Quantifying Representation Collapse: Structural Entanglement in Document Image Unlearning

Amisha Mishra[1] and Chiranjoy Chattopadhyay[1]

School of Computing and Data Sciences, FLAME University, Pune, India
{amisha.mishra, chiranjoy.chattopadhyay}@flame.edu.in

Abstract Machine unlearning (MU) formalizes the process of removing the influence of a forget-set $\mathcal{D}_f$ from a trained model parameter space θ to approximate a retrain-from-scratch (RfS) distribution $P(\theta_{rfs}|\mathcal{D} \setminus \mathcal{D}_f)$. Although distribution-matching and sensitivity-based unlearning algorithms exhibit bounded error in natural image datasets through localized texture unlearning, the Document Image Understanding (DIU) domain relies on structural topologies that fundamentally violate the feature-independence assumptions of these metrics. In this paper, we establish the theoretical and empirical failure of canonical MU metrics—specifically the Fisher Information Matrix, a parameter sensitivity measure, and distributional distance metrics - Total Variation (TV) and Wasserstein-1 divergence (W_1) — when applied to document representations. Through multi-cardinality evaluations on RVL-CDIP and Devanagari datasets, we quantify *structural entanglement*, demonstrating that the geometric basis vectors required for $\mathcal{D}_f$ (e.g., script primitives, layout grids) are inherently shared with $\mathcal{D}_r$. Utilizing Centered Kernel Alignment (CKA) and manifold mapping, we prove that targeted unlearning of document structures induces a global representation collapse rather than localized forgetting, rendering standard MU algorithms theoretically invalid for sparse, topology-driven domains.

Keywords: Machine Unlearning · Document Image Analysis · Wasserstein Distance · Fisher Information · Feature Entanglement

1 Introduction

Machine unlearning (MU) seeks to map a pre-trained parameter state θ_{orig}, optimized over a dataset $\mathcal{D} = \mathcal{D}_r \cup \mathcal{D}_f$ (where $\mathcal{D}_r$ and $\mathcal{D}_f$ denote the retain and forget sets, respectively), to an unlearned state θ_u. Formally, MU bounds the expected distance D between the unlearned model and the theoretical retrain-from-scratch (RfS) state θ_{rfs} generated by learning algorithm $\mathbb{A}$: $\mathbb{E}[D(\mathbb{A}(\mathcal{D}, \mathcal{D}_f), \mathbb{A}(\mathcal{D}_r, \emptyset))] \leq \epsilon$, given an error tolerance ϵ. Because exact MU methods(e.g., [1, 7]) impose prohibitive retraining costs, approximate methods pose MU as a parameter-pruning optimization problem. These utilize distinct sensitivity and distributional metrics—such as the Fisher Information Matrix (FIM) - a parameter sensitivity measure [4], Total Variation Distance (TV),

and Wasserstein metric (W_1) — distributional distance metrics [2,13], providing critical defense against Membership Inference Attacks (MIA) [15]. The efficacy of these sensitivity-based methods is based on a fundamental vision-centric premise: that class-specific representations in the latent space are localized and orthogonal. This holds true for natural images (e.g., CIFAR-10 [11]), where convolutional kernels act as probabilistically independent texture extractors [3, 4].

In contrast, Document Image Understanding (DIU) operates on sparse, topologically rigid signals where semantic classes are defined by global spatial configurations rather than localized textures [6]. Despite this fundamental divergence in manifold topology, the rapid operationalization of privacy-preserving DIU has relied on the unverified assumption that texture-biased MU algorithms safely transfer to structured documents. This introduces a critical systemic failure mode that we term Structural Entanglement. In the DIU feature space, for OCR-based tasks [16], geometric basis vectors—such as the *Shirorekha* (horizontal headline) in Devanagari script from the Devanagari Character Dataset (Deva-C-Set) or the tabular grid structures in financial invoices from RVL-CDIP act as universally shared structural primitives.

When distance-based metrics try to remove $\mathcal{D}_f$, they end up attributing unnaturally large gradients to the underlying feature extractors, inadvertently suppressing shared high-magnitude structural filters. In MU for document images [8], global distribution-shift metrics do not yield a cleanly partitioned submanifold; instead, removing these primitives causes catastrophic structural forgetting [9] and collapses the geometric priors needed to preserve the utility of $\mathcal{D}_r$. This collapse is absent in textured natural-image datasets such as CIFAR-10, where distributed texture statistics reduce reliance on a small set of shared structural primitives. The Centered Kernel Alignment (CKA) [10] was proposed to evaluate the similarity of the internal network independently of the classifier.

The primary contributions of this work are threefold:

1. We empirically establish and quantify the failure of canonical sensitivity metrics – Fisher Information Matrix (FIM), TV, W_1, to achieve bounded-error MU in topological document representations.
2. We formalize the concept of *structural entanglement*, utilizing layer-wise CKA to prove that standard MU filter-pruning on DIU datasets induces global representation collapse rather than localized forgetting.
3. We provide a fine-grained analysis of this failure across macro-structures (high-cardinality layouts in RVL-CDIP [5]) and micro-structures (connected script primitives in Devanagari [14]), establishing stringent baseline requirements for future document-native unlearning architectures.

The remainder of this paper is organized as follows. Section 2 develops the theoretical divergence between texture- and topology-based representations. Section 3 formalizes the approximate unlearning methods and distance metrics. Section 4 analyzes unlearning performance, quantifying catastrophic collapse across macro- and micro-document scales. Section 5 empirically validates the collapse of the representational manifold across scales, while Sec. 6 concludes.

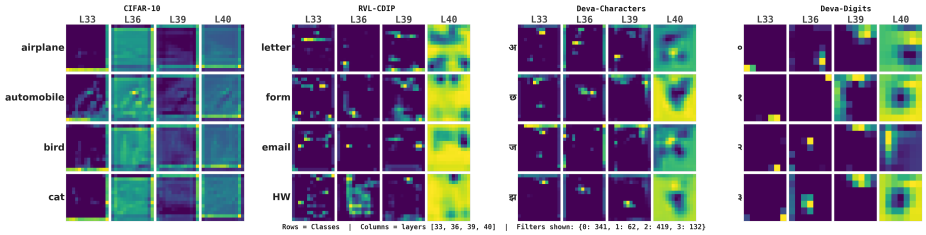


Figure 1: [a-d] Representative layer activations (VGG-16, L33, L36, L39, L40) illustrating the topological gap: (a) natural images depend on dense, localized textures, while (b–d) document domains depend on sparse, global geometric primitives. Filters shown are selected randomly.

2 Analysis of Domain-Specific Feature Disparity

To understand the catastrophic failure of these canonical unlearning metrics for DIU, we distinguish the representation manifolds of natural images from those of structured documents. The fundamental incompatibility stems from the difference between texture orthogonality and topological entanglement.

2.1 Texture vs. Topology in CNN Representations

In standard computer vision classification (e.g., CIFAR-10), Convolutional Neural Networks (CNNs) converge to localized high-frequency texture representations. The receptive fields in intermediate layers act as quasi-independent basis vectors for class-specific textures (e.g., extracting canine fur versus metallic gradients). Because the representational sub-manifolds for these classes are predominantly orthogonal, the parameters encoding class A have minimal intersection with the parameters encoding class B. Consequently, sensitivity metrics such as the diagonal of the FIM, successfully isolate the subset of parameters $\theta_f \subset \theta$ responsible for $\mathcal{D}_f$. Excising θ_f achieves unlearning because the spatial independence of the features ensures that the representation of $\mathcal{D}_r$ remains largely intact.

Document images are information-sparse, high-contrast matrices that lack discriminative texture. To isolate the structural root of the failure of unlearning, we qualitatively analyze the filters present at different layers of the network (Figure 1). While texture-dense control (a) exhibits spatially smooth activations, document (b) and script (c), (d) domains are characterized by extreme activation sparsity. The convolutional filters in DIU optimize for long-range spatial dependencies—horizontal text rulings, tabular grids, and morphological anchors like the *Shirorekha*. This qualitative effect is supported by a quantitative layer-wise metric analysis (Figure 2). The DIU domains show increased activation sparsity (Figure 2a) and a shift toward low–spatial-frequency components (Figure 2b) in the deeper structural layers. In contrast, for the CIFAR-10 baseline, the filter response entropy stays high across layers (Figure 2c).

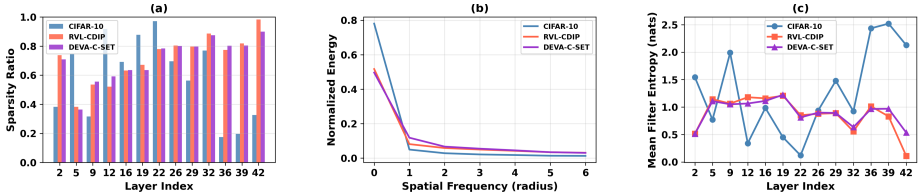


Figure 2: Quantitative layer-wise profiling of representational divergence in VGG-16: (a) Activation Sparsity, (b) Spatial Frequency Profile(Across all layers), and (c) Filter Response Entropy.

2.2 The Primitive Problem: Shared Basis Vectors

As seen in Figure 2, the extreme sparsity and low-frequency nature of DIU signals introduce a severe feature-overlap bottleneck, which we term the *Primitive Problem*. In the latent space of a document-trained model, classes are not constructed from distinct orthogonal feature sets. Instead, they are combinatorial arrangements of a finite set of universally shared geometric basis vectors.

Mathematically, this shared dependence breaks the core assumption of sensitivity based unlearning. Consider a pruning mechanism guided by the empirical Fisher Information, F. For a specific filter weight θ_i, its importance to $\mathcal{D}_f$ is proportional to its Fisher Information $F_i(\mathcal{D}_f)$. In a natural image dataset, a high $F_i(\mathcal{D}_f)$ compared to $F_i(\mathcal{D}_r)$, implies that θ_i is highly discriminative and exclusive to $\mathcal{D}_f$. However, in the DIU domain, if θ_i encodes a foundational primitive—such as an invoice tabular grid—its activation magnitude will be universally high. Consequently, the FIM calculates a massive sensitivity score $F_i(\mathcal{D}_r)$ for this foundational filter when prompted to unlearn a specific layout class.

Standard MU algorithms prioritize weight removal with high sensitivity F_i. In DIU, this leads to a fundamental conflict because $\theta(\mathcal{D}_f) \cap \theta(\mathcal{D}_r) \approx \theta(\mathcal{D}_r)$. Since foundational structural primitives (e.g., rectilinear grids or headline strokes) are shared across classes, targeting weights associated with $\mathcal{D}_f$ inadvertently removes the geometric priors necessary for $\mathcal{D}_r$. Consequently, the resulting state θ_u suffers from a global loss of structural utility rather than a targeted class-forgetting.

3 Methodology and Unlearning Baselines

All experiments are implemented in PyTorch and executed on a single NVIDIA RTX A4500 GPU (20 GB VRAM) with CUDA 13.2. The backbone is VGG16 network with Batch Normalization (VGG16-BN), initialized with weights pre-trained on ImageNet [8], and subsequently fine-tuned on RVL-CDIP [5] for 10 epochs using the Adam optimiser with a learning rate of 10^{-4} and a batch size of 32. Input images are resized to 224×224. Same experimental conditions were followed for all the datasets. To empirically validate the hypothesis of structural

entanglement, we formalize the MU objective and explicitly define the metrics utilized to approximate forgetting. We evaluate three canonical metrics widely used in the parameter pruning space—FIM, TV distance, and W_1 distance.

3.1 The Unlearning Objective and Pruning Framework

Let $\theta_{orig} \in \mathbb{R}^d$ represent the parameter space of a CNN optimized over the complete dataset $\mathcal{D} = \mathcal{D}_r \cup \mathcal{D}_f$. The objective of filter-based MU is to compute a binary pruning mask $M \in \{0,1\}^d$, such that the unlearned model $\theta_u = \theta_{orig} \odot M$ minimizes the representational distance to a RfS model θ_{rfs} exclusively trained on $\mathcal{D}_r$. Here, $\odot$ denotes the element-wise (Hadamard) product.

The mask M is derived by thresholding a sensitivity metric $S(\theta_i, \mathcal{D}_f)$ for each filter i [12]. For a given pruning ratio $\tau \in [0,1]$, the mask is defined as:

$$M_i = \begin{cases} 0, & \text{if } S(\theta_i, \mathcal{D}_f) \geq P_\tau(S(\theta, \mathcal{D}_f)) \\ 1, & \text{otherwise} \end{cases} \tag{1}$$

where P_τ denotes the τ-th percentile of the sensitivity distribution. In our framework, the sensitivity metric S is instantiated using three distinct probabilistic distances, as discussed below.

1. Fisher Information Matrix (FIM): The FIM measures the sensitivity of the model's log-likelihood to variations in its parameters. To ensure computational tractability across deep document architectures, we utilize the diagonal approximation of the empirical Fisher Information for a filter i over $\mathcal{D}_f$:

$$F_{i,i}(\mathcal{D}_f) = \frac{1}{|\mathcal{D}_f|} \sum_{x \in \mathcal{D}_f} \left(\nabla_{\theta_i} \log p(y|x; \theta_{orig})\right)^2 \tag{2}$$

Under the standard MU assumption, the parameters with maximum $F_{i,i}$ encode $\mathcal{D}_f$ exclusively.

2. Total Variation (TV) Distance: As a stricter bound for bounding worst-case mass shifts, TV distance is utilized to prune filters that directly dictate the maximum absolute difference in class probabilities for $\mathcal{D}_f$:

$$\delta(P_{\theta_{orig}}, P_{\theta_u}) = \frac{1}{2} \sum_{x \in \mathcal{D}_f} |P(x|\theta_{orig}) - P(x|\theta_u)| \tag{3}$$

3. Wasserstein-1 (W_1) Distance: Unlike TV which operates on output probabilities, the Wasserstein distance evaluates the geometric shift in the latent feature space Z. For representations $z \sim Z(\mathcal{D}_f)$, we measure the optimal transport cost to map the unlearned manifold to the original manifold:

$$W_1(\mathbb{P}_{orig}, \mathbb{P}_u) = \inf_{\gamma \in \Pi(\mathbb{P}_{orig}, \mathbb{P}_u)} \int_{Z \times Z} \|z_{orig} - z_u\| \, d\gamma(z_{orig}, z_u) \tag{4}$$

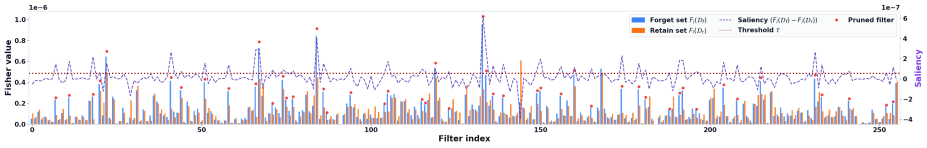


Figure 3: Magnitude-based filter pruning for a target class in a convolutional layer. Sensitivity score calculated using the FIM.

Filters whose ablation minimizes this optimal cost are aggressively pruned. Figure 3 shows this process on [5], where convolutional filters are compared to a sensitivity threshold (τ), and standard algorithms pruned 49 with $F_i > \tau$, where $\mathcal{D}_f = \{Specification\}$. Although mathematically optimal for disjoint domains, we hypothesize that with Equations 2 through 4, topologically entangled document datasets will fail to isolate $\mathcal{D}_f$, triggering structural collapse.

4 Quantitative Analysis of Performance Degradation

We structure our analysis to first establish the baseline validity of our unlearning pipeline on texture-dense natural images, followed by an exhaustive analysis of its failure on document layouts (macro) and script (micro) primitives.

4.1 Baseline Validation: Texture Unlearning (CIFAR-10)

To ensure that the observed failure in the DIU datasets is the product of domain topology rather than pipeline execution, we benchmarked the three canonical sensitivity metrics on the CIFAR-10 dataset. As depicted in the radar plots (Figure 4 (right)) and class-wise retention charts (Figure 5), the unlearning trajectory strictly adheres to theoretical expectations. For example, when targeting the 'Airplane' class via Fisher-based pruning, the forget accuracy $\mathcal{A}_f$ successfully converges to baseline random chance, while the retain accuracy $\mathcal{A}_r$ across

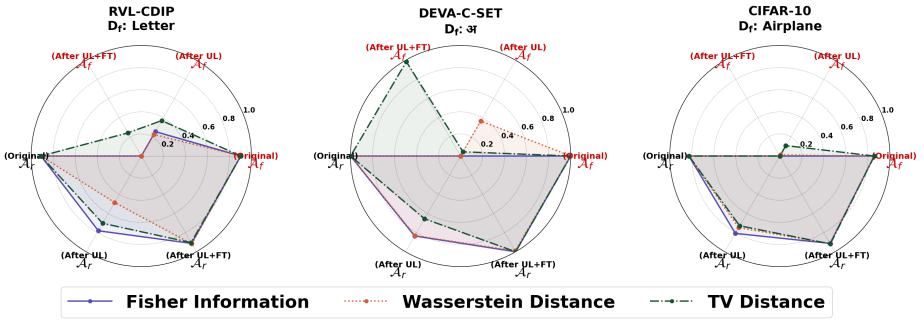


Figure 4: Unlearning performance comparison across datasets.

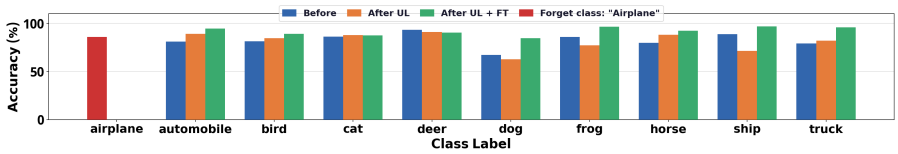


Figure 5: CIFAR-10 dataset class-wise retention performance of the pruned model. The forgotten class is 'Airplane'.

the remaining 9 classes suffers negligible degradation. Also, fine-tuning the unlearned model (1 epoch) on the retain set further suppressed $\mathcal{A}_f$ while recovering $\mathcal{A}_r$. This confirms that in a texture-dense, probabilistically independent feature space, convolutional filters act as localized class-specific extractors that can be excised without triggering global representation collapse.

4.2 Catastrophic Entanglement in Macro-Topology (RVL-CDIP)

When an identical unlearning optimization pipeline is applied to the RVL-CDIP dataset, the assumption of feature independence breaks down. Figure 4(left) provides an immediate visual representation of this instability, which shows unlearning performance on the 'Letter' class, exhibiting severe distortion in the retention manifold compared to the tightly bounded CIFAR-10 baseline. The empirical crash is rigorously quantified in Fig. 6. Using the diagonal FIM, we unlearn another randomly selected class, 'Questionnaire', and found the target class accuracy drops from 0.88 to 0.25. However, the $\mathcal{D}_r$ average accuracy drops symmetrically from 0.91 to 0.57. This secondary utility loss in the retain-set manifold is not uniform but is dictated by *Structural Entanglement*. This foun-

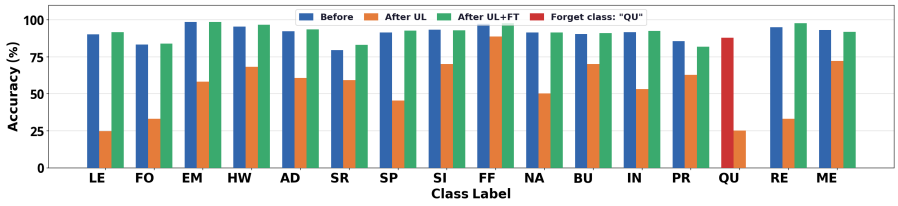


Figure 6: RVL-CDIP dataset class-wise retention performance of the pruned model. The three bars for each class represent the 'Original', 'after pruning' and 'after fine-tuning' performances. Abbreviations: LE: Letter, FO: Form, EM: Email, HW: Handwritten, AD: Advertisement, SR: Scientific Report, SP: Scientific Publication, SI: Specification, FF: File Folder, NA: News Article, BU: Budget, IN: Invoice, PR: Presentation, QU: Questionnaire, RE: Resume, ME: Memo. The forgotten class is 'LE'.

Table 1: Performance Before(Original) and After UL for two forget-set sizes. Classes are abbreviated as in Fig. 6.

(a) $|\mathcal{D}_f| = 2$, ($\mathcal{D}_f$ = SR, SI)

Class	Original	After UL
Forget Set Accuracy $\mathcal{A}_f$		
SR	0.80	0.16
SI	0.93	0.06
Retain Set Accuracy $\mathcal{A}_r$		
LE	0.90	0.05
FO	0.83	0.08
EM	0.99	0.08
HW	0.96	0.38
AD	0.92	0.27
SP	0.91	0.12
FF	0.97	0.93
NA	0.91	0.49
BU	0.90	0.38
IN	0.92	0.02
PR	0.86	0.13
QU	0.88	0.12
RE	0.95	0.81
ME	0.93	0.47

(b) $|\mathcal{D}_f| = 3$, ($\mathcal{D}_f$ = LE, HW, QU)

Class	Original	After UL
Forget Set Accuracy $\mathcal{A}_f$		
LE	0.90	0.31
HW	0.96	0.00
QU	0.88	0.40
Retain Set Accuracy $\mathcal{A}_r$		
FO	0.83	0.76
EM	0.99	0.47
AD	0.92	0.73
SR	0.80	0.55
SP	0.91	0.90
FF	0.97	0.85
NA	0.91	0.89
BU	0.90	0.78
IN	0.92	0.72
PR	0.86	0.59
SI	0.93	0.59
RE	0.95	0.93
ME	0.93	0.77

dational problem of structural entanglement is catastrophically amplified by the high class cardinality inherent to document and script domains. As the total number of classes increases, the filters associated with these geometric primitives are reused, and thus linked to multiple retain-set categories.

As shown in Table 1, pruning structurally shared parameters to forget specific document classes causes measurable degradation between semantically related retain classes. With $|\mathcal{D}_f| = 2$, the forget classes 'SR' and 'SI' are substantially suppressed, dropping from 0.80 to 0.16 and from 0.93 to 0.06, respectively. However, this suppression also affects $\mathcal{D}_r$ classes that rely on similar layout priors and text-structure representations. A Catastrophic drop is observed in classes such as 'LE', 'FO', 'EM'. In contrast, visually distinct classes such as 'FF' remain preserved, indicating that their representations occupy comparatively isolated regions of the feature manifold. The effect is also observed with $|\mathcal{D}_f| = 3$. These observations support our formulation in Section 3: the FIM-based pruning mechanism removes high-saliency structural filters that are shared across many document categories, not specific to the forget classes. As a result, suppressing these shared geometric and textual representations degrades performance in adjacent retain classes, showing that unlearning in document models is fundamentally limited by representation overlap in the learned manifold.

Table 2: Unlearning results on RVL-CDIP across different metrics. Classes are abbreviated as Fig. 6. Performance is reported as Accuracy ($\mathcal{A} \in [0, 1]$).

Class	**Original**		**Fisher (FIM)**		**Wasserstein (W_1)**		**Total Variation (TV)**	
	Forget	**Retain**	**Forget**	**Retain**	**Forget**	**Retain**	**Forget**	**Retain**
LE	0.90	0.91	0.11	0.72	0.10	0.56	0.48	0.68
FO	0.83	0.92	0.35	0.80	0.19	0.69	0.52	0.71
EM	0.99	0.91	0.20	0.78	0.62	0.83	0.94	0.74
HW	0.96	0.91	0.16	0.75	0.21	0.68	0.95	0.76
AD	0.92	0.91	0.34	0.67	0.18	0.78	0.83	0.62
SR	0.80	0.92	0.29	0.66	0.78	0.66	0.62	0.80
SP	0.91	0.91	0.28	0.79	0.24	0.79	0.28	0.73
SI	0.93	0.91	0.74	0.69	0.50	0.76	0.49	0.60
FF	0.97	0.91	0.69	0.56	0.79	0.78	0.98	0.30
NA	0.91	0.91	0.59	0.76	0.37	0.77	0.42	0.67
BU	0.90	0.91	0.23	0.68	0.42	0.72	0.52	0.77
IN	0.92	0.91	0.43	0.77	0.16	0.63	0.57	0.69
PR	0.86	0.91	0.48	0.62	0.75	0.81	0.83	0.79
QU	0.88	0.91	0.25	0.57	0.32	0.70	0.07	0.62
RE	0.95	0.91	0.50	0.75	0.06	0.70	0.94	0.56
ME	0.93	0.91	0.14	0.64	0.31	0.70	0.70	0.59

4.3 Metric Invariance of the Structural Collapse

A potential critique of the observed representation collapse is that the Fisher Information Matrix (FIM) constitutes a suboptimal sensitivity metric for sparse domains. To systematically eliminate this variable, we executed an exhaustive ablation across three fundamentally distinct sensitivity metrics: FIM (gradient sensitivity), Wasserstein-1 distance (latent optimal transport cost), and Total Variation (output probability mass shift).

The empirical results demonstrate that the structural collapse is invariant across evaluation metrics. As reported in Table 2, targeting the ‘Letter‘ class induces severe degradation on the retain set regardless of the underlying mathematical formulation. Under the Fisher Information criterion, the retain accuracy $\mathcal{A}_r$ decreases from 0.91 to 0.71. When employing the Wasserstein-1 distance, the model exhibits the most pronounced collapse, with $\mathcal{A}_r$ dropping from 0.91 to 0.56. A similar trend is observed for the Total Variation metric, where $\mathcal{A}_r$ declines to 0.68, while the method fails to effectively erase the targeted information from the forget set, as evidenced by the persistently high forget-set accuracy $\mathcal{A}_f = 0.48$. Similar behavior is also observed for several other classes. For instance, the ‘Form‘ (FO) class experiences a substantial retain-set degradation under Fisher Information, where $\mathcal{A}_r$ decreases from 0.92 to 0.75 while the forget accuracy remains relatively high at 0.45. The ‘Handwritten‘ (HW) and ‘Email‘ (EM) classes exhibit even stronger instability under the Total Variation metric, with forget accuracies rising to 0.95 and 0.94, respectively, despite noticeable reductions in retain performance. Likewise, the ‘Form Feed‘ (FF) class under Total Variation demonstrates an extreme imbalance, achieving near-complete

forgetting failure with $\mathcal{A}_f = 0.98$ while simultaneously collapsing the retain accuracy to 0.30. Collectively, these findings indicate that the structural crash is a robust phenomenon that persists across different unlearning objectives and is not an artifact of any particular metric choice.

The poor performance of the Wasserstein-1 metric (W_1) provides critical insight into the geometry of document representations. Because W_1 calculates the optimal transport cost to map the unlearned feature space to the original space, it heavily penalizes any deviation in dense, high-magnitude features. In DIU, these high-magnitude features are precisely the shared topological priors (grids, margins). Consequently, optimal transport-based unlearning extensively prunes essential structural basis vectors, leading to the lowest retain accuracy.

4.4 Micro-Level Analysis: Devanagari Scripts

We experimented with the Devanagari character and digit datasets to prove this phenomenon persists at the micro-topological level. The Devanagari script is geometrically defined by shared, continuous sub-primitives, most notably the *Shirorekha* (horizontal headline) and vertical anchoring stems.

Figure 7 illustrates the retention performance in class after unlearning the target character 'अ'. Unlike the CIFAR-10 baseline (where unlearning 'Airplane' leaves 'Automobile' perfectly intact), unlearning 'अ' triggers a generalized performance degradation across topologically adjacent classes. Figure 7 reveals that the accuracy of topologically adjacent classes drops by over 50%. For example, upon forgetting class अ, the retain performance of classes इ and झ dropped substantially from their original performance to as low as 0.11 and 0.02, respectively.

This structural degradation is deterministic, driven by shared morphology. Sensitivity metrics assign highest importance to deep filters detecting primitives like the *Shirorekha* and the vertical stem of 'अ', these filters are aggressively pruned. Yet these primitives are also needed to classify geometrically adjacent characters, causing a global loss of feature discriminability. This micro-level analysis reveals in DIU unlearning: when classes are compositional arrangements of shared primitives, orthogonal feature isolation is mathematically impossible.

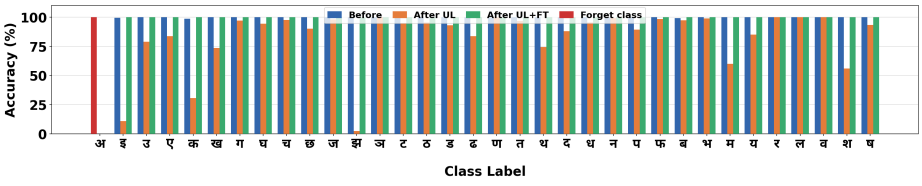


Figure 7: Devanagari character dataset class-wise retention performance of the pruned model. The three bars for each class represent the 'Original', 'after pruning' and 'after fine-tuning' performances. The forgotten class is 'अ'.

Table 3: Layer-wise Fisher Statistics and Pruning Details - VGG16 - RVL-CDIP

Metric	L17	L20	L24	L27	L30	L34	L37	L40
μ	-4.9647×10^{-8}	-3.0370×10^{-8}	4.4060×10^{-8}	3.8715×10^{-8}	8.5751×10^{-8}	1.2782×10^{-6}	3.8444×10^{-8}	2.0242×10^{-6}
σ	6.3522×10^{-8}	1.8444×10^{-7}	4.3298×10^{-7}	1.3533×10^{-7}	3.5216×10^{-7}	3.3121×10^{-6}	1.5051×10^{-6}	1.4162×10^{-5}
Threshold	1.1708×10^{-9}	1.1718×10^{-7}	3.9044×10^{-7}	1.4698×10^{-7}	3.6748×10^{-7}	3.9278×10^{-6}	1.2425×10^{-6}	1.3354×10^{-5}
Filters Pruned	28/256	29/256	37/512	42/512	55/512	39/512	44/512	18/512

5 Analyzing the Representation Collapse

Although the empirical evaluation in Section 4 quantifies the degradation of the network output-level, it does not isolate the internal mechanism of the failure. To definitively prove the *structural entanglement* hypothesis, we analyze the performance with respect to the model's internal parameter representations.

5.1 Deep-Layer Pruning and Structural Ablation

In CNNs, hierarchical feature extraction dictates that early layers (e.g., L1–L10) capture localized, high-frequency textures, while deep layers (e.g., L24–L40) encode macro-spatial relationships and topological priors. The empirical layer-wise pruning statistics directly contradict the texture assumption. As detailed in Table 3 (Fisher Information) and Table 4 (W_1), the sensitivity metrics compute peak importance exactly within the deep structural layers. Under the W_1 metric (Table 4), the pruning mechanism ablates up to 18.4% of the filters in L30 (94/512 filters) and L34 (94/512 filters). Similarly, Fisher-based pruning (Table 3) exhibits maximum filter removal in L30 and L37. Because layers L24–L40 build the document's global topological grid, removing these filters physically alters the deep architecture responsible for global topological modeling, rather than cleanly removing a single class manifold.

5.2 Consistency of Representation Collapse

Table 5 empirically validates that representation collapse is a domain-specific consequence of structural entanglement. In the texture-based CIFAR-10, unlearning is highly selective; excising the "Airplane" class yields an 85.30% drop in forget accuracy while maintaining a negligible 0.59% impact on the retain set. Conversely, both the macro-layout (RVL-CDIP) and micro-script (DEVA-C-SET) domains exhibit severe degradation. For instance, successfully unlearning the Devanagari character 'क' (98.70% forget drop) triggers a massive 52.94%

Table 4: Layer-wise W_1 Statistics and Pruning Details - VGG16 : RVL-CDIP

Metric	L17	L20	L24	L27	L30	L34	L37	L40
μ	4.1633×10^{-17}	1.3878×10^{-17}	-7.8063×10^{-17}	4.1633×10^{-17}	-1.0408×10^{-17}	-6.5919×10^{-17}	8.2399×10^{-17}	-1.0408×10^{-17}
σ	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Threshold	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8
Filters Pruned	47/256	49/256	88/512	88/512	94/512	94/512	92/512	31/512

Table 5: Comparison of Forget-Retain Drop across RVL-CDIP, DEVA-C-SET, and CIFAR-10 datasets. CIFAR-10 classes are abbreviated as Fig. 6.

RVL-CDIP			DEVA-C-SET			CIFAR-10		
Class Name	**Forget Drop**	**Retain Drop**	**Class Name**	**Forget Drop**	**Retain Drop**	**Class Name**	**Forget Drop**	**Retain Drop**
PR	85.58	67.65	इ	99.30	26.10	Airplane	85.30	0.59
QU	62.67	34.54	क	98.70	52.94	Automobile	76.00	3.20
ME	86.52	28.26	ह	100.00	29.07	Bird	71.70	-0.54

reduction in retain accuracy. Similarly, unlearning the 'PR' layout in RVL-CDIP induces a 67.65% retain drop. This scale-independent failure confirms that current distance-based algorithms cannot untangle class-specific identities from the shared topological load-bearers required by the broader document manifold.

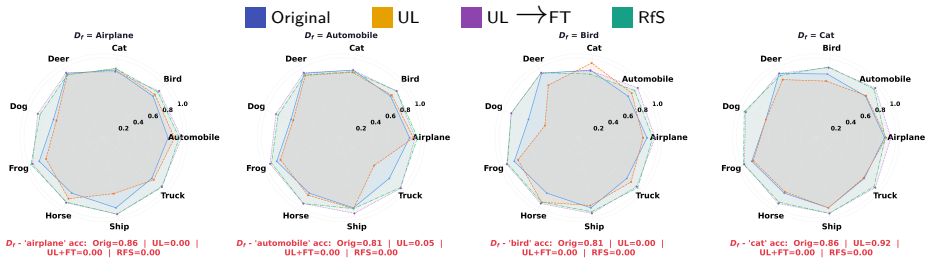
5.3 Visualizing the Collapse: The Topology Grid

The systemic failure caused by structural entanglement is illustrated by measuring $\mathcal{A}_r$ across independent ablation targets (Figure 8). In texture-dense control (Figure 8a, CIFAR-10), radar geometries remain symmetric, demonstrating successful orthogonal excision where removing a target class preserves the utility of the remaining classes. In contrast, Rows 2–4 reveal a severe utility collapse across all DIU domains. Unlearning a macro-layout (Figure 8b, RVL-CDIP) forces the radar inward, indicating the degradation of the foundational grid filters. Crucially, this identical collapse occurs at the micro-structural level for both script primitives (Figure 8c) and curvilinear strokes (Figure 8d). This uniformity shows that distance-based unlearning fails to produce localized forgetting and instead erases the geometric priors needed by the entire manifold.

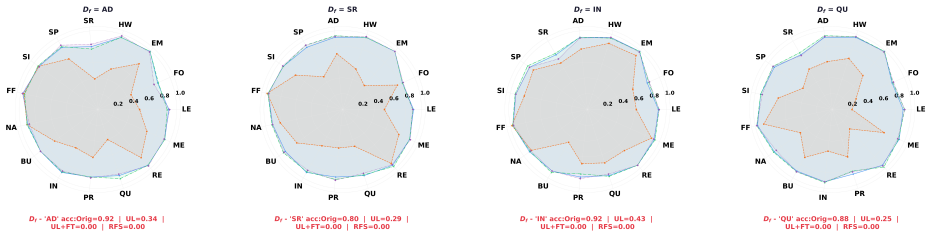
5.4 Cross-Domain Layer-wise CKA Divergence

To quantify the mathematical severity of structural degradation, we compute the linear Centered Kernel Alignment (CKA) between the unlearned model θ_u and the optimal RfS baseline θ_{rfs}. A theoretically sound unlearning pipeline must maintain high representational similarity (CKA ≈ 1.0) across all network depths. Figure 9 plots this layer-wise CKA trajectory across the VGG-16 architecture. The CIFAR-10 maintains global structural integrity (CKA > 0.80 at L40), confirming that orthogonal texture classes can be safely excised without affecting the latent manifold. In all domains, the early convolutional layers (L0–L24) maintain high similarity (CKA > 0.90). This aligns with architectural expectations: early layers extract universal, low-level gradients that are rarely penalized during class-specific unlearning.

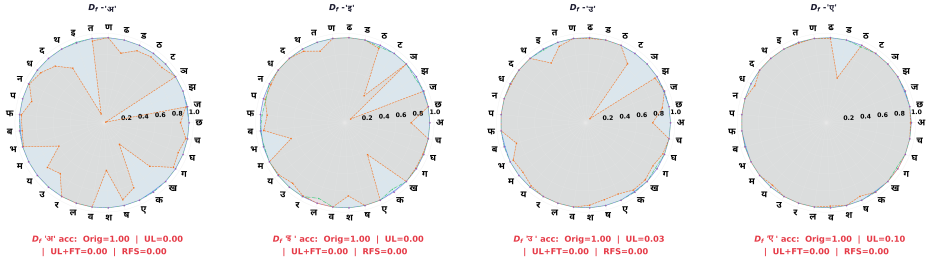
In the DIU domains, the deep feature layers show a sharp breakdown. From Layer 27, the point where the network first combines basic edges into more complex shapes, the CKA similarity drops to 0.171 for RVL-CDIP and 0.339 for Devanagari letters. This strongly supports the structural entanglement hypothesis. Because the algorithm unintentionally disrupts shared geometric features, the unlearned model no longer matches the true data distribution. Instead, it falls into a poor parameter state and loses its learned structural knowledge.



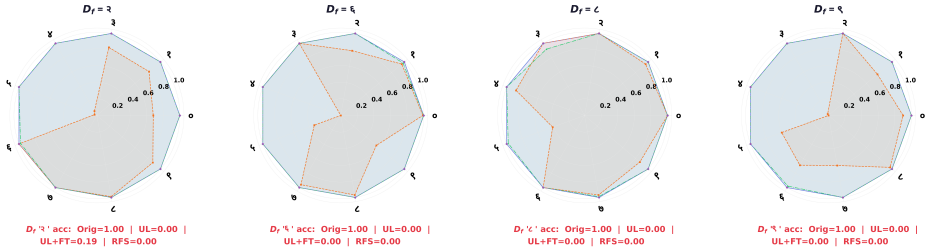
(a) CIFAR-10



(b) RVL-CDIP



(c) Devanagari-Characters



(d) Devanagari-Digits

Figure 8: Radar projections to validate the structural entanglement hypothesis.

5.5 The Privacy vs. Utility Trade-off (MIA)

Both CIFAR-10 and RVL-CDIP pruned models satisfy the MIA privacy threshold (CIFAR-10: AUC = 0.485; RVL-CDIP: AUC = 0.490), yet the underlying mechanism differs critically. In CIFAR-10, privacy follows from genuine selective erasure: forget-class accuracy collapses to 0.50% (random baseline: 10%), retain confidence remains high ($\bar{c}_{\text{retain}} = 0.828$), and deep-layer CKA diverges

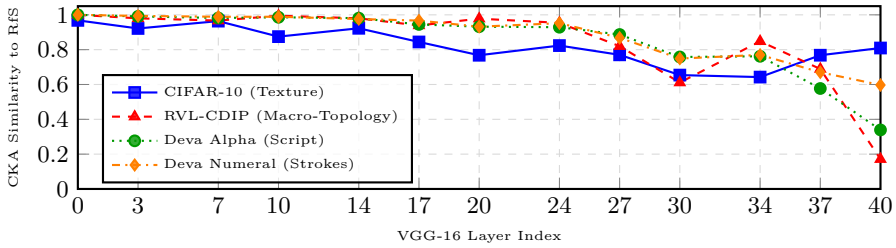


Figure 9: Layer-wise CKA similarity between unlearned and RfS model.

cleanly between forget and retain inputs ($\Delta\text{CKA}_{L40} = 0.606$), confirming that member and non-member distributions are indistinguishable because the forget class is genuinely absent. In RVL-CDIP, the identical AUC is achieved through global representational blindness: forget-class accuracy remains at 27.19%, i.e. four times the 6.2% random baseline, yet member and non-member confidence distributions collapse to near-identical values (0.358 ± 0.183 vs. 0.365 ± 0.189) because the model loses discriminative capacity uniformly across all classes, evidenced by a retain accuracy drop of 14.15pp and a CKA divergence of merely $\Delta\text{CKA}_{L40} = 0.002$ between forget and retain inputs. Structural entanglement forces this outcome: shared geometric primitives co-localize $\mathcal{D}_f$ and $\mathcal{D}_r$ parameters, such that any pruning aggressive enough to suppress membership signals simultaneously ablates the topological priors of the entire manifold, rendering distance-based MU theoretically invalid for document image understanding.

6 Conclusion and Future Work

Privacy-preserving Document Image Understanding cannot rely on vision-centric MU. We quantitatively demonstrate that the sensitivity metrics (FIM, W_1, TV) are theoretically invalid for documents. Cross-scale evaluations on RVL-CDIP and Devanagari establish that document classes rely on shared geometric primitives—a phenomenon we define as structural entanglement. Consequently, current algorithms fail to achieve selective forgetting; instead, they induce global representation blindness, as evidenced by layer-wise CKA divergence and retain-set utility collapse. To enable bounded, privacy-compliant unlearning, the field must design document-native architectures that separate topological priors from semantic features. Future work should focus on graph-theoretic models, dual-stream networks, and dynamic attention-head scrubbing in Vision Transformers. Effective document unlearning also requires dropping orthogonal texture assumptions to preserve structural integrity in entangled manifolds.

Acknowledgment

The authors acknowledge the financial support provided by the iHub Drishti Foundation, IIT Jodhpur, under project (TIH/iHubDrishti/Project/2023-24/47).

References

1. Bourtoule, L., Chandrasekaran, V., Choquette-Choo, C.A., Jia, H., Travers, A., Zhang, B., Lie, D., Papernot, N.: Machine unlearning. In: 2021 IEEE Symposium on Security and Privacy (SP). pp. 141–159 (2021)
2. Chundawat, V.S., Tarun, A.K., Mandal, M., Kankanhalli, M.: Can bad teaching induce forgetting? unlearning in deep networks using an incompetent teacher. In: AAAI Conference on Artificial Intelligence (2023)
3. Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F.A., Brendel, W.: ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In: International Conference on Learning Representations (2019)
4. Golatkar, A., Achille, A., Soatto, S.: Eternal sunshine of the spotless net: Selective forgetting in deep networks. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9304–9312 (2020)
5. Harley, A.W., Ufkes, A., Derpanis, K.G.: Evaluation of Deep Convolutional Nets for Document Image Classification and Retrieval. In: International Conference on Document Analysis and Recognition (2015)
6. Kang, L., Kumar, J., Ye, P., Li, Y., Doermann, D.: Convolutional Neural Networks for Document Image Classification. In: 2014 22nd International Conference on Pattern Recognition. pp. 3168–3172 (2014)
7. Kang, L., Fu, X., Gomez, L., Fornés, A., Valveny, E., Karatzas, D.: Preserving privacy without compromising accuracy: Machine unlearning for handwritten text recognition. Pattern Recognition p. 112411 (2025)
8. Kang, L., Souibgui, M.A., Yang, F., Gomez, L., Valveny, E., Karatzas, D.: Machine unlearning for document classification. In: International Conference on Document Analysis and Recognition. pp. 90–102 (2024)
9. Kirkpatrick, J., et al.: Overcoming catastrophic forgetting in neural networks. Proceedings of the National Academy of Sciences **114**(13), 3521–3526 (2017)
10. Kornblith, S., Norouzi, M., Lee, H., Hinton, G.: Similarity of neural network representations revisited. In: International Conference on Machine Learning. pp. 3519–3529 (2019)
11. Krizhevsky, A.: Learning Multiple Layers of Features from Tiny Images. University of Toronto (05 2012)
12. Murti, C., Bhattacharyya, C.: DisCEdit: Model Editing by Identifying Discriminative Components. In: Advances in Neural Information Processing Systems. vol. 37, pp. 47261–47296 (2024)
13. Nguyen, T.T., Huynh, T.T., Ren, Z., Nguyen, P.L., Liew, A.W.C., Yin, H., Nguyen, Q.V.H.: A Survey of Machine Unlearning. ACM Trans. Intell. Syst. Technol. **16**(5) (2025)
14. PATE, S., Ramteke, P.: Handwritten Devanagari Characters Dataset –(Vowels, Consonants and Numerals) of 44,000 images for Devanagari CAPTCHA Generation and Recognition. (2022)
15. Shokri, R., et al.: Membership inference attacks against machine learning models. In: IEEE Symposium on Security and Privacy (SP). pp. 3–18 (2017)
16. Tien, D.N., Le, D.D.: Robustness Evaluation of OCR-based Visual Document Understanding under Multi-Modal Adversarial Attacks. arXiv preprint arXiv:2506.16407 (2025)