ESC: Erasing Space Concept for Knowledge Deletion

Supplementary Material

A. Pesudo Code of ESC-T

Algorithm 1 Erasing Space Concept with Training (ESC-T)

Input: original feature extractor h_ψ, original classification head g_ϕ, forgetting dataset $\mathcal{D}_f$, ground truth label y, and left singular vectors $\mathbf{U}$.

Parameter: number of epochs T, learning rate η, and threshold τ.

Output: learned binary mask $\mathbf{M}_R$.

Initialize $\mathbf{M}_0 = \mathbf{1} \leftarrow$ matrix of ones with the same size as left singular vectors $\mathbf{U}$.
for t in $1...T$ **do**
  for $(x, y) \in D_f$ **do**
    $\mathbf{U}'_{t-1} = \mathbf{U} \odot \mathbf{M}_{t-1}$ ▷ Equation 5.
    $f_{t-1}(x) = g_\phi \circ h_{\psi_{t-1}}(x)$ ▷ Equation 6.
    $\mathcal{L}_{PCE}(f_{t-1}(x), y)$ ▷ Equation 4.
    $\mathbf{M}_t \leftarrow \mathbf{M}_{t-1} - \eta \nabla \mathcal{L}_{PCE}$
    $\mathbf{M}_t = \min(1, \max(0, \mathbf{M}_t))$
    ▷ Limit the mask to a range [0,1].
  end for
end for

$$\mathbf{M}_R(i,j) = \begin{cases} 1, & \text{if } \mathbf{M}_R(i,j) > \tau. \\ 0, & \text{otherwise.} \end{cases}$$

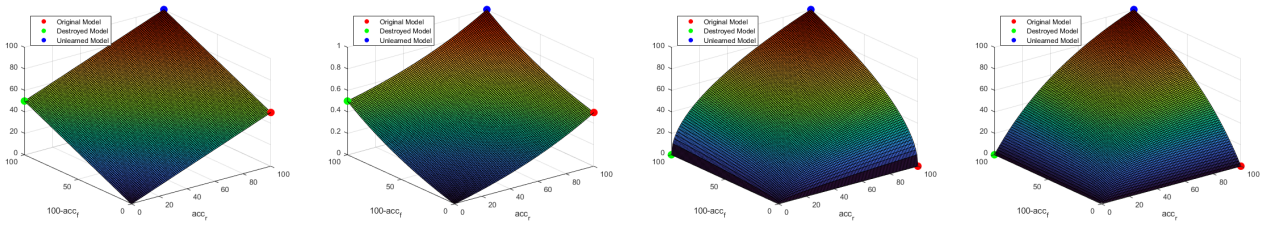
B. Additional Details

B.1. Baselines

- **Negative Gradient (NG)** [16]: From the original model, NG performs gradient ascent on the forgetting data, which is in the opposite direction of the original model's training.
- **Random Label (RL)** [8]: RL starts from the original model and fine-tunes it in a manner similar to the original training process, *i.e.*, using cross-entropy loss, but with randomly assigned labels. In our experiments, RL only used the randomly labeled forgetting data and did not utilize any remaining data. However, in the case of random data forgetting, RL also leverages the remaining data to maintain model performance.
- **Boundary Shrink (BS)** [4]: To advance RL, BS identifies the closest incorrect label for each forgetting sample based on an adversarial attack method and uses this label for unlearning, similar to the approach used in RL.
- **Layer Attack Unlearning (LAU)** [9]: LAU uses Partial-PGD to perform adversarial attacks on the features of the forgetting data, training them to be predicted differently from the original model, while employing knowledge distillation to maintain decision boundaries for the remaining data.
- **Fisher** [8]: The Fisher (Forgetting) identifies the parameters that significantly influence the forgetting data and introduces noise to neutralize their effects from the original model.
- **BadT** [5]: BadT sets the original model as a competent teacher and the randomly initialized model as an incompetent teacher. The unlearned model is obtained by minimizing the KL divergence between its output and the competent teacher's output on the remaining data, while aligning with the incompetent teacher on the forgetting data.
- **SCRUB** [11]: SCRUB is also an unlearning method based on the distillation approach, which also utilizes the original model as a teacher. The objective of SCRUB is to minimize the KL divergence with the teacher on the remaining data while maximizing the KL divergence on the forgetting data. Additionally, cross-entropy loss is used on the remaining data to maintain model performance.
- **ℓ_1-sparse** [13]: ℓ_1-sparse infuses weight sparsity into the unlearning process. The objective of ℓ_1-sparse is to minimize the cross-entropy loss on the remaining data with an ℓ_1 norm-based sparsity penalty. It directly removes knowledge from the feature extractor using a regularization term, achieving slightly better results in KD compared to existing methods. However, it necessarily requires remaining data for unlearning.
- **SalUn** [7]: SalUn consists of a two-step process: finding the weight saliency map and performing the unlearning process. SalUn leverages the gradient-based weight saliency map to identify important parameters for unlearning using the NG loss. Based on this, SalUn updates only the top-k% of parameters using any other forgetting loss, typically using the RL loss in practice.

B.2. Linear Probing

In Section 3.2.3., we introduced a novel benchmark KR as an effective metric for evaluating the degree of KD. KR performs linear probing using the feature extractor after KD, and then measures the utility, *i.e.*, accuracy. We conduct linear probing with frozen feature extractor of the unlearned model and randomly initialized classification head. As is typically done, we only optimize the classification head with all training data $D = D_f \cup D_r$. We trained the classifier for 10 epochs using the SGD optimizer with a learning rate of 0.001 and a batch size of 64. We also used linear probing in the experiments presented in Figure 1 of our



(a) Arithmetic Mean (b) Unlearning Score [9] (c) Geometric Mean (d) Harmonic Mean

Figure 1. Visualization of the distribution of various mean methods. The red dots represent the ideal original model, the green dots represent the destroyed model, and the blue dots represent the completely unlearned model.

Arithmetic Mean	Unlearning Score	Geometric Mean	Harmonic Mean
$\frac{acc_r+(100-acc_f)}{2}$	$\frac{\exp(\frac{acc_r}{100})+\exp(1-\frac{acc_f}{100})-2}{2\cdot(\exp(1)-1}$	$\sqrt{acc_r \cdot (100-acc_f)}$	$\frac{2\cdot(100-\text{acc}_f)\cdot\text{acc}_r}{(100-\text{acc}_f)+\text{acc}_r}$

Table 1. The equation of each types of mean methods.

Method	CIFAR-100					
	$D_{ft}(\downarrow)$	$D_{rt}(\uparrow)$	$AM_{\text{t}}(\uparrow)$	$US_{\text{t}}(\uparrow)$	$GM_{\text{t}}(\uparrow)$	$HM_{\text{t}}(\uparrow)$
Original	93.30	93.12	49.91	0.4676	24.98	12.50
Finetune [19]	42.20	90.89	74.35	0.6588	72.48	70.66
NG [16]	87.80	92.44	52.32	0.4802	33.58	21.56
RL [8]	13.60	23.46	54.93	0.4764	45.02	36.90
Boundary Shrink [4]	7.60	24.17	58.29	0.5217	47.26	38.32
LAU [9]	0.00	93.13	96.57	0.9475	96.50	96.44
ESC (Ours)	0.10	93.22	96.56	0.9474	96.50	96.45
ESC-T (Ours)	0.00	93.18	96.59	0.9479	96.53	96.47
Retrain	0.00	93.16	96.58	0.9477	96.52	96.46

Table 2. Comparison of various methods: Arithmetic Mean (AM), Unlearning Score (US), Geometric Mean (GM), and Harmonic Mean (HM).

main paper.

C. Additional Discussion

C.1. HM Metric

For a comprehensive evaluation, we employ the Harmonic Mean in our main paper. An appropriate comprehensive evaluation can detect that the unlearned model can effectively remove the forget knowledge while preserving the remain knowledge. To achieve this goal, we consider some failure cases of unlearning. The original model should receive a low score in comprehensive evaluation because this model is completely not unlearned, and the destroyed model, which has accuracy 0% in both remain (acc_r) and forget (acc_f) data in an extreme case, should also receive a low score. On the other hand, ideal unlearned models, such as retrain model, should have high score in comprehensive evaluation. Based on this case analysis, we seek the most appropriate mean method. To find the most suitable method, we compare the Arithmetic Mean, Unlearning Score [9], Geometric Mean, and Harmonic Mean. Each type of mean is calculated as illustrated in Table 1 and visualized in Figure 1.

In the first case, the Arithmetic Mean increases with larger acc_r and smaller acc_f (Figure 1 (a)). However, it remains fixed at 50, even when the model exhibits complete failure with 0% acc_r and 0% acc_f, or in scenarios where unlearning completely fails, resulting in acc_r and acc_f both being 100%. These characteristics mean that it is not a good metric for evaluating unlearning as we mentioned. Unlearning Score [9] also has the same problem because it applies an exponential to acc but follows the Arithmetic Mean (Figure 1 (b)). To complement this, we can think of Geometric Mean and Harmonic Mean (Figure 1 (c), (d)). Both have a value of zero in situations where the model is completely

$p(\%)$	CIFAR-100						CIFAR-100$-KR$					
	$D_f(\downarrow)$	$D_r(\uparrow)$	$D_{ft}(\downarrow)$	$D_{rt}(\uparrow)$	$HM(\uparrow)$	$HM_{\text{t}}(\uparrow)$	$D_f(\downarrow)$	$D_r(\uparrow)$	$D_{ft}(\downarrow)$	$D_{rt}(\uparrow)$	$HM(\uparrow)$	$HM_{\text{t}}(\uparrow)$
1	35.20	98.13	32.50	93.09	78.06	78.26	34.42	97.00	33.30	92.12	78.25	77.38
1.7	0.02	98.05	0.10	93.22	**99.01**	**96.45**	0.36	97.00	0.40	92.09	**98.30**	**95.70**
5	0.22	97.41	0.10	92.40	98.58	96.00	0.42	95.13	0.40	90.24	97.30	94.69
10	0.16	96.81	0.10	91.53	98.30	95.53	0.44	91.18	0.30	86.47	95.19	92.61
30	0.12	94.23	0.10	89.38	96.97	94.35	0.90	77.34	0.90	73.71	86.88	84.54
50	0.16	91.99	0.10	86.97	95.75	92.99	0.80	62.46	0.80	59.07	76.66	74.05
70	0.02	84.11	0.00	79.63	91.36	88.66	1.02	37.23	1.10	35.30	54.11	52.03
90	0.02	44.77	0.00	42.38	61.85	59.53	0.78	7.89	1.30	7.30	14.62	13.59

Table 3. Ablation study of hyperparameter p in ESC. We report accuracy and KR using ViT. The results show that our methods are robust to various values of the hyperparameter p. However, if p is too large, ESC results in over-deletion, causing both the forgetting knowledge and some of the remaining knowledge to be lost.

broken or not unlearned, making them ideal for evaluating unlearning. However, as the Harmonic Mean is more sensitive to small values compared to the Geometric Mean, it yields a lower value (12.5) than the Geometric Mean (24.98) when the model is the original model. Therefore, it is better suited for comprehensive evaluation in unlearning scenarios according to all casess. Table 2 shows that when applying the metrics to an unlearning experiment with Vision Transformer (ViT) [6] on CIFAR-100 [10], the Harmonic Mean is the best at differentiating small values. The Harmonic Mean reflects the characteristics of unlearning well, with the original model being close to 0% and the unlearned model being close to 100%.

C.2. What is Knowledge in Deep Learning Model?

Our methods are designed to remove forgetting knowledge from the original model, even at the feature level. In this section, we further discuss the nature of knowledge in deep learning models. Deep learning models accumulate learned knowledge as data passes through the network, with the embedding features containing the most accumulated knowledge. This accumulated knowledge consists of sub-features that represent lower-level knowledge compared to the target knowledge, *i.e.*, class.

For example, the identity of person A might be defined by a combination of an oval face shape, straight eyebrows, blue eyes, and a Roman nose. Similarly, the identity of person B could be characterized by a long face shape, arched eyebrows, black eyes, and a Nubian nose. While these sub-features may be present across different classes and can be captured from the remaining knowledge, it is the specific combination of these features that allows the model to identify particular knowledge. As mentioned in Section 2.3.1., even if the model operates based on the the remaining knowledge, this is not sufficient for KD. Therefore, it is crucial to remove the model's ability to effectively combine these features for identification. This is why our methods also outperform the retrained model in KR.

C.3. Setting of Random Data Forgetting

We present the random data forgetting experiments in Table 5 of our main paper. The goal of random data forgetting is to remove the influence of specific data from the original model, and the optimal point for achieving this is the retrained model. This is distinct from our primary focus in KD, which involves knowledge removal requests from users. For this reason, unlike KD, we evaluate the effectiveness of random data forgetting by comparing it to the retrain-from-scratch model. Furthermore, we used the average gap as an integrated value for each metric, similar to [7, 13].

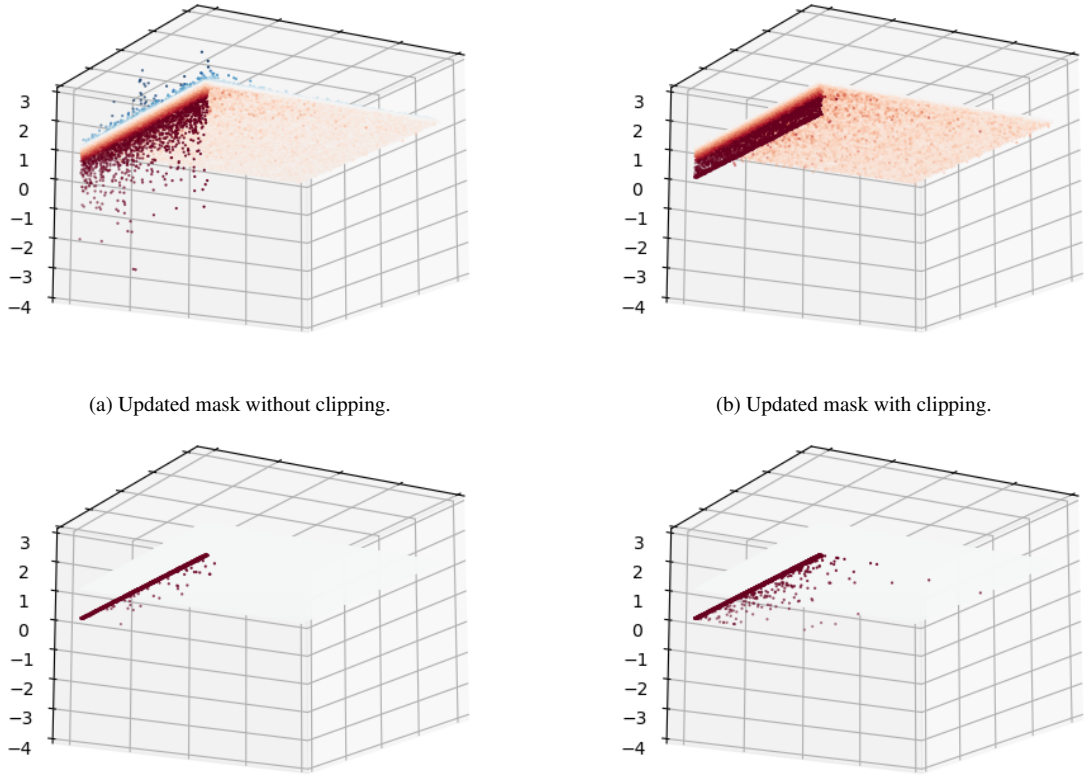
D. Additional Ablation

D.1. p Ablation

Table 3 shows the full results of the hyperparameter p ablation study. This shows that as p gets larger, more knowledge is erased. When we erased very little, like 1%, model retained a lot of forget knowledge, and when we erased a lot, like 90%, model lost a lot of remain knowledge. ESC achieves relatively robust results across various values of p. Despite the performance is good for most p, we empirically found that ESC achieves the most stable results around 1.7%. For this reason, we selected the pruning hyperparameter p between 1 and 3 in our experiments. Some variation is needed because each model has a different embedding dimension size, and each principal direction contains varying levels of knowledge intensity. However, we consistently obtain comparable results for any value of p within this range.

D.2. The Effect of Clipping

In our additional experiments from Table 3 in our main paper, we investigated the impact of restricting masks to values between 0 and 1 during the update process in ESC-T. As we mentioned in Section 4.3. in our main paper, when we



(a) Updated mask without clipping.

(b) Updated mask with clipping.

(c) After threshold without clipping.

(d) After threshold with clipping.

Figure 2. Visualization of clipping effect on the unlearned mask after training and the threshold. The results show that the unlearned mask utilizes more elements when the clipping method is employed. This also means that more elements can be used, after threshold, for unlearning, allowing ESC-T to efficiently remove more forget knowledge while preserving the remaining knowledge.

introduced clipping to confine mask values within the range of 0 to 1, we observed an improvement in KR performance. To understand this, we visualized the unlearned mask $\mathbf{M}_T$ in Figure 2. As shown in Figure 2 (a), without clipping, some elements are updated to have out of range values, leading to overconfidence about forgetting data in the classification head. This overconfidence triggers an early stopping of our training process, as it meets the criterion of our strategy where all forgetting data are misclassified. Consequently, as shown in Figure 2 (c), this approach fails to fully evaluate the influence of each element, resulting in the missing of crucial elements pertinent to the KD of forgetting data. On the other hand, with clipping, our training process relatively considers the full range of mask as shown in Figure 2 (b) and (d). Clipping makes it possible for ESC-T to capture more important elements, enabling us to effectively eliminate forgetting knowledge. Furthermore, these results demonstrate the validity of our design choice for ESC, as only a few directional elements play crucial roles for unlearning process.

We presented the mask after threshold in Figure 2 (c) and (d). Threshold ensures that crucial forgetting knowledge is completely erased (0), while less important knowledge is fully retained (1), allowing for the preservation of remaining knowledge. This clear distinction based on importance significantly enhances the performance of KR, leading to improved outcomes. By applying clipping and threshold together, we can get the optimal KR performance while maintaining good utility.

E. Additional Experiments

E.1. Additional Results for KD

Additional in ViT. We also conducted the KD with ViT on the CIFAR-100 dataset, and Table 4 provides the results. In this experiments, we utilized the ViT and removed 10 classes from the trained model. Except the LAU, they suffer from the over-forgetting or under-forgetting in the utility, and feature-level knowledge deletion is also restricted. LAU has comparable results in utility, but it also fail to re-

Method	CIFAR-100						CIFAR-100$-KR$					
	$D_f(\downarrow)$	$D_r(\uparrow)$	$D_{ft}(\downarrow)$	$D_{rt}(\uparrow)$	$HM(\uparrow)$	$HM_t(\uparrow)$	$D_f(\downarrow)$	$D_r(\uparrow)$	$D_{ft}(\downarrow)$	$D_{rt}(\uparrow)$	$HM(\uparrow)$	$HM_t(\uparrow)$
Original	98.42	98.10	93.30	93.12	3.11	12.50	97.78	97.01	91.70	91.83	4.34	15.22
Finetune [19]	43.60	98.31	42.20	90.89	71.68	70.66	76.14	96.62	70.70	89.11	38.27	44.10
NG [16]	46.92	66.87	42.60	64.30	59.18	60.65	95.14	96.32	88.20	91.41	9.25	20.90
RL [8]	14.58	24.73	13.60	23.46	38.36	36.90	11.76	17.04	10.80	16.47	28.56	27.81
Boundary Shrink [4]	8.26	25.06	7.60	24.17	39.37	38.32	19.30	30.21	17.10	29.96	43.96	44.01
LAU [9]	0.00	97.99	0.00	93.13	98.98	96.44	97.78	97.01	91.70	91.83	4.34	15.22
ESC (Ours)	0.02	98.05	0.10	93.22	<u>99.01</u>	<u>96.45</u>	0.36	97.00	0.40	92.09	**98.30**	**95.70**
ESC-T (Ours)	0.00	98.11	0.00	93.18	**99.05**	**96.47**	1.80	96.93	1.60	92.04	<u>97.56</u>	<u>95.11</u>
Retrain	0.00	98.48	0.00	93.16	99.23	96.46	71.34	97.17	71.20	91.90	44.26	43.86

Table 4. Accuracy and KR using ViT on CIFAR-100. Bold at the best value and underlined at the second.

Method	Lacuna-10						Lacuna-10$-KR$					
	$D_f(\downarrow)$	$D_r(\uparrow)$	$D_{ft}(\downarrow)$	$D_{rt}(\uparrow)$	$HM(\uparrow)$	$HM_t(\uparrow)$	$D_f(\downarrow)$	$D_r(\uparrow)$	$D_{ft}(\downarrow)$	$D_{rt}(\uparrow)$	$HM(\uparrow)$	$HM_t(\uparrow)$
Original	100.0	100.0	92.00	86.56	0.00	14.65	100.0	100.0	89.00	85.89	0.00	19.50
LAU [9]	0.00	100.0	0.00	86.22	100.0	92.60	100.0	100.0	89.00	85.89	0.00	19.50
ESC (Ours)	1.50	100.0	0.00	88.00	99.24	93.62	12.25	100.0	10.00	87.22	**93.48**	88.59
ESC-T (Ours)	0.00	100.0	0.00	88.11	**100.0**	**93.68**	12.25	100.0	9.00	87.22	**93.48**	**89.07**
Retrain	0.00	100.0	0.00	89.33	100.0	94.36	81.00	100.0	64.00	84.78	91.93	50.54

Table 5. Accuracy and KR using All-CNN on Lacuna-10.

Method	Lacuna-100						Lacuna-100$-KR$					
	$D_f(\downarrow)$	$D_r(\uparrow)$	$D_{ft}(\downarrow)$	$D_{rt}(\uparrow)$	$HM(\uparrow)$	$HM_t(\uparrow)$	$D_f(\downarrow)$	$D_r(\uparrow)$	$D_{ft}(\downarrow)$	$D_{rt}(\uparrow)$	$HM(\uparrow)$	$HM_t(\uparrow)$
Original	94.58	94.88	90.60	90.30	10.25	17.03	95.48	94.98	93.10	90.61	8.63	12.82
LAU [9]	0.00	88.55	0.00	83.97	93.93	91.29	95.48	94.98	93.10	90.61	8.63	12.82
ESC (Ours)	0.40	86.54	0.10	81.09	92.61	89.52	0.10	89.65	0.40	85.10	94.50	91.78
ESC-T (Ours)	0.00	93.38	0.00	89.87	96.58	94.66	58.68	94.69	54.90	90.50	57.53	60.20
Retrain	40.53	96.24	38.60	91.30	73.51	73.42	92.05	95.42	89.20	91.27	13.91	19.31

Table 6. Accuracy and KR using ViT on Lacuna-100.

move the feature knowledge. In contrast, our methods were still outperform to other comparisons both utility and KR.

Experiments of Facial Domains. To further validate and strengthen the concept of KD, we conducted experiments with facial domain datasets. Given the risks associated with extensive personal data in face recognition systems, our experiments with these datasets are particularly relevant. Lacuna-10 [8, 11] consists of face images of 10 celebrities from VGGFaces2 [2], randomly sampled with at least 500 images each. It was split into a test set of 100 samples per class, while the remaining samples formed the training set. In Figure 6 of our main paper, we visualize the effectiveness of our methods using Grad-CAM on this dataset. In Table 5, we present the quantitative results of these experiments. Our methods continue to demonstrate strong effectiveness for KD in the facial domain, successfully mitigating privacy concerns while maintaining model performance. In addition, we also conducted experiments with a larger facial domain dataset, Lacuna-100, which contains randomly sampled images of 100 celebrities with at least 500 images each. We selected the ViT model as it is more appropriate for the larger dataset, and the results are shown in Table 6. These results again highlight the effectiveness of our proposed methods in protecting individual privacy by removing specific identity knowledge from the trained models.

Large Scale Dataset. Our methods are simple and significantly efficient for KD. For this reason, they can be easily applied to larger-scale datasets, such as ImageNet-1K [14]. In Table 7, our methods achieve superior performance compared to LAU in both utility and KR. These results indicate that our methods are robust to dataset size, making them highly useful in real-world applications, where most deep learning models are built on large-scale datasets. In these experiments, we only use LAU for comparison because it is the most efficient method among the existing ones.

Fine-Grained Dataset. In real-world scenario, deep learning model need to consider fine-grained datasets, dis-

Method	ImageNet-1K						ImageNet-1K$-KR$					
	$D_f(\downarrow)$	$D_r(\uparrow)$	$D_{ft}(\downarrow)$	$D_{rt}(\uparrow)$	$HM(\uparrow)$	$HM_t(\uparrow)$	$D_f(\downarrow)$	$D_r(\uparrow)$	$D_{ft}(\downarrow)$	$D_{rt}(\uparrow)$	$HM(\uparrow)$	$HM_t(\uparrow)$
Original	87.80	87.40	80.90	80.25	21.41	30.86	87.18	86.92	80.96	79.78	22.34	30.74
LAU [9]	0.00	67.01	0.00	62.09	80.25	76.61	87.18	86.92	80.96	79.78	22.34	30.74
ESC (Ours)	0.11	83.81	0.06	79.15	<u>91.15</u>	<u>88.34</u>	0.04	84.72	0.00	80.15	**91.71**	**88.98**
ESC-T (Ours)	0.01	84.81	0.00	80.14	**91.78**	**88.98**	56.97	86.35	50.40	80.27	<u>57.44</u>	<u>61.31</u>

Table 7. Accuracy and KR using ViT on ImageNet-1k. Bold at the best and underlined at the second.

Method	CUB-200-2011						CUB-200-2011$-KR$					
	$D_f(\downarrow)$	$D_r(\uparrow)$	$D_{ft}(\downarrow)$	$D_{rt}(\uparrow)$	$HM(\uparrow)$	$HM_t(\uparrow)$	$D_f(\downarrow)$	$D_r(\uparrow)$	$D_{ft}(\downarrow)$	$D_{rt}(\uparrow)$	$HM(\uparrow)$	$HM_t(\uparrow)$
Original	94.82	95.66	83.81	85.03	9.82	27.20	86.81	92.60	78.42	82.04	23.09	34.17
LAU [9]	0.00	92.90	0.18	82.51	96.62	90.34	86.81	92.60	78.42	82.04	23.09	34.17
ESC (Ours)	0.00	89.51	0.00	79.19	94.46	88.39	0.83	76.18	0.18	68.52	86.17	81.26
ESC-T (Ours)	0.00	92.55	0.00	82.72	96.13	90.54	1.50	85.73	1.08	77.05	91.97	86.63
Retrain	48.25	98.74	45.86	86.18	67.91	66.50	76.79	93.46	69.24	80.53	37.19	44.42

Table 8. Accuracy and KR using ViT on CUB-200-2011.

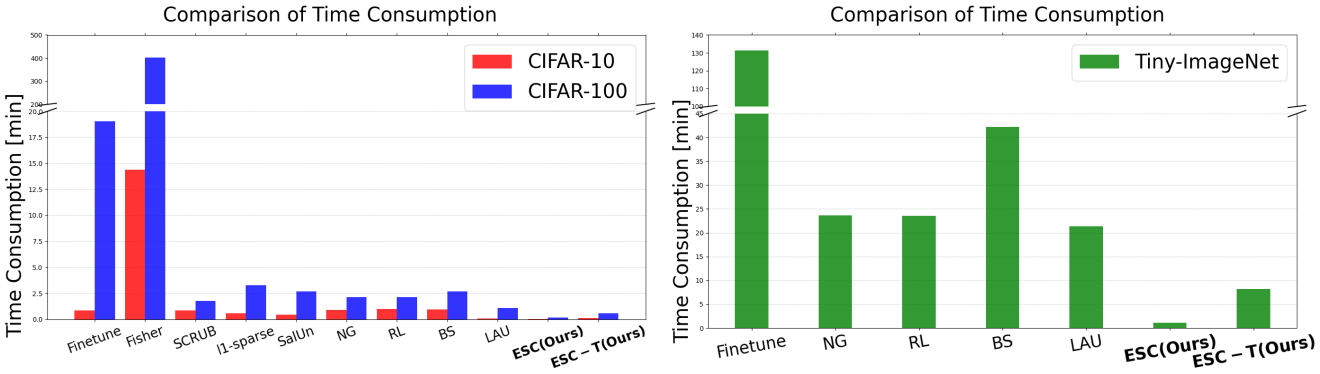


Figure 3. Comparison of time consumption. In addition to Figure 5 of our main paper, we present further results from other comparisons using the remaining data on the left. On the right side, we also illustrate the results on the Tiny-ImageNet dataset.

tinguishing subtle differences between objects is often necessary. Such datasets enable the model to learn complex objects with high precision, ensuring reliable performance that aligns with real-world needs. Furthermore, KD must effectively address these demands by appropriately handling subtle differences between objects. This is challenging because the objects share much of the general knowledge, differing only in fine details. To verify our methods in this setting, we conducted KD using the CUB-200-2011 dataset [18], which includes 200 different bird species, each annotated in detail. Approximately 60 images represent each species, creating a diverse and rich dataset. Despite these challenges, the results demonstrated the generalizability of our methods in a fine-grained setting.

E.2. Additional Time Consumption

To extend Figure 5 from our main paper, we present the time consumption for all comparisons on the CIFAR-10 and CIFAR-100 datasets in Figure 3. When using the remaining data, significantly more time is typically required for KD since the remaining data is much larger than the forgetting data (almost nine times larger). In contrast, ESC requires only a single forward pass, achieving the lowest time consumption. ESC-T also involves lightweight training, updating only the mask, making it faster than other methods. This efficiency becomes even more remarkable as the dataset and model size grow. As shown in Figure 3 left, where both the model (ViT) and dataset (Tiny-ImageNet) are larger, the efficiency of our methods is significantly higher than other methods, with a difference of more than 10 minutes.

E.3. Analysis by the Number of Target Classes

In this section, we explore the impact of the number of forgetting classes on KD performance. We conducted this experiment on the CIFAR-100 with ViT, varying the number of forgetting classes from 1 to 60. In Figure 4, we visualized the variation of accuracy with other baselines that exhibit poor multi-class unlearning performance, as discussed

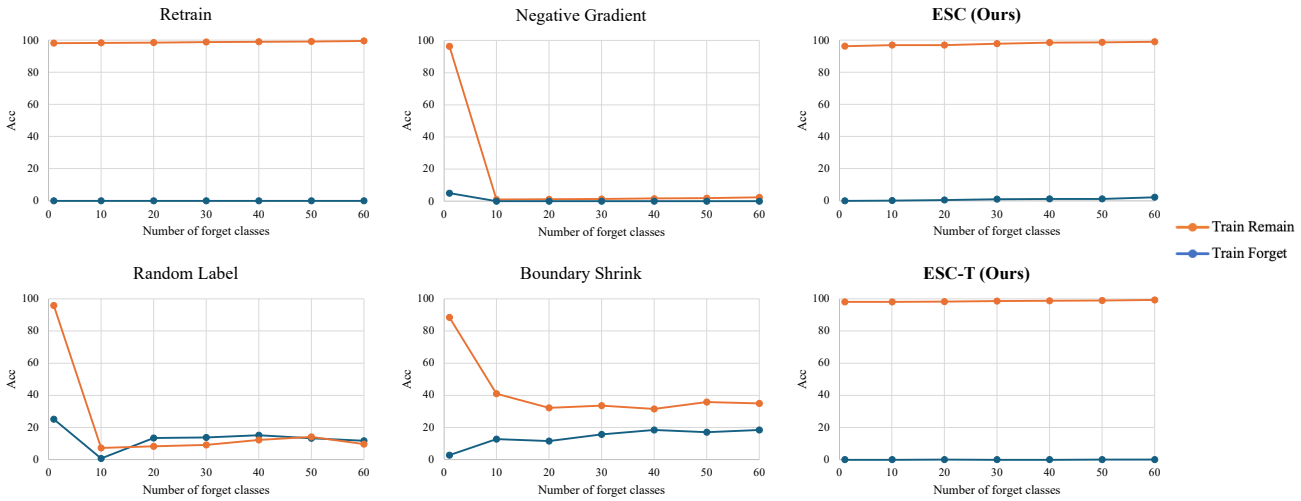


Figure 4. Multi-class unlearning experiments on diverse baselines. We use the results of train accuracy on CIFAR-100 with ViT.

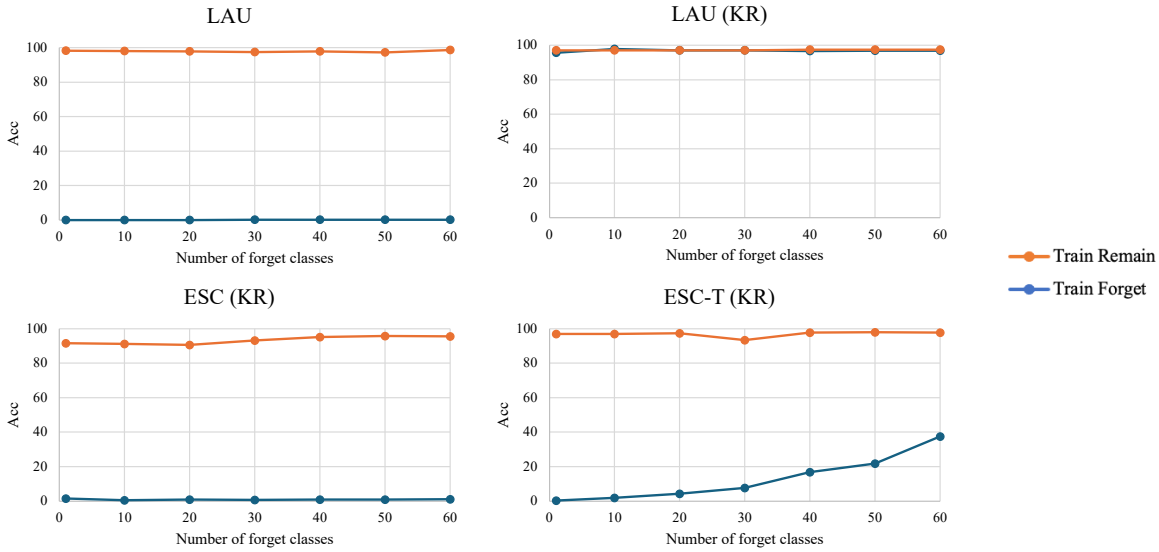


Figure 5. Multi-class unlearning experiments on LAU [9] and ours. While LAU performs well in utility, it completely recover the forgetting knowledge in KR. In contrast, ESC and ESC-T still preserve the performance in KR.

in our main paper. For all baselines, we searched for the best hyperparameter for single-class unlearning and adopted this for other multi-class unlearning scenarios. Our methods use the same hyperparameter settings as detailed in our main paper.

The results of NG and RL indicate significant degradation of the model. In the case of NG, as unlearning progresses, the unlearned model tends to converge to a scratch model due to gradient ascent. NG requires more unlearning steps in a multi-class unlearning scenario because an increase in the number of classes leads to an increase in the amount of training data. Furthermore, when using only forgetting data, NG has no way to preserve the remaining knowledge, causing additional unlearning steps to further degrade the model. Similarly, RL induces confusion in the model about the forgetting data, potentially leading to unintended outcomes. This confusion, stemming from the model's attempt to adapt to the randomly assigned new labels, may inadvertently reinforce incorrect associations between features and these labels. In multi-class settings, randomly reassigning labels from multiple classes hinders effective unlearning by scattering the model's focus, resulting in the deterioration of the unlearned model. The impact of BS, is mitigated when the unlearned class is adjacent in fea-

Method	CIFAR-10						CIFAR-10$-KR$					
	$D_f(\downarrow)$	$D_r(\uparrow)$	$D_{ft}(\downarrow)$	$D_{rt}(\uparrow)$	$HM(\uparrow)$	$HM_t(\uparrow)$	$D_f(\downarrow)$	$D_r(\uparrow)$	$D_{ft}(\downarrow)$	$D_{rt}(\uparrow)$	$HM(\uparrow)$	$HM_t(\uparrow)$
Original	99.96	99.91	99.20	99.02	0.08	1.59	99.98	99.89	99.40	98.92	0.04	1.19
ESC(Ours)	1.92	99.56	2.30	98.36	98.81	98.03	73.48	99.72	70.40	98.63	41.90	45.53
ESC-T(Ours)	0.86	99.92	0.50	99.06	99.53	99.28	82.84	99.90	81.70	98.96	29.29	30.89
Method	CIFAR-100						CIFAR-100$-KR$					
	$D_f(\downarrow)$	$D_r(\uparrow)$	$D_{ft}(\downarrow)$	$D_{rt}(\uparrow)$	$HM(\uparrow)$	$HM_t(\uparrow)$	$D_f(\downarrow)$	$D_r(\uparrow)$	$D_{ft}(\downarrow)$	$D_{rt}(\uparrow)$	$HM(\uparrow)$	$HM_t(\uparrow)$
Original	98.42	98.10	93.30	93.12	3.11	12.50	97.78	97.01	91.70	91.83	4.34	15.22
ESC(Ours)	0.20	95.61	0.00	90.67	97.66	95.11	31.70	97.93	28.30	86.74	80.47	78.51
ESC-T(Ours)	0.98	98.01	0.50	93.09	98.51	96.19	40.64	96.54	36.70	92.01	73.52	75.00
Method	Tiny-ImageNet						Tiny-ImageNet$-KR$					
	$D_f(\downarrow)$	$D_r(\uparrow)$	$D_{ft}(\downarrow)$	$D_{rt}(\uparrow)$	$HM(\uparrow)$	$HM_t(\uparrow)$	$D_f(\downarrow)$	$D_r(\uparrow)$	$D_{ft}(\downarrow)$	$D_{rt}(\uparrow)$	$HM(\uparrow)$	$HM_t(\uparrow)$
Original	96.72	96.47	90.30	90.24	6.34	17.52	95.55	94.80	90.50	88.57	8.50	17.16
ESC(Ours)	0.12	94.94	0.00	89.36	97.35	94.38	1.01	94.76	0.40	89.12	96.83	94.07
ESC-T(Ours)	0.14	96.40	0.00	90.51	98.10	95.02	7.36	94.58	6.30	88.78	93.60	91.17

Table 9. Accuracy and KR performance in ViT, using unseen data (test data) for KD. Our methods shows comparable results only with unseen data on various datasets.

Method	# of sequence	CIFAR-100						CIFAR-100$-KR$					
		$D_f(\downarrow)$	$D_r(\uparrow)$	$D_{ft}(\downarrow)$	$D_{rt}(\uparrow)$	$HM(\uparrow)$	$HM_t(\uparrow)$	$D_f(\downarrow)$	$D_r(\uparrow)$	$D_{ft}(\downarrow)$	$D_{rt}(\uparrow)$	$HM(\uparrow)$	$HM_t(\uparrow)$
ESC (Ours)	1	0.02	98.05	0.10	93.22	99.01	96.45	0.36	97.00	0.40	92.09	98.30	95.70
	2	0.04	98.04	0.20	93.04	98.99	96.30	0.98	96.73	1.80	91.87	97.86	94.93
	5	0.16	97.34	0.50	92.30	98.57	95.76	0.30	95.06	0.20	90.26	97.32	94.79
	10	1.58	97.87	1.30	93.00	98.14	95.77	1.46	96.43	1.10	91.60	97.47	95.11
ESC-T (Ours)	1	0.00	98.11	0.00	93.18	99.05	96.47	1.80	96.93	1.60	92.04	97.56	95.11
	2	0.00	98.09	0.00	93.19	99.04	96.47	2.10	96.71	3.50	91.93	97.30	94.16
	5	0.00	98.09	0.00	93.14	99.04	96.45	1.54	96.78	2.50	92.11	97.61	94.73
	10	0.04	98.10	0.00	93.2	99.02	96.48	1.82	96.70	2.70	92.03	97.43	94.59

Table 10. Experiments of incremental unlearning using ViT on CIFAR-100.

ture space, resulting in limited performance in multi-class unlearning. However, both ESC and ESC-T consistently perform well regardless of the number of forgetting classes. This is because our methods directly eliminate the knowledge of forgetting data by using modified feature space, thereby minimizing the impact of unlearning steps or the relationships between each forgetting class.

As illustrated in Figure 5, similar to the results presented in Table 1 of our main paper, LAU performs well in multi-class unlearning scenarios. However, similar to the original model, it suffers from the complete recovery of forgetting knowledge, as shown in KR. This means that LAU still has enormous forgetting knowledge and it cannot fulfill the KD. Conversely, both ESC and ESC-T effectively remove forgetting knowledge, enabling them to perform well in KR. Although ESC-T slightly improve the forgetting accuracy as growing the number of forgetting classes, they are still reasonable with respect to the forgetting accuracy in KR of retrain model (71.34% in ViT, 41.28% in All-CNN).

E.4. Unlearning with Unseen Data

To demonstrate the expandability of our methods, we performed unlearning with unseen forgetting data (test data), and the results presented in Table 9. Our methods continue to exhibit reasonable performance when utilizing the unlearned model process with forgetting data (seen data) for unlearning. In terms of accuracy, the gap between using seen data and unseen data is less than 1%. Similarly, in KR, this gap remains below 4%, except for CIFAR-100. The forget accuracy slightly improved in KR from CIFAR-100, it is also a reasonable result. These results indicate that our methods can be applied in more practical scenarios where the original training data is inaccessible.

E.5. Incremental Unlearning with ESC and ESC-T

In Table 10, we applied our method repeatedly, i.e. incremental unlearning scenario. We conducted this using ViT on CIFAR-100 and unlearned 10 classes. For the sequential setting, we divided the unlearning task as 1, 2, 5, and 10. For example, in case of 5, we divided 10 classes into 5 sub-

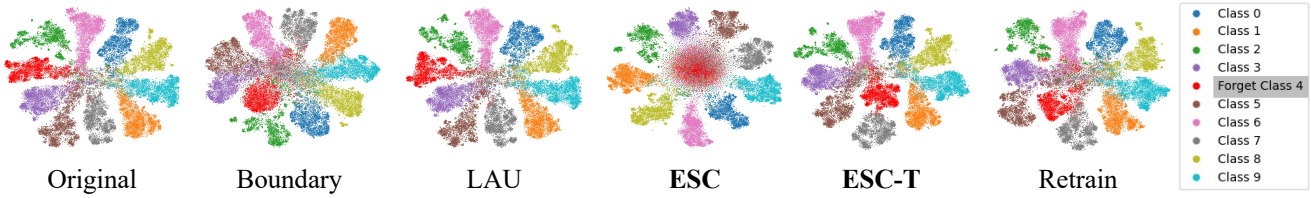


Figure 6. t-SNE [17] visualization of each unlearning method on the CIFAR-10 dataset. We removed the knowledge about *'deer'*, which corresponds to class 4, and is represented by a red dot. Each point represents a sample colored with the ground truth class.

Method	Forgetting Ratio	10%	20%	30%	40%	50%
ESC	Avg. Gap with Retrain (↓)	1.45	1.44	1.61	1.88	1.70
ESC-T		2.77	2.68	3.04	2.27	3.22

Table 11. Experiments on the impact of various forgetting ratios.

Method	All-CNN	ViT	
	CIFAR-10	CIFAR-10	CIFAR-100
Original	0.6717	0.8265	0.9802
Finetune [19]	0.8212	0.8939	0.9847
Boundary Shrink [4]	0.8719	0.9679	0.9983
LAU [9]	0.7773	0.9513	0.9869
ESC(Ours)	**0.8949**	**0.9952**	**0.9989**
ESC-T(Ours)	0.8699	0.9882	**0.9989**
Retrain	0.7946	0.9378	0.9897

Table 12. Zero Retrain Forgetting performance using All-CNN and ViT on CIFAR-10 and CIFAR-100. The larger the value, the closer the unlearned model is to the scratch.

sets and conducted disjoint unlearning for each subset. Our methods also worked well in incremental setting. Because each U_P can be merged into $U_{P_{total}} = \cdots U_{P_2} U_{P_1}$, no additional parameters are needed for incremental unlearning.

E.6. Additional Random Data Forgetting.

To demonstrate the robustness of our method to different forgetting ratios, we conducted experiments under various settings. As shown in Table 11, our method remains robust across a wide range of forgetting ratios.

E.7. Zero Retrain Forgetting

In addition to the MIA, we further assessed the privacy guarantees of our methods by employing the Zero Retrain Forgetting (ZRF) [5]. The ZRF calculates the Jensen–Shannon (JS) divergence [12] between the unlearned model and a random initialized model. In Table 12, the ZRF score of our methods is the highest compared to others, reaching almost 0.99 in the ViT experiments. These results indicate that our methods successfully remove forgetting knowledge in privacy-focused perspective. This effectiveness is attributed to our approach's ability to erase knowledge at the feature level.

Figure 7. Qualitative results.

Method	ID (↓)	FID_{pre} (↓)
GUIDE	0.059	7.921
ESC	**-0.029**	**5.690**

Table 13. Quantitative results.

E.8. Generative Unlearning

We also apply ESC to Generative Identity Unlearning (GUIDE) [15]. To obtain latent representations, GUIDE uses GOAE[20] for inversion and EG3D [3] as the generative model. Based on those, GUIDE computes the unlearned latent w_u as below:

$$w_u = w_0 + w_i d, \tag{1}$$

where w_{id} is obtained using GOAE. To remove individual identity while preserving overall generation quality, we apply ESC only to the identity latent feature (w_{id}). As shown in Figure 7 and Table 13, the results demonstrate the applicability to generative tasks.

F. Additional Visualization

In this section, we illustrated additional visualization results. In Figure 8, we present the Grad-CAM results for each class in the CIFAR-10 dataset using the All-CNN, complementing the results to Section 4.4. in our main paper. The findings emphasize once again that both ESC and

ESC-T successfully eliminate attention to class object compared with the original model. Our methods consistently focus strongly on the background regardless of the dataset, indicating that unlearning has been successfully achieved at the feature level.

Furthermore, we utilized t-SNE [17] for visualizing the manifold space of features, in Figure 6. Upon applying ESC, the clusters for the forgetting class not only centralized among the remaining class clusters but also manifested in a dispersed Gaussian form. In fact, the activations of the forgetting class are close to zero, leading to these results. This deactivation also removes the linear separation in the feature space, ensuring the effective deletion of the model's ability to capture distinct features of forgetting knowledge. When comparing the distributions of Retrain and ESC-T, they appeared quite similar, but ESC-T performed slightly better than Retrain on KR. This suggests that the pre-trained information was strongly entangled at the feature level, preventing KD from being effectively applied through retraining. This suggests that the pre-trained information is strongly entangled at the feature level, and it means that retrained model is not sufficient for KD.

G. Broad Impacts and Limitations

We reveal a new insight that existing unlearning methods cannot fully fulfill the knowledge removal request by users. Furthermore, our findings suggest a potential problem: existing methods cannot eliminate almost forgetting knowledge, and we finally suggest a novel perspective for knowledge removal considering user's requests and feature-level, called Knowledge Deletion. We also propose a novel evaluation setting for this issue. Our insights provide a benefit to the related research community. Furthermore, we introduce simple yet effective methods for KD. Our methods demonstrate remarkable performance in various KD scenarios, including incremental and facial domains, even in the random data forgetting scenario. In addition to their efficiency, our methods are suitable for real-world AI deployments.

ESC and ESC-T utilize an additional layer after the penultimate layer, i.e., after the feature extractor. Our methods are particularly useful when the service provider releases their model as a black-box, such as Chat-GPT [1]. However, if the model is released as a white-box, we need to integrate this part into the original model architecture. If we have a simple MLP layer, it can be directly merged with the existing weights, but merging with the entire model remains an open question. We expect that this issue could be addressed through methods such as distillation, and we plan to conduct follow-up research on this challenge.

Furthermore, our methods has generalizability and robustness in discriminative tasks, we need to extend this to other domains, such as diffusion model and language model. Our methods directly edit the feature space, while the latent space of generative models remains highly sensitive. Consequently, further investigation is needed to determine the applicability of our methods to generative models. We encourage future research to address these gaps.

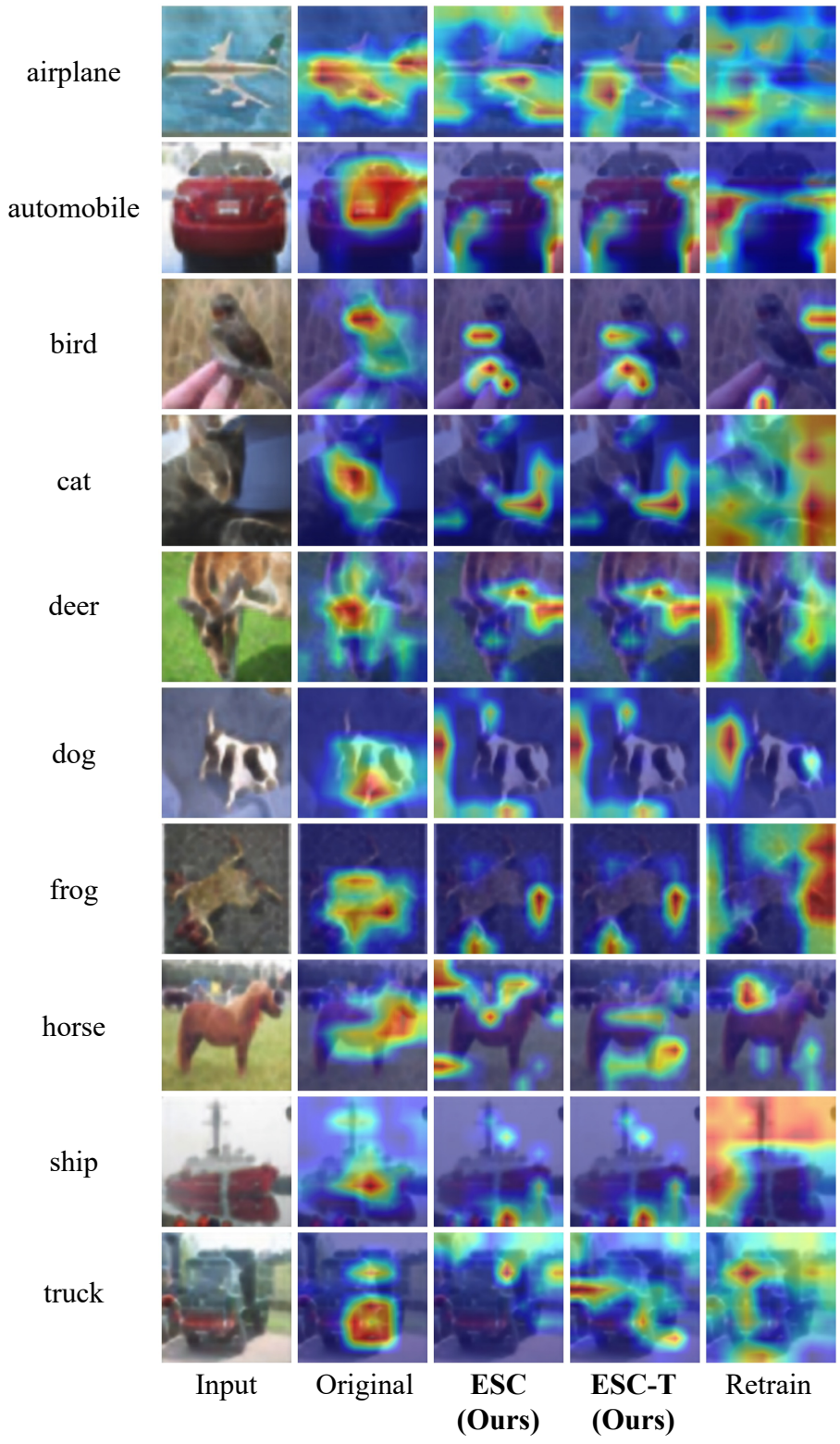


Figure 8. The activation map using Grad-CAM on the All-CNN model on the CIFAR-10 dataset.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 10

[2] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74. IEEE, 2018. 5

[3] Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3D generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 9

[4] 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. 1, 2, 5, 9

[5] 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. 1, 9

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 3

[7] Chongyu Fan, Jiancheng Liu, Yihua Zhang, Eric Wong, Dennis Wei, and Sijia Liu. Salun: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. In *The Twelfth International Conference on Learning Representations*, 2024. 1, 3

[8] 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*, pages 9304–9312, 2020. 1, 2, 5

[9] Hyunjune Kim, Sangyong Lee, and Simon S Woo. Layer attack unlearning: Fast and accurate machine unlearning via layer level attack and knowledge distillation. In *The 38th Annual AAAI Conference on Artificial Intelligence*. AAAI, 2024. 1, 2, 5, 6, 7, 9

[10] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 3

[11] Meghdad Kurmanji, Peter Triantafillou, Jamie Hayes, and Eleni Triantafillou. Towards unbounded machine unlearning. *Advances in Neural Information Processing Systems*, 36, 2023. 1, 5

[12] Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151, 1991. 9

[13] Jiancheng Liu, Parikshit Ram, Yuguang Yao, Gaowen Liu, Yang Liu, PRANAY SHARMA, Sijia Liu, et al. Model sparsity can simplify machine unlearning. *Advances in Neural Information Processing Systems*, 36, 2023. 1, 3

[14] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015. 5

[15] Juwon Seo, Sung-Hoon Lee, Tae-Young Lee, Seungjun Moon, and Gyeong-Moon Park. Generative unlearning for any identity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9151–9161, 2024. 9

[16] Anvith Thudi, Gabriel Deza, Varun Chandrasekaran, and Nicolas Papernot. Unrolling sgd: Understanding factors influencing machine unlearning. In *2022 IEEE 7th European Symposium on Security and Privacy (EuroS&P)*, pages 303–319. IEEE, 2022. 1, 2, 5

[17] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (11), 2008. 9, 10

[18] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 6

[19] Alexander Warnecke, Lukas Pirch, Christian Wressnegger, and Konrad Rieck. Machine unlearning of features and labels. In *Proc. of the 30th Network and Distributed System Security (NDSS)*, 2023. 2, 5, 9

[20] Ziyang Yuan, Yiming Zhu, Yu Li, Hongyu Liu, and Chun Yuan. Make encoder great again in 3d gan inversion through geometry and occlusion-aware encoding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2437–2447, 2023. 9