;=\; A \Sigma_b^\star A. \end{aligned} \tag{31}$$

Define

$$\begin{aligned} a_{\min} \;&=\; \min_{1 \leq c \leq d} a_c, \\ a_{\max} \;&=\; \max_{1 \leq c \leq d} a_c, \end{aligned} \tag{32}$$

so that $0 \leq a_{\min} \leq a_{\max} \leq 1$.

**Bounds on eigenvalues.** For any symmetric positive semidefinite matrix $\Sigma \succeq 0$, the eigenvalues of $A \Sigma A$ lie between $a_{\min}^2$ and $a_{\max}^2$ times the eigenvalues of $\Sigma$:

$$a_{\min}^2 \, \lambda_{\min}(\Sigma) \;\leq\; \lambda_{\min}(A \Sigma A) \;\leq\; \lambda_{\max}(A \Sigma A) \;\leq\; a_{\max}^2 \, \lambda_{\max}(\Sigma). \tag{33}$$

Indeed, for any unit vector $u$,

$$\begin{aligned} u^\top A\Sigma A u &= (Au)^\top \Sigma (Au) \\ &\le \lambda_{\max}(\Sigma)\,\|Au\|_2^2 \\ &\le \lambda_{\max}(\Sigma)\,a_{\max}^2, \end{aligned} \tag{34}$$

and similarly

$$\begin{aligned} u^\top A\Sigma A u &\ge \lambda_{\min}(\Sigma)\,\|Au\|_2^2 \\ &\ge \lambda_{\min}(\Sigma)\,a_{\min}^2. \end{aligned} \tag{35}$$

Taking maxima and minima over all unit $u$ yields the stated bounds. Applied to $\Sigma_t^\star$ and $\Sigma_b^\star$, this shows that CSE cannot increase the spectral norms of the target or non-target covariances in standardized space; they are both scaled by factors between $a_{\min}^2$ and $a_{\max}^2$.

**Directional contraction under diagonal target covariance.** In the special case where the population target covariance is diagonal, $\Sigma_t^\star = \mathrm{diag}(\lambda_1^t, \dots, \lambda_d^t)$, the variance along any unit direction $v$ after attenuation is

$$\begin{aligned} v^\top \Sigma_t' v &= v^\top A \Sigma_t^\star A v \\ &= \sum_{c=1}^{d} a_c^2 \lambda_c^t\, v[c]^2. \end{aligned} \tag{36}$$

If a generalized eigenvector $v_j^\star$ has most of its mass on channels with strong attenuation ($a_c \ll 1$), then

$$v_j^{\star\top} \Sigma_t' v_j^\star \;\le\; \Big(\max_{c: v_j^\star[c] \neq 0} a_c^2\Big) v_j^{\star\top} \Sigma_t^\star v_j^\star, \tag{37}$$

and conversely, if it is supported on lightly attenuated channels ($a_c \approx 1$), the variance along $v_j^\star$ is almost preserved. This formalizes the intuition that CSE acts locally in channel space: it contracts variance more strongly along directions that are heavily supported on high-salience channels, and less along directions supported on low-salience channels.

## B.4. Finite-Sample Error Bounds

The derivation above assumes access to population covariances $\Sigma_t^\star$, $\Sigma_b^\star$. In practice, CSE operates with empirical covariances $\Sigma_t$, $\Sigma_b$ estimated from $n_t$ and $n_b$ samples. Here we collect standard finite-sample bounds for these estimates and for their generalized eigenpairs.

**Covariance estimation.** Assume that standardized features $\tilde{H}_t$ and $\tilde{H}_b$ are zero-mean and subgaussian with parameter $\kappa > 0$; that is, for any unit vector $u \in \mathbb{S}^{d-1}$,

$$\mathbb{E}\exp\big(\langle u, \tilde{H}_t\rangle^2/\kappa^2\big) \;\le\; 2, \tag{38}$$

and similarly for $\tilde{H}_b$. Let

$$\begin{aligned} E_t &= \Sigma_t - \Sigma_t^\star, \\ E_b &= \Sigma_b - \Sigma_b^\star. \end{aligned} \tag{39}$$

Then there exists a constant $C > 0$ (depending only on the subgaussian parameter) such that, for any $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\begin{aligned} \|E_t\|_2 &\le C\kappa^2 \sqrt{\frac{d + \log(1/\delta)}{n_t}}, \\ \|E_b\|_2 &\le C\kappa^2 \sqrt{\frac{d + \log(1/\delta)}{n_b}}. \end{aligned} \tag{40}$$

These are standard matrix concentration bounds for empirical covariances of subgaussian vectors.

**Generalized eigenvalues and eigenvectors.** Let $\{(\rho_j^\star, v_j^\star)\}$ denote the population generalized eigenpairs of $(\Sigma_t^\star, \tilde{\Sigma}_b^\star)$, where $\tilde{\Sigma}_b^\star = \Sigma_b^\star + \delta I$, and $\{(\rho_j, v_j)\}$ the empirical generalized eigenpairs of $(\Sigma_t, \tilde{\Sigma}_b)$, where $\tilde{\Sigma}_b = \Sigma_b + \delta I$. Assume the top $k$ population eigenvalues have a positive generalized eigengap

$$\gamma \;=\; \min_{1 \le j \le k} \big(\rho_j^\star - \rho_{j+1}^\star\big) \;>\; 0. \tag{41}$$

By reducing to an ordinary eigenproblem in the whitened basis $\tilde{\Sigma}_b^{\star -1/2} \Sigma_t^\star \tilde{\Sigma}_b^{\star -1/2}$, classical perturbation theory (Davis–Kahan-type results) implies that there exist constants $C_1, C_2 > 0$ (depending on $\tilde{\Sigma}_b^\star$ and $\delta$) such that, for all $j \le k$,

$$|\rho_j - \rho_j^\star| \;\le\; C_1\big(\|E_t\|_2 + \|E_b\|_2\big), \tag{42}$$

and

$$\sin\angle(v_j, v_j^\star) \;\le\; C_2\, \frac{\|E_t\|_2 + \|E_b\|_2}{\gamma}. \tag{43}$$

Combining these with the covariance bounds above yields explicit finite-sample error bounds on eigenvalues and eigenvectors in terms of $d$, $n_t$, $n_b$, and the eigengap $\gamma$.

**Salience scores and channel selection.** Let $s_c^\star$ and $s_c$ denote the population and empirical salience scores for channel $c$, computed from $\{(\rho_j^\star, v_j^\star)\}_{j=1}^k$ and $\{(\rho_j, v_j)\}_{j=1}^k$, respectively. By expanding $s_c - s_c^\star$, applying the triangle inequality, and using the eigenpair perturbation bounds above, one can show that there exists a constant $C_3 > 0$ such that

$$\max_{1 \le c \le d} |s_c - s_c^\star| \;\le\; C_3 k \Big(\max_{j \le k} |\rho_j - \rho_j^\star| + \max_{j \le k} \|v_j - v_j^\star\|_2\Big), \tag{44}$$

which is

$$O\left(k\, \frac{\|E_t\|_2 + \|E_b\|_2}{\gamma}\right) \tag{45}$$

under the assumptions above.

Suppose the population scores have a margin between in-subnet and out-of-subnet channels:

$$\Delta \;=\; \min_{c\in C^\star} s_c^\star \;-\; \max_{c\notin C^\star} s_c^\star \;>\; 0, \tag{46}$$

where $C^\star$ is the population minimal set of channels achieving the coverage threshold. If

$$\max_c |s_c - s_c^\star| \;\le\; \Delta/2, \tag{47}$$

then the empirical greedy selection recovers $C^\star$ exactly. Thus, provided the eigengap $\gamma$ and margin $\Delta$ are not too small and $n_t, n_b$ are sufficiently large, CSE's subnet selection is stable under sampling noise.

### B.5. Idealized Projection and Approximation by CSE

It is useful to compare CSE's channel-wise attenuation with an idealized projection that directly removes generalized eigen-directions.

**Idealized eigen-projection in whitened space.** Consider the population covariances $\Sigma_t^\star$ and $\Sigma_b^\star$ and define the whitened target covariance

$$C_t \;=\; (\tilde\Sigma_b^\star)^{-1/2}\Sigma_t^\star(\tilde\Sigma_b^\star)^{-1/2}. \tag{48}$$

Let $(\rho_j^\star, u_j)$ be the eigenpairs of $C_t$, with

$$u_j^\top u_j \;=\; 1, \qquad C_t u_j \;=\; \rho_j^\star u_j, \qquad j = 1,\dots,d. \tag{49}$$

The whitened background covariance is the identity, so $\rho_j^\star$ are exactly the generalized eigenvalues of $(\Sigma_t^\star, \tilde\Sigma_b^\star)$.

Define the projector onto the orthogonal complement of the top $k$ eigendirections in whitened space:

$$Q_\perp \;=\; I_d - \sum_{j=1}^{k} u_j u_j^\top. \tag{50}$$

The idealized transformed whitened features

$$\begin{aligned} Z_t' \;&=\; Q_\perp Z_t, \\ Z_b' \;&=\; Q_\perp Z_b, \end{aligned} \tag{51}$$

where

$$\begin{aligned} Z_t \;&=\; (\tilde\Sigma_b^\star)^{-1/2}\tilde H_t, \\ Z_b \;&=\; (\tilde\Sigma_b^\star)^{-1/2}\tilde H_b, \end{aligned} \tag{52}$$

have covariances

$$\begin{aligned} C_t^{\mathrm{proj}} \;&=\; Q_\perp C_t Q_\perp, \\ C_b^{\mathrm{proj}} \;&=\; Q_\perp I Q_\perp \;=\; Q_\perp. \end{aligned} \tag{53}$$

Because $Q_\perp$ annihilates $u_1,\dots,u_k$ but leaves $u_{k+1},\dots,u_d$ unchanged, the nonzero eigenvalues of $C_t^{\mathrm{proj}}$ are exactly $\rho_{k+1}^\star,\dots,\rho_d^\star$. Consequently, if we consider the Rayleigh quotient

$$R(v) \;=\; \frac{v^\top C_t^{\mathrm{proj}} v}{v^\top C_b^{\mathrm{proj}} v}, \qquad v\neq 0,\; v\in \mathrm{range}(Q_\perp), \tag{54}$$

then

$$\max_{\substack{v\neq 0 \\ v\in\mathrm{range}(Q_\perp)}} R(v) \;=\; \rho_{k+1}^\star. \tag{55}$$

In whitened coordinates, an ideal eigen-projection can therefore reduce the maximum target-to-background variance ratio from $\rho_1^\star$ to $\rho_{k+1}^\star$ exactly.

**Channel-wise attenuation as a diagonal approximation.** CSE does not implement $Q_\perp$ directly. Instead, it applies a diagonal attenuation $A$ in the original standardized coordinates. Let

$$\begin{aligned} \tilde H_t' \;&=\; A\tilde H_t, \\ \tilde H_b' \;&=\; A\tilde H_b, \end{aligned} \tag{56}$$

and let $C_t^{\mathrm{CSE}}, C_b^{\mathrm{CSE}}$ denote the corresponding whitened covariances:

$$\begin{aligned} C_t^{\mathrm{CSE}} \;&=\; (\tilde\Sigma_b^\star)^{-1/2} A\Sigma_t^\star A(\tilde\Sigma_b^\star)^{-1/2}, \\ C_b^{\mathrm{CSE}} \;&=\; (\tilde\Sigma_b^\star)^{-1/2} A\Sigma_b^\star A(\tilde\Sigma_b^\star)^{-1/2}. \end{aligned} \tag{57}$$

If the top generalized eigenvectors $u_1,\dots,u_k$ are close to sparse vectors in the canonical basis (for example, strongly aligned with a subset of channels), and CSE assigns $a_c \approx 0$ on the corresponding channels while keeping $a_c \approx 1$ elsewhere, then $A$ approximately zeroes out those eigendirections. In such regimes we can view $C_t^{\mathrm{CSE}}$ and $C_b^{\mathrm{CSE}}$ as diagonal approximations to $C_t^{\mathrm{proj}}$ and $C_b^{\mathrm{proj}}$, and expect the maximum target-to-background variance ratio under CSE to be close to $\rho_{k+1}^\star$ (up to factors depending on the quality of this alignment).

This perspective clarifies that CSE approximates an ideal projection onto the complement of the most target-salient generalized eigen-directions, but does so using only channel-wise attenuation, which is architecture agnostic and can be folded into existing weights without changing the model's computational graph.

## C. Extended Experimental Setup

This appendix provides the details needed to reproduce all experiments: dataset protocols and class mappings, exact splits for target / non-target / evaluation sets, metric and attack definitions, and the hyperparameters used for CSE and all baselines.

## C.1. Datasets and Cross Dataset Protocols

All experiments use standard vision benchmarks. We consider CIFAR-10, which contains 10 classes with 50,000 training and 10,000 test images; CIFAR-100, which contains 100 classes with 50,000 training and 10,000 test images; ImageNet-1K, which contains 1,000 classes with approximately 1.28M training images and 50,000 validation images (treated as test); and LFW, which contains 13,233 images of 5,749 identities and is used for identity forgetting. Unless stated otherwise, we use the standard train/test splits provided with each dataset. All backbones (ResNet-18, EfficientNet-B0, Swin-T) are initialized from ImageNet-1K pretraining. For CIFAR experiments, images are resized to $224\times224$; training uses random crops and horizontal flips, and evaluation uses center crops only.

For class-level forgetting, a semantic class family is selected and aligned across datasets using fixed label mappings. Unlearning is applied to a *source* dataset and forgetting is evaluated on a disjoint *evaluation* dataset that shares the same semantic class but not the same images. Each dataset appears as both source and evaluation domain across the probes.

The semantic alignments used across all cross–dataset experiments are summarized in Table 3. When an exact class name is not present in ImageNet-1K, the closest included class is used.

The main single-class probes are CIFAR-10 $\rightarrow$ ImageNet (airplane family), ImageNet $\rightarrow$ CIFAR-10 (truck family), and ImageNet $\rightarrow$ CIFAR-100 (shark family). In addition, multi-class forgetting on CIFAR-100 is considered by constructing target sets of size $\{2, 3, 4, 5\}$ from object-like categories such as `castle`, `telephone`, `television`, and `lawn_mower`.

For identity-level forgetting, the LFW dataset is split by identities into disjoint train and test sets, with 80% of identities used for training and 20% for testing. A single identity with at least 50 images is selected as the target, and all remaining identities are treated as retain classes. A ResNet-18 backbone is first fine-tuned on the full LFW training identities for face recognition; unlearning then targets only the selected identity.

## C.2. Splits and Sample Counts

Let $D$ denote a dataset with training split $D^{\text{train}}$ and test split $D^{\text{test}}$. For a given set of target classes $C_t$, we define the target and retain subsets

$$\begin{aligned} D_t^{\text{train}} &= \{(x, y) \in D^{\text{train}} : y \in C_t\}, \\ D_r^{\text{train}} &= \{(x, y) \in D^{\text{train}} : y \notin C_t\}, \\ D_t^{\text{test}} &= \{(x, y) \in D^{\text{test}} : y \in C_t\}, \\ D_r^{\text{test}} &= \{(x, y) \in D^{\text{test}} : y \notin C_t\}. \end{aligned}$$

For single-class forgetting on CIFAR-10, this yields $|D_t^{\text{train}}| = 5{,}000$, $|D_r^{\text{train}}| = 45{,}000$, $|D_t^{\text{test}}| = 1{,}000$, and $|D_r^{\text{test}}| = 9{,}000$. For single-class forgetting on CIFAR-100, the corresponding counts are $|D_t^{\text{train}}| = 500$, $|D_r^{\text{train}}| = 49{,}500$, $|D_t^{\text{test}}| = 100$, and $|D_r^{\text{test}}| = 9{,}900$. On ImageNet-1K, class sizes vary; for a typical target class we have $|D_t^{\text{train}}| \approx 1{,}300$ and $|D_t^{\text{test}}| = 50$, with all remaining images assigned to $D_r^{\text{train}}$ and $D_r^{\text{test}}$.

CSE requires a target set $D_t$ and a non-target set $D_b$. The target set is always the full target training subset, i.e.,

$$D_t = D_t^{\text{train}}.$$

The non-target set $D_b$ is formed by sampling a small subset of semantically related non-target classes drawn from the evaluation dataset. Concretely, we select two or three classes that are nearby in concept space (for example, `bird` and `ship` when forgetting `airplane`) and sample 10% of their training images to form $D_b$. Thus, for CIFAR-10 single-class forgetting, $D_b$ typically contains $0.1 \times 5{,}000 = 500$ samples from each chosen related class, i.e., between 500 and 1,500 images depending on the number of classes used. In ablations, we vary this fraction (for example, 5%, 10%, 15%), replace related classes by distant ones, or set $D_b = \varnothing$.

For evaluation, we distinguish four subsets. The forget-train set $D_{\text{ft}}$ consists of all target training samples $D_t^{\text{train}}$. The forget-test set $D_{\text{ft}}^{\text{test}}$ consists of all target test samples $D_t^{\text{test}}$. The retain-train set $D_{\text{rt}}$ consists of all non-target training samples $D_r^{\text{train}}$, and the retain-test set $D_{\text{rt}}^{\text{test}}$ consists of all non-target test samples $D_r^{\text{test}}$. These subsets are used to compute all reported metrics.

## C.3. Metrics and Membership Inference Attack

Given a model $f$ and a labeled dataset $S = \{(x_i, y_i)\}$, the classification accuracy is defined as

$$\text{Acc}(S; f) = \frac{1}{|S|} \sum_{(x_i, y_i) \in S} \mathbb{1}[f(x_i) = y_i].$$

We report four accuracy metrics:

$$\text{Accf} = \text{Acc}(D_{\text{ft}}; f), \qquad \text{Accft} = \text{Acc}(D_{\text{ft}}^{\text{test}}; f),$$

$$\text{Accr} = \text{Acc}(D_{\text{rt}}; f), \qquad \text{Accrt} = \text{Acc}(D_{\text{rt}}^{\text{test}}; f),$$

corresponding respectively to forget accuracy on train, forget accuracy on test, retain accuracy on train, and retain accuracy on test. Effective unlearning corresponds to low Accf and Accft, while maintaining high Accr and Accrt.

To summarize the trade-off between forgetting and retention, we convert forget-test accuracy into a forget-success score

$$F = 1 - \text{Accft},$$

Table 3. Semantic class mappings used in cross–dataset experiments. When an exact ImageNet label is unavailable, the nearest included class is used (e.g., `airliner` may be approximated by a related aircraft class).

| Family | CIFAR class | ImageNet-1K class(es) |
|---|---|---|
| Airplane | CIFAR-10 `airplane` | `airliner` or nearest aircraft class |
| Truck | CIFAR-10 `truck` | {`garbage_truck`, `tow_truck`, `trailer_truck`} |
| Ship | CIFAR-10 `ship` | `container_ship` |
| Cat | CIFAR-10 `cat` | `tabby_cat` |
| Frog | CIFAR-10 `frog` | `bullfrog` |
| Shark | CIFAR-100 `shark` | {`white_shark`, `tiger_shark`} |
| Castle | CIFAR-100 `castle` | `castle` |
| Keyboard | CIFAR-100 `keyboard` | `computer_keyboard` |
| Telephone | CIFAR-100 `telephone` | `cellular_telephone`, `dial_telephone` |
| Television | CIFAR-100 `television` | `television` |
| Lawn mower | CIFAR-100 `lawn_mower` | `lawn_mower` |

where $F = 1$ indicates perfect forgetting (zero test accuracy on the target). The reported harmonic mean is

$$\text{H-Mean} = \frac{2F \cdot \text{Accrt}}{F + \text{Accrt}},$$

which lies in $[0, 1]$ and is large when both forgetting ($F$) and retention ($\text{Accrt}$) are simultaneously high.

Membership inference is used to quantify residual memorization on the forget set. We consider a simple loss-threshold attack. The member set consists of all samples in $D_{\text{ft}}$ (target training), and the non-member set consists of an equally sized subset of $D_{\text{ft}}^{\text{test}}$ (target test). For each sample $(x, y)$ in this balanced pool we compute the loss

$$\ell(x, y) = -\log p_f(y \mid x),$$

where $p_f(\cdot \mid x)$ is the model's predictive distribution. The pool is split in half; on the first half a scalar threshold $\tau$ is chosen to maximize membership classification accuracy for the decision rule

$$\hat{m}(x, y) = \mathbb{1}[\ell(x, y) < \tau],$$

where $\hat{m} = 1$ indicates "member" and $\hat{m} = 0$ indicates "non-member". The membership inference attack success rate (MIA) is the classification accuracy of $\hat{m}$ on the held-out half. On a balanced mixture, $\text{MIA} \approx 0.5$ corresponds to random guessing.

## C.4. Hyperparameters

CSE uses the same hyperparameters across all backbones and datasets unless specified otherwise. For each block $\ell$, the background covariance $\Sigma_b^{(\ell)}$ is regularized as

$$\Sigma_b^{(\ell)} \leftarrow \Sigma_b^{(\ell)} + \alpha \cdot \frac{\text{tr}(\Sigma_b^{(\ell)})}{d_\ell} I, \quad \alpha = 0.01.$$

The number of eigenvectors used in salience computation is

$$k_\ell = \min(k_{\max}, \lfloor \beta \, d_\ell \rfloor),$$

with $k_{\max} = 50$ and $\beta = 0.3$. The subnet at each block is chosen as the smallest set of channels whose cumulative salience reaches the coverage threshold $\tau_{\text{cov}} = 0.85$. Salience values are mapped to attenuation strengths with a smooth transfer function parameterized by $\tau_0 = 0.1$ and $\lambda_0 = 0.5$, followed by clipping of the resulting attenuation coefficients to the interval $[0, 1]$. The per-channel standard deviation estimates include a small constant $\varepsilon = 10^{-6}$ inside the square root for numerical stability. Unless otherwise indicated, 10% of images per semantically similar non-target class are used to form $D_b$. These values are used for ResNet-18, EfficientNet-B0, and Swin-T without any per-backbone tuning.

All training-based unlearning baselines share a common fine-tuning schedule. We use stochastic gradient descent (SGD) with momentum 0.9, learning rate $10^{-5}$, batch size 64, and 10 fine-tuning epochs. This schedule is applied to DELETE, BU (Boundary Unlearning), SCAR, ESC-T, and SCRUB+R. For DELETE, the encoder and classifier are fine-tuned jointly with the DELETE objective on the retain training data (and any auxiliary loss terms defined by the method). For BU, only the classifier head is fine-tuned with a regularizer that shrinks the decision boundary associated with the target class. For SCAR, the encoder and head are fine-tuned with the SCAR objective that selectively corrects predictions on target samples while preserving retain performance. ESC-T is initialized from the ESC-edited representation and then fine-tuned for 10 epochs using the same schedule. SCRUB+R applies SCRUB (described below), followed by 10 epochs of head fine-tuning on retain training data to recover any loss in retain accuracy.

Closed-form baselines use their recommended hyperparameters from their respective implementations. ESC

is applied to the encoder representation with default rank and regularization parameters; no additional training is performed. LEACE is implemented as a linear projection that removes directions associated with the target concept, using a maximum projection rank (for example, 64 directions) per layer. SCRUB removes a fixed number of concept directions per layer (for example, up to $r = 32$ directions), with default regularization. Targeted CLIP performs text-guided editing driven by the target class name; the number of optimization steps and learning rate follow the defaults from the official implementation. Unless explicitly noted, these hyperparameters are kept fixed across datasets and backbones to isolate the effect of the different unlearning strategies.

## D. Runtime and Resource Usage

We also compare the runtime and resource footprint of CSE against the main training-based unlearning baselines. As a representative configuration, we consider single-class forgetting on CIFAR-10 (airplane family) with a ResNet-18 backbone. All methods are run on a single NVIDIA A100 GPU with 40 GB memory, using the same dataloader, batch size, and mixed-precision settings. Training-based methods perform 10 epochs of fine-tuning on the retain data, whereas CSE and other analytic methods perform a single offline pass to collect features and compute their respective projections or transformations. Table 4 reports approximate wall-clock unlearning time and peak GPU memory usage.

In this setting, CSE completes unlearning in under ten minutes with a peak memory footprint of about 4.5 GB, comparable to other analytic methods (ESC, LEACE, SCRUB), which also require only a single pass over the data. By contrast, training-based approaches (DELETE, BU, SCAR, ESC-T, SCRUB+R) require between roughly 48 and 72 minutes due to multiple epochs of optimization, and consistently use more GPU memory (around 5–6 GB) to store gradients and optimizer state. Thus, CSE achieves effective unlearning while being substantially faster than training-based baselines and using comparable or lower GPU memory.

Table 4. Approximate runtime and peak GPU memory usage for single-class forgetting on CIFAR-10 (airplane) with a ResNet-18 backbone on a single NVIDIA A100 (40 GB). Times are wall-clock minutes for the full unlearning procedure.

| Method | Type | Time (min) | Peak GPU mem (GB) |
|---|---|---|---|
| CSE | analytic | 8 | 4.5 |
| ESC | analytic | 7 | 4.3 |
| LEACE | analytic | 6 | 4.1 |
| SCRUB | analytic | 9 | 4.7 |
| DELETE | training-based | 60 | 5.8 |
| BU | training-based | 48 | 5.2 |
| SCAR | training-based | 65 | 6.0 |
| ESC-T | training-based | 55 | 5.7 |
| SCRUB+R | training-based | 72 | 6.1 |

## E. Ablations and Qualitative Examples

This appendix presents additional empirical evidence for CSE. First, it summarizes critical ablations on the non-target set design and on key CSE hyperparameters, including tabulated results. It then provides qualitative Grad-CAM visualizations for cross-dataset unlearning, together with clear failure cases that illustrate the limits of the method.

### E.1. Non-Target Set Design

CSE relies on a small non-target set $D_b$ to provide contrastive structure for subnet discovery. To understand how the design of $D_b$ affects the trade-off between forgetting and retention, we consider single-class forgetting on CIFAR-10 (forgetting `airplane`) with a ResNet-18 backbone under five variants of $D_b$. In each case, the target set is the full CIFAR-10 `airplane` training class, and the non-target set is constructed from varying subsets of the remaining classes. Table 5 reports forget-test accuracy (Accft; lower is better) and retain-test accuracy (Accrt; higher is better) for each configuration.

Table 5. Impact of non-target set design on CSE for CIFAR-10 single-class forgetting (`airplane`, ResNet-18). Accft: forget-test accuracy (lower is better). Accrt: retain-test accuracy (higher is better).

| ID | Non-target set $D_b$ design | Accft ↓ | Accrt ↑ |
|---|---|---|---|
| (1) | Semantic, 10% / class (`bird`, `ship`) | 0.02 | 0.93 |
| (2) | Semantic, 5% / class (`bird`, `ship`) | 0.02 | 0.85 |
| (3) | Semantic, 15% / class (`bird`, `ship`) | 0.02 | 0.94 |
| (4) | No semantic overlap (`truck`, `cat`) | 0.02 | 0.81 |
| (5) | No non-target set ($D_b = \varnothing$) | 0.02 | 0.76 |

The default configuration uses semantically similar non-target classes, sampling 10% of training images from `bird` and `ship`. This yields Accft $= 0.02$ and Accrt $= 0.93$, indicating nearly complete forgetting of the target with only a small drop in retain accuracy relative to the original model. Reducing the sampling rate to 5% (configuration 2) halves the size of $D_b$; Accft remains at $0.02$, but Accrt decreases to $0.85$, showing that too few non-target examples make the covariance and eigenanalysis less stable and lead to more collateral damage.

Increasing $D_b$ to 15% per class (configuration 3) yields Accft $= 0.02$ and Accrt $= 0.94$, a slight improvement over the default. This suggests that larger non-target sets can marginally improve retention, but the gain beyond 10% is modest and comes at additional data and compute cost. Us-

ing semantically distant classes for $D_b$ (configuration 4, `truck` and `cat`) keeps Accft at 0.02 but reduces Accrt to 0.81. In this case, the non-target set no longer shares the same visual factors as the target (wings, fuselage, sky background), so the generalized eigenanalysis cannot reliably isolate directions that are truly target-specific, and more shared structure is inadvertently attenuated.

Finally, removing the non-target set entirely (configuration 5) causes the method to degenerate into a purely target-driven attenuation scheme. Forgetting remains strong (Accft $= 0.02$), but retention collapses to Accrt $= 0.76$, comparable to aggressive training-based unlearning. This demonstrates that the contrastive formulation is essential: without a meaningful background set, the subnet cannot distinguish target-salient channels from channels that also support non-target classes.

Overall, these ablations show that CSE critically depends on a small, but semantically aligned, non-target set. Too few samples (5% per class) or non-overlapping classes significantly reduce Accrt, and removing $D_b$ entirely causes severe collateral damage. A design based on 10% of images per semantically related non-target class produces strong forgetting and near-baseline retention at modest computational cost.

## E.2. Sensitivity to Coverage and Sample Budget

The subnet discovered by CSE is controlled by two main hyperparameters: the coverage threshold $\tau_{\text{cov}}$, which determines how much discriminative mass the subnet must capture, and the number of non-target samples per class, $n_b$, used to estimate the covariances. In addition, we can either greedily select channels based on sorted salience or use an (idealized) oracle that chooses the best subset in hindsight.

To characterize sensitivity, we perform a controlled ablation on CIFAR-10 (forgetting `airplane` with ResNet-18) where we vary $\tau_{\text{cov}}$, the selection strategy, and the sample budget $n_b$. For each configuration we report: the fraction of channels selected per block (Ch.%), the forget-test accuracy (Accft), the retain-test accuracy (Accrt), the resulting H-Mean between test-time forgetting and retention, and the wall-clock time to run CSE on this setting. Results are summarized in Table 6.

Varying the coverage threshold shows that more aggressive subnet selection ($\tau_{\text{cov}} = 0.70$) reduces the number of channels (Ch.% $= 12$) and slightly harms forgetting (Accft $= 0.08$) and retention (Accrt $= 0.90$), leading to H-Mean $= 0.89$. Raising the threshold to 0.85 increases the selected subnet to 18% of channels and yields Accft $= 0.02$, Accrt $= 0.93$, and H-Mean $= 0.95$, a strong operating point. Pushing the coverage to 0.95 increases the subnet to 31% of channels and slightly improves forgetting (Accft $= 0.01$) at nearly unchanged retention (Accrt $= 0.92$), but at a marginally higher runtime. This suggests that discriminative information is somewhat distributed, but a coverage of 0.85 already captures enough mass to match or exceed the performance gains of higher coverage without unnecessary subnet growth.

When comparing selection strategies, greedy selection (sorting channels by salience and taking the smallest prefix that meets the coverage constraint) significantly outperforms random selection at the same coverage. Random selection yields Accft $= 0.15$ and Accrt $= 0.84$ with H-Mean $= 0.82$, whereas greedy selection achieves Accft $= 0.02$, Accrt $= 0.93$, and H-Mean $= 0.95$ while selecting the same fraction of channels. The oracle selection, which is allowed to search over subsets in hindsight, improves H-Mean only marginally (0.96) at the cost of an extreme computational overhead (8,420 seconds in this experiment), illustrating that the greedy heuristic is nearly optimal while being orders of magnitude faster.

Finally, varying the non-target sample budget shows that CSE remains stable even with relatively few samples per class. Using $n_b = 5$ samples per class yields Accft $= 0.02$, Accrt $= 0.85$, and H-Mean $= 0.87$, indicating some degradation but still reasonable performance. Increasing to $n_b = 10$ achieves Accft $= 0.02$, Accrt $= 0.93$, and H-Mean $= 0.95$, and further increases to $n_b = 20$ or 50 do not significantly change the metrics. Runtime grows only slightly with $n_b$. Overall, $\tau_{\text{cov}} = 0.85$, greedy selection, and $n_b = 10$ form a robust and efficient default that balances forgetting strength, retention, and computational cost.

## E.3. Qualitative Grad-CAM Examples

To complement the quantitative metrics, we examine Grad-CAM visualizations before and after applying CSE. For class-level forgetting, we consider examples where the model is originally trained on a source dataset (for instance, CIFAR-10 or ImageNet) and evaluated on a different dataset that shares the same semantic class.

A typical cross-dataset airplane example consists of three rows: the original image (an airplane in a novel pose or background), the Grad-CAM heatmap before unlearning, and the Grad-CAM heatmap after CSE. Before unlearning, Grad-CAM strongly highlights the fuselage, wings, and tail, indicating that the classifier relies on these regions to recognize the airplane. After CSE, the heatmap on the airplane body collapses to a nearly uniform, low-intensity pattern; activation shifts either to irrelevant background textures or disappears entirely. This corresponds to a drop in classification confidence for the airplane class to near-random levels. Importantly, when the same procedure is applied to non-target classes such as warplanes or birds, the post-CSE Grad-CAM maps remain sharply focused on the relevant object parts (engines, wings, or bodies), indicating that the underlying representation for related but non-target concepts is preserved.

Table 6. Ablation on subnet coverage, selection strategy, and non-target sample budget for CIFAR-10 single-class forgetting (`airplane`, ResNet-18). Ch.%: fraction of channels selected per block. Accft: forget-test accuracy (lower is better). Accrt: retain-test accuracy (higher is better). H-Mean: harmonic mean of test-time forgetting and retention (higher is better). Time: end-to-end CSE runtime in seconds for this configuration. Default settings are emphasized.

| | $\tau_{\text{cov}}$ | | | Selection | | | $n_b$ (samples per class) | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Setting | 0.70 | **0.85** | 0.95 | Random | **Greedy** | Oracle | 5 | **10** | 20 | 50 |
| Ch.% | 12 | **18** | 31 | 18 | **18** | 19 | – | – | – | – |
| Accft ↓ | 0.08 | **0.02** | 0.01 | 0.15 | **0.02** | 0.01 | 0.02 | **0.02** | 0.02 | 0.01 |
| Accrt ↑ | 0.90 | **0.93** | 0.92 | 0.84 | **0.93** | 0.94 | 0.85 | **0.93** | 0.93 | 0.94 |
| H-Mean ↑ | 0.89 | **0.95** | 0.94 | 0.82 | **0.95** | 0.96 | 0.87 | **0.95** | 0.95 | 0.96 |
| Time (s) | 8.2 | **8.5** | 9.1 | 0.3 | **8.5** | 8420 | 7.8 | **8.5** | 8.9 | 9.2 |

Similar behavior is observed in cross-dataset truck and shark experiments. For trucks, CSE suppresses saliency on the target truck in evaluation images while retaining strong, localized heatmaps on non-target automobiles. For sharks, CSE attenuates activations on shark bodies and fins, while non-target fish such as trout and rays continue to produce coherent Grad-CAM maps centered on the fish. Across these examples, the visual evidence supports the interpretation of CSE as a localized erasure mechanism: target saliency is removed while non-target structure and geometry are largely maintained.

In an identity-forgetting setting on faces, CSE is applied to a pre-trained face recognition model. For the target identity, Grad-CAM before unlearning shows strong saliency concentrated on distinctive facial features (eyes, nose, mouth, hairstyle). After CSE, the corresponding heatmaps become diffuse and low-intensity, and verification accuracy for the target identity drops to near chance. For non-target identities, the Grad-CAM maps remain crisp and well-localized on faces, indicating that CSE avoids global disruption of the face embedding space.

## F. Additional Related Work

Beyond the methods benchmarked in our main experiments, several recent works address class-centric unlearning in vision models from complementary angles. **Distillation-based approaches** use teacher–student frameworks to selectively forget: Bad Teacher [8] pairs a competent and an incompetent teacher to guide forgetting via knowledge distillation, while SCRUB [23] leverages teacher outputs on remaining data to preserve non-target knowledge. [19] originally introduced knowledge distillation for model compression; its adaptation to unlearning highlights how "dark knowledge" in logits can be repurposed for selective erasure. **Saliency- and instance-level methods** offer finer-grained control: SalUn [12] identifies gradient-based weight saliency maps to steer unlearning in both classification and generation, Learn to Unlearn [6] derives per-instance replacement labels via adversarial attacks against the frozen model, and [34] propose a fast post-hoc unlearning scheme using noise-based weight perturbation without retraining. In the **generative domain**, [14] erase semantic concepts from diffusion models by editing cross-attention layers, extending concept removal beyond discriminative classifiers. Machine unlearning has also been applied to **backdoor defense**: [25] show that targeted unlearning can neutralize data-poisoning attacks without full retraining. Finally, [37] study continual forgetting in pre-trained vision models, addressing the sequential multi-class setting where classes are forgotten incrementally over time. In contrast to all of the above, CSE is training-free, operates directly in channel space via contrastive eigenanalysis, and introduces no inference-time overhead through algebraic fold-in—properties that make it particularly suited to deployment-constrained settings.

## G. Scalability and Representation Geometry

**Same-dataset stress test.** To evaluate CSE under aggressive forgetting within a single dataset, we forget $k \in \{6, 7, 8\}$ classes simultaneously on CIFAR-10 (ResNet-18). As shown in Table 7, CSE maintains high retention accuracy even as the number of forgotten classes grows ($\text{Acc}_{\text{rt}} = 0.90$ at $k = 8$), closely tracking the Retrain baseline (0.89), while DELETE degrades more noticeably (0.82). This stability arises because CSE edits the encoder selectively—attenuating only a compact, contrastive set of target-salient channels and preserving shared features.

Table 7. Same-dataset stress test (CIFAR-10): forget $k \in \{6, 7, 8\}$ classes. Entries are $\text{Acc}_{\text{ft}}/\text{Acc}_{\text{rt}}$.

| **Method** | $k$=6 | $k$=7 | $k$=8 |
|---|---|---|---|
| Original | 0.95/0.94 | 0.95/0.94 | 0.95/0.94 |
| Retrain | 0.00/0.91 | 0.00/0.90 | 0.00/0.89 |
| DELETE | 0.15/0.86 | 0.17/0.84 | 0.19/0.82 |
| **CSE** | **0.04/0.92** | **0.05/0.91** | **0.06/0.90** |

**Retain-geometry audit.** Beyond accuracy, we verify that

CSE preserves the geometric structure of non-target representations. We compare model embeddings before and after unlearning using two metrics: (i) linear CKA (Centered Kernel Alignment) [21] and (ii) correlation of pairwise class-centroid distances (DistCorr), each computed separately on retained and forgotten class subsets. Table 8 confirms that CSE best preserves non-target geometry (highest $\text{CKA}_r$ and $\text{DistCorr}_r$) while most strongly disrupting the forgotten concept (lowest $\text{CKA}_f$ and $\text{DistCorr}_f$), consistent with the accuracy trends reported in the main paper.

Table 8. Geometry audit. Representation similarity before vs. after unlearning on retained ($r$) and forgotten ($f$) sets.

| **Method** | $\textbf{CKA}_r \uparrow$ | $\textbf{CKA}_f \downarrow$ | $\textbf{DistCorr}_r \uparrow$ | $\textbf{DistCorr}_f \downarrow$ |
|---|---|---|---|---|
| DELETE | 0.86 | 0.62 | 0.84 | 0.59 |
| ESC | 0.88 | 0.58 | 0.86 | 0.55 |
| **CSE** | **0.96** | **0.31** | **0.95** | **0.28** |

## References


[1] Deniz Bayazit, Negar Foroutan, Zeming Chen, Gail Weiss, and Antoine Bosselut. Discovering knowledge-critical subnetworks in pretrained language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6549–6583, 2024. 4

[2] Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. Leace: Perfect linear concept erasure in closed form. *Advances in Neural Information Processing Systems*, 36:66044–66063, 2023. 2

[3] Jacopo Bonato, Marco Cotogni, and Luigi Sabetta. Is retain set all you need in machine unlearning? restoring performance of unlearned models with out-of-distribution images. In *European Conference on Computer Vision*, pages 1–19. Springer, 2024. 2, 3

[4] Lucas Bourtoule, Varun Chandrasekaran, Christopher A Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *2021 IEEE symposium on security and privacy (SP)*, pages 141–159. IEEE, 2021. 1, 8

[5] Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *2015 IEEE symposium on security and privacy*, pages 463–480. IEEE, 2015. 1

[6] 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*, pages 11186–11194, 2024. 19

[7] 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*, pages 7766–7775, 2023. 2, 3, 8

[8] Vikram S Chundawat, Ayush K Tarun, Murari Mandal, and Mohan Kankanhalli. Can bad teaching induce forgetting? unlearning in deep networks using an incompetent teacher. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7210–7217, 2023. 19

[9] Kate Crawford and Trevor Paglen. Excavating ai: The politics of images in machine learning training sets. *Ai & Society*, 36(4):1105–1116, 2021. 1

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. IEEE, 2009. 5

[11] Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012. 3

[12] 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. *arXiv preprint arXiv:2310.12508*, 2023. 19

[13] Jack Foster, Stefan Schoepf, and Alexandra Brintrup. Fast machine unlearning without retraining through selective synaptic dampening. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024. 8

[14] Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2426–2436, 2023. 19

[15] Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9304–9312, 2020. 8

[16] Laura Graves, Vineel Nagisetty, and Vijay Ganesh. Amnesiac machine learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11516–11524, 2021. 8

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[18] Alvin Heng and Harold Soh. Selective amnesia: A continual learning approach to forgetting in deep generative models. *Advances in Neural Information Processing Systems*, 36: 17170–17194, 2023. 2

[19] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 8, 19

[20] Gary B Huang, Manu Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In *Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition*, 2008. 8

[21] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pages 3519–3529. PMLR, 2019. 20

[22] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 5

[23] Meghdad Kurmanji, Peter Triantafillou, Jamie Hayes, and Eleni Triantafillou. Towards unbounded machine unlearning. *Advances in neural information processing systems*, 36: 1957–1987, 2023. 19

[24] Tae-Young Lee, Sundong Park, Minwoo Jeon, Hyoseok Hwang, and Gyeong-Moon Park. Esc: Erasing space concept for knowledge deletion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5010–5019, 2025. 2, 3

[25] Yang Liu, Mingyuan Fan, Cen Chen, Ximeng Liu, Zhuo Ma, Li Wang, and Jianfeng Ma. Backdoor defense with machine unlearning. In *IEEE INFOCOM 2022-IEEE conference on computer communications*, pages 280–289. IEEE, 2022. 19

[26] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 5

[27] Martin Pawelczyk, Jimmy Z Di, Yiwei Lu, Ayush Sekhari, Gautam Kamath, and Seth Neel. Machine unlearning fails to remove data poisoning attacks. *arXiv preprint arXiv:2406.17216*, 2024. 2

[28] Achyuta Rajaram, Neil Chowdhury, Antonio Torralba, Jacob Andreas, and Sarah Schwettmann. Automatic discovery of visual circuits. *arXiv preprint arXiv:2404.14349*, 2024. 4

[29] Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. *arXiv preprint arXiv:2004.07667*, 2020. 1

[30] Shauli Ravfogel, Michael Twiton, Yoav Goldberg, and Ryan D Cotterell. Linear adversarial concept erasure. In *International Conference on Machine Learning*, pages 18400–18421. PMLR, 2022. 1

[31] Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22522–22531, 2023. 1

[32] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE, 2017. 5

[33] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 5

[34] 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, 2023. 19

[35] Paul Voigt and Axel Von dem Bussche. The eu general data protection regulation (gdpr). *A practical guide, 1st ed., Cham: Springer International Publishing*, 10(3152676):10–5555, 2017. 1

[36] Zeliang Zhang, Gaowen Liu, Charles Fleming, Ramana Rao Kompella, and Chenliang Xu. Targeted forgetting of image subgroups in clip models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9870–9880, 2025. 2, 3

[37] Hongbo Zhao, Bolin Ni, Junsong Fan, Yuxi Wang, Yuntao Chen, Gaofeng Meng, and Zhaoxiang Zhang. Continual forgetting for pre-trained vision models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28631–28642, 2024. 19

[38] 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 *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 20350–20359, 2025. 2, 3, 5, 8