LoTUS: Large-Scale Machine Unlearning with a Taste of Uncertainty

Supplementary Material

7. Reproducibility and Transparency

We provide the code to reproduce the results presented in this paper. Additionally, all tables and figures included in the paper are replicated in Jupyter notebooks to enhance transparency and reproducability. We conducted the experiments using Python 3.11 and CUDA 12.1. For ImageNet1k experiments, we used an NVIDIA RTX A6000 48GB GPU. The remaining experiments were performed on an NVIDIA RTX 4080 16GB GPU. We also used an Intel i7-12700K CPU and 32GB of RAM.

8. Extended Analysis on the Accuracy Metrics

In Tab. 4 we present the scores used to calculate the Avg Gap defined in Eq. (15). These results demonstrate that LoTUS succeeds superior performance in MIA accuracy, as well as in accuracies on both the retain and test sets, securing the best score in all but one measure, where it ranks second. Regarding retention performance (*i.e*., preserving the utility of the pre-trained model) LoTUS clearly outperforms state-of-the-art approaches by achieving the best scores on both the retain and test sets. Evaluating unlearning effectiveness, however, requires a more nuanced analysis. While LoTUS achieves the best scores in MIA accuracy, its accuracy on the forget set exceeds that of the gold standard model (*i.e*., the model retrained solely on the retain set). This apparent discrepancy necessitates incorporating the JSD metric, a more robust measure of unlearning effectiveness, as detailed in Sec. 5. The JSD metric captures distributional-level differences and is more sensitive than the accuracy on the forget set [9]. Therefore, the overall results indicate that LoTUS excels in unlearning effectiveness as demonstrated by its superior JSD and MIA accuracy scores, without requiring as significant a degradation in the model's utility on the forget set compared to the gold standard model. Despite this disproportionate penalty associated with higher forget accuracy, LoTUS achieves the best Avg Gap in all but one benchmark, where it ranks second.

9. Detailed Analysis on the Time Complexity

This section provides an in-depth analysis demonstrating why LoTUS achieves superior efficiency compared to state-of-the-art approaches, as observed in Tabs. 1 to 3 and discussed in Sec. 5. We define the time complexity of model updates in DNNs, generalized across architectures like ResNet18 and ViT, as:

$$O(E \cdot \frac{n_f + n_r}{B} \cdot N_p \cdot N_i) \tag{18}$$

where E represents the total number of epochs, n_f and n_r are the number of instances in D_f (forget set) and D_r (retain set) used during unlearning, respectively, B is the batch size, N_p is the total number of model parameters, and N_i is the input dimensionality. While this definition abstracts away architectural-specific details and optimizations, it provides a meaningful framework for comparing methods on shared benchmarks.

The main advantage of LoTUS over Finetuning, NegGrad+, Random Labeling and SCRUB is that it requires significantly fewer instances n_r from the retain set D_r. Specifically, LoTUS can use only 30% of the instances in D_r to preserve the utility of the model. All other factors (E, n_f, B, N_p, N_i) are the same for all the unlearning baselines in our benchmarks.

As the number of instances n_f in the forget set increases, the execution time of LoTUS increases, aligned with Eq. (18). Thus, in the extreme scenario where 50% of the forget set is designated for unlearning, we observe that the efficiency of Finetuning, NegGrad+, and Random Labeling may exceed that of LoTUS, as shown in Tab. 2. In Tab. 5 we present the scores succeded by these basic unlearning methods, that are not presented in Tab. 2, and we show that they may be better in terms of efficiency but LoTUS remains the best in terms of unlearning effectiveness.

Next, we compare the time complexity of the auxiliary computations between LoTUS and other unlearning baselines that use equal or fewer samples from the retain set D_r:

LoTUS: $O(n_f + n_v)$, where n_v is the total number of instances in the validation set, for computing τ_d.

Bad Teacher [9]: $O\big((n_f + n_r) \cdot k\big)$, where k is the total number of classes, for calculating the $\mathcal{KL}$ divergences between the student and the teacher.

UNSIR [31]: $O(E_{noise} \cdot n_f \cdot N_i)$, where E_{noise} are the epochs for noise optimization, and N_i represents the total input dimensionality (product of channels, width and height of the images).

SSD [13]: $O(n_f \cdot N_p^2)$ for computing the Fisher Information Matrix.

In this analysis, we exempt the complexity of the feedforward process which is considered the same. We observe that LoTUS is the only one with auxiliary computations of linear complexity.

10. Detailed Comparison of RF-JSD and ZRF

The computation of ZRF score [9] involves computing the Jensen-Shannon Divergence (JSD) score twice: once be-

	Metric (↓)	Gold Std	Finetuning	NegGrad+ [22]	Rnd Labeling [16]	Bad Teacher [9]	SCRUB [22]	SSD [13]	UNSIR [31]	LoTUS
ViT CIFAR-100	MIA Acc.	0.72$_{\pm0.00}$	0.77$_{\pm0.00}$(0.05)	0.79$_{\pm0.02}$(0.07)	0.74$_{\pm0.01}$(0.02)	0.66$_{\pm0.01}$(0.06)	0.75$_{\pm0.00}$(0.03)	0.75$_{\pm0.01}$(0.03)	0.78$_{\pm0.01}$(0.06)	0.73$_{\pm0.00}$**(0.01)**
	Forget Acc.	0.92$_{\pm0.00}$	0.95$_{\pm0.01}$(0.03)	0.97$_{\pm0.02}$(0.05)	0.94$_{\pm0.00}$**(0.02)**	0.90$_{\pm0.01}$**(0.02)**	0.97$_{\pm0.00}$(0.05)	0.96$_{\pm0.01}$(0.04)	0.94$_{\pm0.00}$**(0.02)**	0.96$_{\pm0.01}$(0.04)
	Retain Acc.	0.96$_{\pm0.00}$	0.98$_{\pm0.00}$(0.02)	0.97$_{\pm0.02}$(0.01)	0.98$_{\pm0.00}$(0.02)	0.91$_{\pm0.00}$(0.05)	0.96$_{\pm0.00}$()0.00	0.96$_{\pm0.00}$**(0.00)**	0.95$_{\pm0.01}$(0.01)	0.96$_{\pm0.00}$**(0.00)**
	Test Acc.	0.91$_{\pm0.01}$	0.92$_{\pm0.01}$(0.01)	0.91$_{\pm0.01}$**(0.00)**	0.92$_{\pm0.00}$(0.01)	0.89$_{\pm0.01}$(0.02)	0.91$_{\pm0.01}$**(0.00)**	0.91$_{\pm0.00}$**(0.00)**	0.90$_{\pm0.01}$(0.01)	0.91$_{\pm0.00}$**(0.00)**
	Avg Gap		0.0275	0.0325	0.0175	0.0375	0.0200	0.0175	0.0250	**0.0125**
ViT CIFAR-10	MIA Acc.	0.88$_{\pm0.00}$	0.90$_{\pm0.00}$(0.02)	0.91$_{\pm0.00}$(0.03)	0.84$_{\pm0.02}$(0.04)	0.81$_{\pm0.02}$(0.07)	0.88$_{\pm0.00}$**(0.00)**	0.89$_{\pm0.01}$(0.01)	0.90$_{\pm0.00}$(0.02)	0.87$_{\pm0.00}$(0.01)
	Forget Acc.	0.99$_{\pm0.00}$	0.99$_{\pm0.00}$**(0.00)**	1.00$_{\pm0.00}$(0.01)	0.99$_{\pm0.00}$**(0.00)**	0.96$_{\pm0.01}$(0.03)	1.00$_{\pm0.00}$(0.01)	1.00$_{\pm0.01}$(0.01)	0.99$_{\pm0.00}$**(0.00)**	1.00$_{\pm0.00}$(0.01)
	Retain Acc.	1.00$_{\pm0.00}$	1.00$_{\pm0.00}$**(0.00)**	1.00$_{\pm0.00}$**(0.00)**	1.00$_{\pm0.00}$**(0.00)**	0.97$_{\pm0.01}$(0.03)	1.00$_{\pm0.00}$**(0.00)**	1.00$_{\pm0.01}$**(0.00)**	0.99$_{\pm0.00}$(0.01)	1.00$_{\pm0.00}$**(0.00)**
	Test Acc.	0.99$_{\pm0.01}$	0.99$_{\pm0.01}$**(0.00)**	0.99$_{\pm0.01}$**(0.00)**	0.99$_{\pm0.00}$**(0.00)**	0.96$_{\pm0.01}$(0.03)	0.99$_{\pm0.01}$(0.01)	0.99$_{\pm0.01}$(0.01)	0.99$_{\pm0.00}$(0.01)	0.98$_{\pm0.01}$**(0.00)**
	Avg Gap		0.0075	0.0125	0.0125	0.0375	**0.0050**	0.0075	0.0100	**0.0050**
ViT MUFAC	MIA Acc.	0.57$_{\pm0.00}$	0.52$_{\pm0.08}$(0.05)	0.52$_{\pm0.07}$(0.05)	0.52$_{\pm0.10}$(0.05)	0.35$_{\pm0.05}$(0.22)	0.59$_{\pm0.01}$**(0.02)**	0.59$_{\pm0.01}$**(0.02)**	0.47$_{\pm0.08}$(0.10)	0.59$_{\pm0.01}$**(0.02)**
	Forget Acc.	0.57$_{\pm0.01}$	0.61$_{\pm0.01}$(0.04)	0.66$_{\pm0.02}$(0.09)	0.58$_{\pm0.01}$**(0.01)**	0.43$_{\pm0.06}$(0.14)	0.62$_{\pm0.01}$(0.05)	0.59$_{\pm0.04}$(0.02)	0.58$_{\pm0.01}$**(0.01)**	0.63$_{\pm0.00}$(0.06)
	Retain Acc.	0.66$_{\pm0.01}$	0.72$_{\pm0.01}$(0.06)	0.71$_{\pm0.01}$(0.05)	0.67$_{\pm0.02}$(0.01)	0.47$_{\pm0.07}$(0.19)	0.66$_{\pm0.01}$**(0.00)**	0.63$_{\pm0.04}$(0.03)	0.72$_{\pm0.01}$(0.06)	0.66$_{\pm0.01}$**(0.00)**
	Test Acc.	0.65$_{\pm0.01}$	0.66$_{\pm0.01}$(0.01)	0.65$_{\pm0.03}$(0.01)	0.64$_{\pm0.01}$(0.01)	0.50$_{\pm0.08}$(0.15)	0.66$_{\pm0.01}$(0.01)	0.64$_{\pm0.01}$(0.01)	0.63$_{\pm0.02}$(0.02)	0.65$_{\pm0.01}$**(0.00)**
	Avg Gap		0.0400	0.0475	**0.0200**	0.1750	**0.0200**	0.0200	0.0475	**0.0200**
ResNet18 CIFAR-100	MIA Acc.	0.49$_{\pm0.01}$	0.00$_{\pm0.00}$(0.49)	0.00$_{\pm0.00}$(0.49)	0.00$_{\pm0.00}$(0.49)	0.33$_{\pm0.58}$(0.16)	0.78$_{\pm0.05}$(0.29)	0.59$_{\pm0.05}$(0.10)	0.00$_{\pm0.00}$(0.49)	0.55$_{\pm0.01}$**(0.06)**
	Forget Acc.	0.57$_{\pm0.02}$	0.40$_{\pm0.06}$(0.17)	0.41$_{\pm0.06}$(0.16)	0.31$_{\pm0.06}$(0.26)	0.27$_{\pm0.03}$(0.30)	0.93$_{\pm0.03}$(0.36)	0.50$_{\pm0.32}$**(0.07)**	0.40$_{\pm0.07}$(0.17)	0.89$_{\pm0.04}$(0.32)
	Retain Acc.	0.94$_{\pm0.03}$	0.41$_{\pm0.06}$(0.53)	0.41$_{\pm0.06}$(0.53)	0.37$_{\pm0.07}$(0.57)	0.28$_{\pm0.03}$(0.66)	0.93$_{\pm0.03}$**(0.01)**	0.50$_{\pm0.32}$(0.44)	0.41$_{\pm0.07}$(0.53)	0.93$_{\pm0.03}$**(0.01)**
	Test Acc.	0.60$_{\pm0.02}$	0.35$_{\pm0.05}$(0.25)	0.35$_{\pm0.05}$(0.25)	0.31$_{\pm0.06}$(0.29)	0.25$_{\pm0.03}$(0.35)	0.60$_{\pm0.02}$**(0.00)**	0.36$_{\pm0.20}$(0.24)	0.34$_{\pm0.04}$(0.26)	0.62$_{\pm0.01}$(0.02)
	Avg Gap		0.3600	0.3575	0.4025	0.3675	0.1650	0.2125	0.3625	**0.1025**
ResNet18 CIFAR-10	MIA Acc.	0.76$_{\pm0.03}$	0.30$_{\pm0.26}$(0.46)	0.48$_{\pm0.50}$(0.28)	0.48$_{\pm0.50}$(0.28)	0.43$_{\pm0.37}$(0.33)	0.94$_{\pm0.01}$(0.18)	0.81$_{\pm0.11}$(0.05)	0.46$_{\pm0.03}$(0.30)	0.83$_{\pm0.01}$**(0.04)**
	Forget Acc.	0.91$_{\pm0.02}$	0.97$_{\pm0.01}$(0.06)	0.97$_{\pm0.01}$(0.06)	0.96$_{\pm0.01}$(0.05)	0.71$_{\pm0.18}$(0.20)	1.00$_{\pm0.00}$(0.09)	0.86$_{\pm0.16}$(0.05)	0.93$_{\pm0.01}$**(0.02)**	0.99$_{\pm0.01}$(0.08)
	Retain Acc.	0.99$_{\pm0.02}$	0.98$_{\pm0.01}$(0.01)	0.97$_{\pm0.01}$(0.02)	0.97$_{\pm0.01}$(0.02)	0.71$_{\pm0.18}$(0.28)	1.00$_{\pm0.00}$(0.01)	0.87$_{\pm0.16}$(0.12)	0.93$_{\pm0.01}$(0.06)	0.99$_{\pm0.00}$**(0.00)**
	Test Acc.	0.91$_{\pm0.02}$	0.89$_{\pm0.02}$(0.02)	0.88$_{\pm0.02}$(0.03)	0.89$_{\pm0.02}$(0.02)	0.66$_{\pm0.16}$(0.25)	0.93$_{\pm0.01}$(0.02)	0.80$_{\pm0.15}$(0.11)	0.86$_{\pm0.01}$(0.05)	0.91$_{\pm0.01}$**(0.00)**
	Avg Gap		0.1375	0.0975	0.0925	0.2650	0.0750	0.0825	0.1075	**0.0375**
ResNet18 MUFAC	MIA Acc.	0.48$_{\pm0.04}$	0.54$_{\pm0.09}$(0.06)	0.53$_{\pm0.08}$**(0.05)**	0.33$_{\pm0.31}$(0.15)	0.34$_{\pm0.01}$(0.14)	0.70$_{\pm0.05}$(0.22)	0.70$_{\pm0.06}$(0.22)	0.40$_{\pm0.35}$(0.08)	0.53$_{\pm0.04}$**(0.05)**
	Forget Acc.	0.47$_{\pm0.04}$	0.64$_{\pm0.04}$(0.17)	0.68$_{\pm0.04}$(0.21)	0.66$_{\pm0.04}$(0.19)	0.53$_{\pm0.07}$**(0.06)**	0.88$_{\pm0.06}$(0.41)	0.87$_{\pm0.06}$(0.40)	0.71$_{\pm0.03}$(0.24)	0.86$_{\pm0.04}$(0.39)
	Retain Acc.	0.89$_{\pm0.04}$	0.64$_{\pm0.04}$(0.25)	0.66$_{\pm0.03}$(0.23)	0.80$_{\pm0.03}$(0.09)	0.76$_{\pm0.04}$(0.13)	0.89$_{\pm0.04}$**(0.00)**	0.89$_{\pm0.05}$**(0.00)**	0.73$_{\pm0.03}$(0.16)	0.85$_{\pm0.08}$(0.04)
	Test Acc.	0.56$_{\pm0.02}$	0.43$_{\pm0.01}$(0.13)	0.43$_{\pm0.01}$(0.13)	0.47$_{\pm0.02}$(0.09)	0.48$_{\pm0.03}$(0.08)	0.54$_{\pm0.03}$**(0.02)**	0.54$_{\pm0.03}$**(0.02)**	0.46$_{\pm0.01}$(0.10)	0.54$_{\pm0.05}$**(0.02)**
	Avg Gap		0.1525	0.1550	0.1300	**0.1025**	0.1625	0.1600	0.1450	0.1250

Table 4. **Accuracy Metrics after unlearning with MUFAC and 10% of CIFAR-10/100 training sets**. Mean performance and standard deviation ($\mu \pm \sigma$) are reported across three trials with different Forget and Retain sets. Performance gaps relative to the Gold Standard (Std) are noted as (●), with smaller gaps indicating stronger performance. Avg Gap serves as a key indicator, summarizing performance across MIA, Forget, Retain, and Test Accuracy. LoTUS achieves state-of-the-art results in MIA, retain and test accuracies, ranking as the best in most cases and second-best in the remaining.

	Metric (↓)	Finetuning	NegGrad+	Rnd Labeling	LoTUS
ViT C-100	Avg. Gap	0.0400	0.0600	0.0250	**0.0225**
	JSD ×1e-4	0.02$_{\pm0.00}$	0.03$_{\pm0.01}$	**0.01$_{\pm0.01}$**	**0.01$_{\pm0.00}$**
	Time (min)	**6.34$_{\pm0.01}$**	12.68$_{\pm0.02}$	12.63$_{\pm0.02}$	13.79$_{\pm0.02}$
ViT C-10	Avg. Gap	0.0125	0.0200	**0.0050**	**0.0050**
	JSD ×1e-4	**0.00$_{\pm0.00}$**	0.01$_{\pm0.00}$	**0.00$_{\pm0.00}$**	**0.00$_{\pm0.00}$**
	Time (min)	**6.48$_{\pm0.27}$**	12.97$_{\pm0.50}$	12.60$_{\pm0.03}$	14.09$_{\pm0.53}$
ResNet C-100	Avg. Gap	0.3200	0.3150	0.3875	**0.1725**
	JSD ×1e-4	1.39$_{\pm0.10}$	1.38$_{\pm0.08}$	1.03$_{\pm0.23}$	**0.28$_{\pm0.00}$**
	Time (min)	**0.26$_{\pm0.01}$**	0.52$_{\pm0.00}$	0.48$_{\pm0.00}$	0.57$_{\pm0.01}$
ResNet C-10	Avg. Gap	0.1100	0.1475	0.2100	**0.0650**
	JSD ×1e-4	0.31$_{\pm0.00}$	0.31$_{\pm0.01}$	0.73$_{\pm0.22}$	**0.09$_{\pm0.01}$**
	Time (min)	**0.26$_{\pm0.01}$**	0.51$_{\pm0.02}$	0.48$_{\pm0.00}$	0.57$_{\pm0.00}$

Table 5. **Scaling up the Forget set to 50% of the training sets:** LoTUS outperforms basic unlearning methods in unlearning effectiveness, but not in efficiency.

tween the unlearned and a randomly initialized model, and again between the original and the same randomly initialized model. The latter serves as a reference point for the optimal value. In contrast, the RF-JSD simplifies the computation by requiring only a single JSD calculation between the unlearned model and the original model, with its optimal value fixed at zero. This direct alignment with the JSD metric (which also has an optimal value fixed at zero) provides a more straightforward interpretation of unlearning effectiveness. Additionally, the RF-JSD score improves computational efficiency using normalized class-wise mean distribution. This reduces the complexity from $O(n_f \cdot n_u \cdot k)$, where n_f and n_u represent the number of instances in the forget and the test sets respectively, to $O\big((n_f + n_u) \cdot k\big)$, where k is the number of classes. This optimization significantly decreases the computational overhead, especially in scenarios with large datasets. In this analysis we exempt the complexity of the feed-forward process which remains unchanged.

Finally, in Tab. 6, we present the detailed correlation between RF-JSD and JSD as measured by the Pearson correlation coefficient (PCC) in all benchmarks, which exhibits a strong correlation between these two metrics, with RF-JSD offering the additional advantage of not requiring the retrained (gold standard) model.

11. Cleaning the MUFAC Dataset

We identified duplicates within the forget, retain, validation, and test splits of the MUFAC dataset. More critically, we discovered instances of information leakage between these splits. To address this, we used image hashing to detect identical images with different filenames across and within

	Dataset $\left(\frac{\lvert D_f\rvert}{\lvert D_f\rvert+\lvert D_r\rvert}\right)$	PCC (↑)	p-value (↓)
ViT	CIFAR-100 (10%)	0.84	0.0043
	CIFAR-10 (10%)	0.92	0.0005
	MUFAC	0.93	0.0003
	CIFAR-100 (50%)	0.94	0.0001
	CIFAR-10 (50%)	0.99	0.0000
ResNet18	CIFAR-100 (10%)	0.97	0.0000
	CIFAR-10 (10%)	0.90	0.0011
	MUFAC	0.88	0.0018
	CIFAR-100 (50%)	0.91	0.0006
	CIFAR-10 (50%)	0.89	0.0013
	Mean ± Std	$0.92_{\pm 0.04}$	$0.0010_{\pm 0.0016}$

Table 6. **RF-JSD and JSD Correlation** measured with the Pearson correlation coefficient (PCC). A high PCC (closer to 1) indicates a strong correlation, while a low p-value reflects high confidence in the measurement. The table shows that RF-JSD strongly correlates with JSD across datasets and architectures, demonstrating its reliability as a proxy metric.

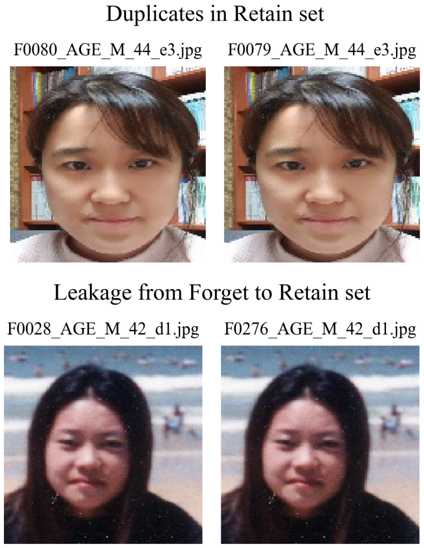

Figure 3. **Duplicates in MUFAC:** An example of a duplicate within the retain set (top) and a critical duplicate shared between the retain and forget set (bottom), which introduces information leakage.

these splits, as shown in Fig. 3. After removing these duplicates, Fig. 4 presents the updated class distribution of the images in each split. We share the code to identify these duplicates and clean MUFAC.

12. Failure Analysis

Unlearning samples from MUFAC (the clean version) presents greater challenges for all unlearning methods, as reflected by the significantly higher JSD scores. Notably, MUFAC is the only benchmark where LoTUS achieves the second-best Avg Gap rather than the best. To explore the dataset's distinct challenges, we analyzed the similarity between samples designated for retention and forgetting. As illustrated in Fig. 5, MUFAC demonstrates a higher similarity between the forget and retain sets, measured using image hash similarities. This increased overlap complicates the unlearning proces

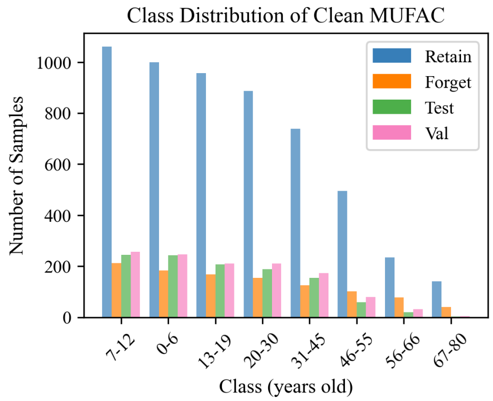

Figure 4. **Number of MUFAC Samples per Class & Split.** Unlike the balanced CIFAR-10/100 splits, MUFAC exhibits imbalanced class distributions of that varies across the retain, forget, validation, test splits.

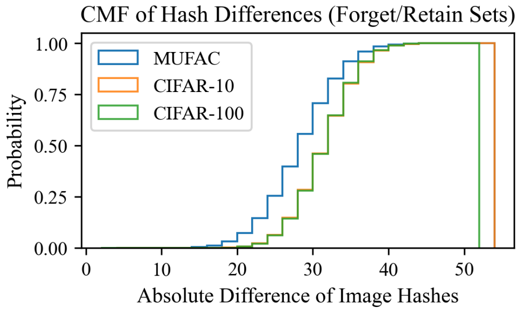

Figure 5. **Orthogonality of Forget/Retain Sets.** We measure the similarity between samples in the forget and retain sets using the absolute difference between their image hashes. MUFAC exhibits significantly higher similarity between forget and retain sets, complicating the unlearning process.

13. Social Impact

The primary evaluation of LoTUS focuses on addressing privacy-related concerns. For instance, it facilitates compliance with user requests to delete their data from a database, ensuring that the knowledge derived from these data points is also erased from the corresponding DNN models. From a security perspective, LoTUS can be applied to unlearn training samples modified by adversaries, which may otherwise compromise the model's performance. Instance-wise unlearning is more consistent with real-world conditions where privacy or security issues arise for specific data points and need to be removed, than class unlearning [5].

14. Typographical Corrections

The first equation in Eq. (9) of the main paper contains a typographical error. The corrected version is:

$$I\big(f_{\text{un}}(X_s), X_s \in D_f\big) = I\big(f_{\textbf{orig}}(X_s), X_{\mathbf{s}} \in D_u\big) \tag{19}$$

The following equations in Eq. (9) are correct as presented. Moreover, in Tab. 1, the JSD $\times 1e-4$ scores for the ResNet18-MUFAC benchmark are: (Finetune: $6.88_{\pm 0.59}$), (NegGrad+: $6.87_{\pm 0.62}$), (Rnd Labeling: $5.84_{\pm 0.98}$), (Bad Teacher: $4.30_{\pm 0.49}$), (SCRUB: $1.87_{\pm 0.08}$), (SSD: $3.04_{\pm 1.55}$), (UNSIR: $3.05_{\pm 0.32}$), (LoTUS: $\mathbf{1.29_{\pm 0.03}}$). These corrections do not affect any of the conclusions or results of the paper. These corrections will be applied in the next version of the main paper, and this section will be omitted.