DiF3CON: Diffusion Forgetting via Continual Unlearning

Cătălin Ciorcîrlan[1], Vlad Vasilescu[1], Ana Nicolae[1]
[1]SIGMA Lab, CAMPUS Research Institute, POLITEHNICA Bucharest
Bucharest, Romania
{catalin.ciocirlan, vlad.vasilescu, ana_antonia.neacsu}@upb.ro

***Abstract*—Diffusion models have emerged as the dominant framework for image generation, offering high fidelity samples and flexible conditioning. However, they may pose risks with regard to generating copyright-sensitive or inappropriate content that may be exploited by malicious actors. Machine unlearning emerged as a counteraction to this phenomenon, seeking to efficiently remove the influence of specific training data without sacrificing the general performance. Prior work on diffusion unlearning mostly targets text-to-image or class-conditional models, while image-to-image settings such as inpainting remain largely unexplored. In this work, we introduce DiF3CON (*DiFFusion Forgetting via CONtinual unlearning*), a continual unlearning framework for diffusion-based inpainting models. Inspired by regularization and stability mechanisms from multi-task learning, DiF3CON simulates iterative, efficient class-wise data removal, without compromising the performance of the base diffusion model. Extensive experiments on the Places365 dataset show that DiF3CON achieves unlearning stability over significantly larger forgetting data volumes, advancing this relatively underexplored area of machine unlearning for diffusion-based image-to-image models. Our code is available at https://github.com/Cata400/dif3con.**

***Index Terms*—Machine Unlearning, Diffusion models, Inpainting, Continual Unlearning.**

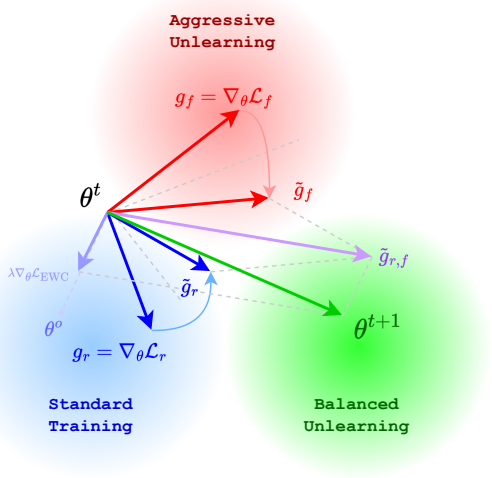


Fig. 1. We devise a *continual diffusion unlearning* algorithm which reliably preserves the performance of the base model θ^0, maintaining balance between significant information loss and inefficient unlearning.

I. Introduction

Generative models [1]–[3] have fundamentally altered the technological landscape, establishing themselves as valuable solutions in a wide range of fields. From entertainment applications in the creative industries [4], [5] to scientific research [6], [7], their seemingly elite performance was made possible primarily through unregulated access to huge amounts of data. By identifying and internalizing patterns in the desired data distribution, generative models are able to create synthetic content that adheres to these incorporated rules.

Along with this unstoppable wave for designing new generative systems, several legal and ethical challenges regarding intellectual property rights have emerged [8], [9]. The legal status of AI-generated data is mostly uncertain, raising questions of ownership and *copyrightability* [10]. Because generative models encode and internalize stylistic content into abstract latent representations, which are then leveraged during decoding to synthesize new samples, their behavior can be viewed as a form of reproduction to which existing regulatory frameworks may apply. To comply with users' *Right to be Forgotten* [11], data forgetting techniques have therefore been promoted as a means of mitigating potential legal exposure The simplest option would be to retrain using only compliant datasets, although this would not necessarily avoid generating undesired content. Moreover, retraining a model which was substantially tuned on multiple data sources raises practical considerations, leading to increased costs and energy consumption.

Machine Unlearning (MU) [12]–[14] is a sub-field which tackles the problem of *erasing* certain concepts from deployed AI systems, while conserving their trustworthiness. Such techniques come as an advantageous alternative to standard retraining, allowing for fast and efficient concept deletion in both legal and ethical-related scenarios. For generative models, machine unlearning has been mostly applied on text-to-image models to erase Not Safe For Work (NSFW) content [15], [16] or entire objects [17], [18]. Nevertheless, previous works mainly deal with erasing single concepts from generative models, which may prove suboptimal in some scenarios. Additionally, unlearning is, in itself, a highly unstable process [19], easily leading to major useful information loss.

In this work we introduce DiF3CON, a continual unlearning framework for image-to-image generative inpainting models. Unlike previous efforts, we push beyond standard 1-concept evaluations and design a general framework tailored

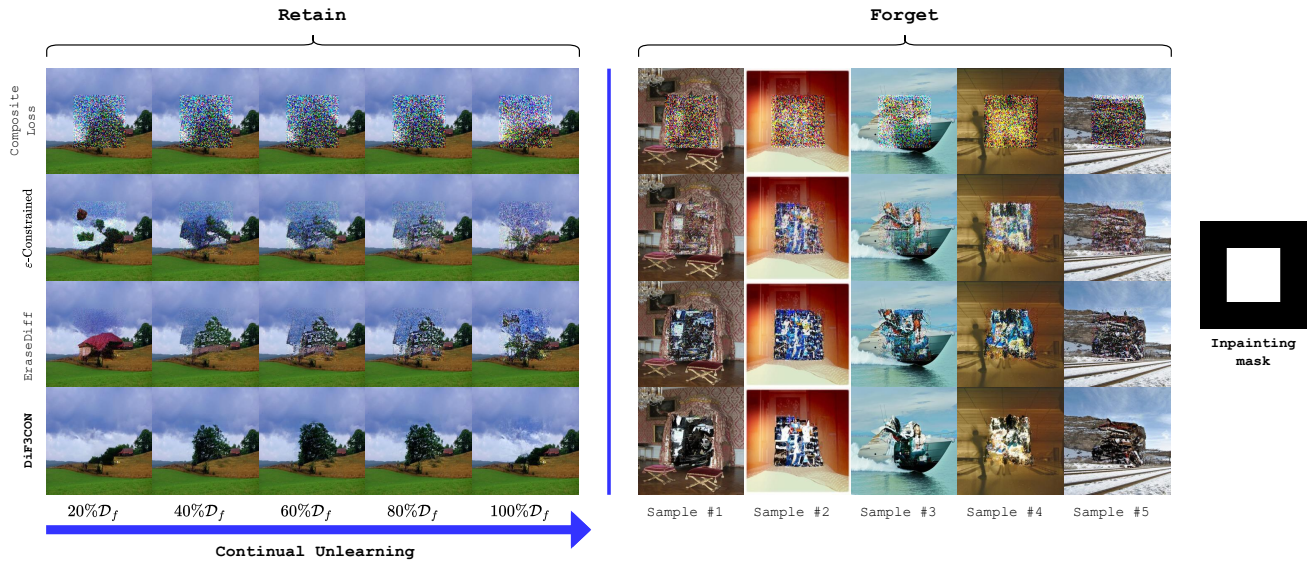


Fig. 2. Performance comparison on a sample from the retain set (left) under increasing forget set size $\mathcal{D}_f$, and on the forget set (right) after complete unlearning ($100\%\mathcal{D}_f$). Same inpainting mask is used for all examples.

for continuous unlearning of significant amounts of data. Our contributions can be summarized as follows:

- We introduce DiF3CON, an unlearning framework for generative inpainting models, designed for achieving balance between preserving original performance and aggressive information forgetting (Fig. 1).
- We conduct comprehensive evaluations in both easy (Sec. IV-B) and challenging unlearning scenarios (Sec. IV-C), proving the reliability of our method in unlearning high amounts of data.
- Finally, we carry out several ablation studies to determine a set of guidelines for developing efficient unlearning systems.

II. Related Work

Generative models. Diffusion-based generative models were first popularized by DDPM [1], which formulates data generation as the reversal of a gradual noising process. This approach demonstrated that iteratively denoising Gaussian noise can yield samples of remarkable quality. This view of generation as stochastic denoising has since become a central paradigm, enabling flexible conditional modeling by appropriately steering the reverse process [20], [21]. Building on this foundation, Ho and Salimans [22] introduced Classifier-Free Guidance (CFG), a conditioning technique that removes the need for an external classifier to guide the diffusion process. By jointly training the model on both conditional and unconditional data, this method simplifies the architecture while substantially improving controllability. A further key milestone was the proposal of Latent Diffusion Models (LDM) [2], which represent the basis for subsequent Stable Diffusion versions [23]. Instead of running the forward-backward diffusion directly in pixel space, LDMs operate in the latent space of largely pretrained auto-encoders, drastically reducing computational cost and enabling scalable high-resolution synthesis. This latent formulation also facilitates multimodal conditioning, allowing the model to be guided by text [24], semantic maps [25], or other images [26]. It has led to a wide range of practical applications in computer vision, among which image restoration has emerged as a particularly prominent one. Among these advancements, Palette [27] is a unified image-to-image translation framework based on conditional diffusion models that achieves great performance across diverse tasks, such as colorization, inpainting, uncropping, and JPEG restoration, without any task-specific auxiliary losses.

Machine Unlearning is a relatively recent research area aiming to provide a safe and reliable mechanism in mitigating risks associated with the potential misuse of generative models. The goal of MU is to remove from a pre-trained model the concepts associated with a subset of data used during training, further referred to as the *forget set* [12]. Since retraining the model without this data may be computationally prohibitive, a variety of more efficient approaches have been proposed to approximate the effect of retraining. Prior work in MU primarily focused on classification tasks, exploring different strategies such as gradient ascent on the forget objective [13], re-labeling the forget data with randomly chosen incorrect labels [14], or leveraging meta-learning frameworks to facilitate rapid unlearning updates [28]. Research on MU for generative models has started to emerge in recent years, primarily focusing on concept erasure in the context of text-to-image or class-conditional models [16], [29], [30]. Finding an effective compromise in this underlying conflicting context remains a challenging open optimization problem, for which only initial solution strategies have been proposed so far.

DualOptim [31] is a plug-and-play algorithm designed for

general MU frameworks, which employs alternating optimization of data forgetting and data retaining, decoupling their momentum dynamics and improving stability and performance. SalUn [32] is a framework guided by gradient weight saliency, leveraging only the most relevant weights for achieving the best unlearning performance. EraseDiff [16] proposes an algorithm for class- and text-conditional diffusion models which reformulates the original bi-objective optimization into a single-constrained problem. By using a forgetting value function as a constraint on the forget loss, it achieves a Pareto-optimal balance between concept preservation and erasure. One limitation of these works is that no further unlearning iterations are evaluated, their performance being solely assessed on single-{class, concept} forgetting.

MU in the context of image-to-image diffusion models remains, however, under-explored. The previous work of Li *et al*. [33] proposes an algorithm that formulates unlearning as a bounded optimization problem, operating solely on the encoder in a teacher-student setup. It aligns latent space embeddings of retained data between the models, while pushing forget embeddings towards Gaussian noise using an L_2-based mutual information approximation. Feng *et al*. [34] introduces a controllable unlearning framework for image-to-image models, framing unlearning as an ε-constrained bi-objective optimization problem, maximizing divergence on forget samples while minimizing it on retain samples. These approaches employ a large number of forget classes during training, aiming to unlearn them simultaneously; however, this setup is likely to degrade the model's utility on the retained data, given their aggressive forgetting strategy.

Continual Learning is a fundamental research field that describes how a model can efficiently acquire knowledge from continuous data streams over time. This paradigm emphasizes the ability to integrate new information while preserving previously learned knowledge, thereby mitigating the phenomenon of catastrophic forgetting, enabling models to adapt to new tasks, and making them more flexible and robust [35]–[38]. In the context of MU, this intersection represents an emerging and challenging research frontier that remains sparsely investigated.

III. Problem Formulation

A. Preliminaries

Inpainting. For RGB images $x \in \mathbb{R}^{H\times W\times C}$ and inpainting masks $m \in \{0,1\}$, the general goal of inpainting is to train a network ϵ_θ that faithfully reconstructs x, provided only the visible pixels [27], [39], [40]:

$$\theta^* = \arg\min_\theta \mathbb{E}_{x,m}\Big[\big\|\epsilon_\theta\big(x\odot(1-m)\big)-x\big\|_2\Big]. \quad (1)$$

In the following, we briefly discuss inpainting in the context of diffusion models, setting the stage for introducing our unlearning framework.

Diffusion Models. DDPMs [1] are generative models trained to reverse a gradual *forward* noising process through step-by-step *backward* denoising operations. The forward process starts with a clean image x_0 and creates a noisy observation $x_t = \sqrt{\bar\alpha_t}x_0 + \sqrt{1-\bar\alpha_t}\eta, \eta \sim \mathcal{N}(0,\mathbf{I})$, for a timestep $t \sim \mathcal{U}(1,T), T \in \mathbb{N}^+$ and scheduling parameters $\{\beta_t, \alpha_t = 1-\beta_t, \bar\alpha_t = \prod_{i=1}^t \alpha_i\}_{t=1}^T$. Then, a denoiser ϵ_θ is trained to predict η given x_t and some optional guidance context c:

$$\theta^* = \arg\min_\theta \|\epsilon_\theta(x_t,t,c)-\eta\|_2. \quad (2)$$

For inpainting applications, the guidance context is usually given by the unmasked pixels $c = x_0 \odot (1-m)$. During inference, starting with a random sample $\widehat{x}_T \sim \mathcal{N}(0,\mathbf{I})$ the denoiser is applied step-by-step until retrieving $\widehat{x}_0$:

$$\widehat{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\big(\widehat{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar\alpha_t}}\epsilon_\theta(\widehat{x}_t,t,c)\big) + \sqrt{\beta_t}z, \quad (3)$$

with $z \sim \mathcal{N}(0,\mathbf{I})$. Equation (3) corresponds to the standard DDPM backward process, which requires transitioning through all the steps $t \in \{1,\dots,T\}$ to arrive at $\widehat{x}_0$. DDPM usually requires a large number of steps T (*e.g.* $T = 1000$) in order for the backward process to properly approximate a Gaussian distribution [41], which makes sampling from the model extremely slow. Several techniques have been developed to significantly reduce the number of sampling steps, using the same network ϵ_θ [20], [42]. We simply refer to the network's prediction as $\epsilon_\theta(x_t)$ when the other arguments are off-topic.

B. Machine Unlearning

Let $\mathcal{D} = \{(x_i,y_i)\}_{i=1}^N$ denote the image dataset on which ϵ_{θ^0} was trained on, where x_i is an input image and y_i is its associated label. We will further refer to ϵ_{θ^0} as the *teacher* model. The goal of MU is to optimize a new set of weights θ^u s.t. ϵ_{θ^u} generates incorrect predictions on a subset $\mathcal{D}_f \subset \mathcal{D}$, while maintaining the performance on $\mathcal{D}_r = \mathcal{D}\setminus\mathcal{D}_f$. We refer to the former as the *forget set* and to the latter as the *retain set*. Although there are several formulations for the goal of unlearning, a general form can be defined as:

$$\theta^u = \arg\min_\theta \mathbb{E}_{(x,y)\sim\mathcal{D}}\big[\mathcal{L}_r(x,y,\theta)+\mathcal{L}_f(x,y,\theta)\big], \quad (4)$$

where $\mathcal{L}_r(x,y,\theta) = \mathbb{1}_{x\in\mathcal{D}_r}\mathcal{L}(x,y,\theta)$ and $\mathcal{L}_f(x,y,\theta) = \mathbb{1}_{x\in\mathcal{D}_f}\mathcal{L}(x,\bar y,\theta)$ are the retain and forget losses, respectively, $\mathcal{L}$ is the loss function previously used to optimize ϵ_{θ^0}, $\mathbb{1}_{(\cdot)}$ is the indicator function, and $\bar y \neq y$ is a proxy target assigned to forget samples. For classification, $\bar y$ is usually randomly sampled from the set of retain labels [14].

Balancing the conflicting objectives of $\mathcal{L}_r$ and $\mathcal{L}_f$ is challenging, even for simple tasks. Deep models often learn complex representations where features are intermingled across different classes or concepts, making them hard to separate [43]. Additionally, $\mathcal{L}_f$ gradients usually have larger magnitudes, due to their inherent opposition to the natural learning trajectory, requiring several consecutive updates to help with optimization stability.

C. Diffusion Unlearning

As is widely practiced, we consider denoising networks of the form $\epsilon_\theta := D_{\theta_D} \circ E_{\theta_E}$, where $D_{\theta_D}, E_{\theta_E}$ are symmetric decoder and encoder models. In one of the first works addressing image-to-image unlearning, Li *et al*. [33] devised the following objectives to optimize θ_E:

$$\mathcal{L}_r = \mathbb{1}_{x_0 \in \mathcal{D}_r} \| E_{\theta_E}(x_t) - E_{\theta_E^0}(x_t) \|_2, \tag{5}$$

$$\mathcal{L}_f = \mathbb{1}_{x_0 \in \mathcal{D}_f} \| E_{\theta_E}(x_t) - E_{\theta_E^0}(z) \|_2, \tag{6}$$

where $z \sim \mathcal{N}(0, \mathbf{I})$, $\alpha > 0$, while the decoder was kept frozen $\theta_D = \theta_D^0$. The goal is to push the embeddings on forget data towards unsystematic information, inducing random behavior on $\mathcal{D}_f$. However, this is counterintuitive with respect to the fine-tuning literature, whereas the decoder is usually modified to allow for data distribution shifts. Additionally, this approach fails to account for the fact that any error introduced by the newly optimized E_{θ_E} would eventually be magnified by the fixed decoder $\mathcal{D}_{\theta_D}$, leading to inadequate reconstruction of both retained and forgotten data.

As previously discussed, a central challenge in unlearning lies in the joint optimization of the competing objectives $\mathcal{L}_r$ and $\mathcal{L}_f$, which induces an inherently multi-objective trade-off between retain and forget. Huang *et al*. [28] address this by introducing a *gradient harmonization* scheme, inspired by multi-task learning, which reprojects conflicting gradients toward a unified optimization goal [44]. Similarly, Ko *et al*. [30] propose restricting each task's gradient to the orthogonal complement of the other, ensuring joint loss reduction, while Meng *et al*. [45] extends this idea through gradient rescaling and adaptive weighting to maintain balanced task contributions. Beyond gradient conflicts, the computational burden of training large models raises concerns about maintaining model utility after unlearning. To this end, Kirkpatrick *et al*. [35] introduce *Elastic Weight Consolidation* (EWC), a continual learning method that regularizes new task updates to prevent drastic deviations from prior models. *Fisher Information Matrix* (FIM) is used to assign importance to each weight, efficiently preserving performance while mitigating catastrophic forgetting.

D. Proposed framework

In the following, we define our general unlearning algorithm DiF3CON, aimed at preserving the performance of the base generative model, while performing unlearning in an iterative, continual manner. Our framework is shown in Algorithm 1.

Given a retain set $\mathcal{D}_r$ containing $K_r \geq 1$ classes to be preserved and a forget set $\mathcal{D}_f$ containing $K_f \geq 1$ classes to be unlearned from a base diffusion model ϵ_{θ^0}, we uniformly partition $\mathcal{D}_f$ over the total unlearning epochs E, decomposing the entire process into multiple similar sub-tasks, allowing for a smoother optimization process. Since the base diffusion model has already undergone extensive training on $\mathcal{D}_r \cup \mathcal{D}_f$, we do not foresee the need to use the entire sets during each unlearning epoch. Instead, we construct the training set for each epoch $i = \overline{1, E}$ as the union between *retain buffer* $\mathtt{r}^i$ and *forget buffer* $\mathtt{f}^i$. At each iteration, we set $\mathtt{r}^i$ as a random subset of images from $\mathcal{D}_r$, while $\mathtt{f}^i$ contains all images from the currently unlearned classes combined with randomly sampled data from previously unlearned classes. The latter's purpose is twofold: to prioritize the currently unlearned classes and to ensure previous information reliably remains forgotten.

Algorithm 1: DiF3CON

Input: Teacher weights θ^0, retain/forget sets $\mathcal{D}_r/\mathcal{D}_f$, buffer fractions p_r/p_f, forget classes K_f, λ, α
Output: Unlearned model $\theta^u = \theta^E$

$\mathtt{r}^0, \mathtt{f}^0 \leftarrow [], []$ `// retain/forget buffers`
$\mathcal{P} \leftarrow \left\{ \overline{i\lfloor \frac{K_f}{E} \rfloor, (i+1)\lfloor \frac{K_f}{E} \rfloor} \right\}_{i=1}^{E}$ `// forget classes partitioning`
for $i \leftarrow 1$ **to** E **do**
  $\theta^i \leftarrow \theta^{i-1}$
  $\mathtt{r}^i = \mathtt{RandomSample}(\mathcal{D}_r, p_r)$
  $\mathtt{f}^i = [\mathtt{f}^{i-1}; \mathcal{D}_f[\mathcal{P}_i]]$
  $\mathcal{D}_{\text{train}}^{(i)} \leftarrow [\mathtt{r}^i; \mathtt{f}^i]$
  for $b \in \mathcal{D}_{\text{train}}^{(i)}$ **do**
    $\ell_r, \ell_f = \mathcal{L}_r(b, \theta^i), \mathcal{L}_f(b, \theta^i)$ `//` (7), (8)
    $\widetilde{g}_{r,f} = \mathtt{GH}_\alpha(\nabla_{\theta^i}\ell_r, \nabla_{\theta^i}\ell_f)$ `//` (9)
    $g_{\text{EWC}} = \nabla_{\theta^i}\mathcal{L}_{\text{EWC}}(\theta^i, \theta^{i-1}, \lambda)$ `//` (11)
    $\theta^i \leftarrow \theta^i - \mathtt{lr} \cdot (\widetilde{g}_{r,f} + g_{\text{EWC}})$
  $\mathtt{f}^i = \mathtt{RandomSample}(\mathtt{f}^i, p_f)$

During initial experiments, we observed that fine-tuning only the encoder E, while keeping D frozen, yielded sub-optimal performance. Hence, we choose instead to keep the encoder frozen, preserving the latent space representation power of the initial model, while fine-tuning D in order to correctly reconstruct $\mathcal{D}_r$ and increase perturbation on $\mathcal{D}_f$. Given this setting, we selected the following objectives:

$$\mathcal{L}_r = \mathbb{1}_{x_0 \in \mathcal{D}_r} \| \epsilon_{\theta^u}(x_t, t, c_r) - \epsilon_{\theta^0}(x_t, t, c_r) \|_2, \tag{7}$$

$$\mathcal{L}_f = \mathbb{1}_{x_0 \in \mathcal{D}_f} \| \epsilon_{\theta^u}(x_t, t, c_f) - \epsilon_{\theta^0}(x_t', t, c_f) \|_2, \tag{8}$$

where c_r, c_f are the corresponding context tensors, and x_t' is a noisy retain sample within the same batch. Since the reconstruction of ϵ_{θ^0} on x_t' will lead to a mismatch with c_f, this will drive ϵ_{θ^u} to forget the information contained within c_f, implicitly from $\mathcal{D}_f$, while preserving the reconstruction performance of ϵ_{θ^0}. On the other hand, $\mathcal{L}_r$ maintains the initial objective, hence pushing towards knowledge distillation over the domain of $\mathcal{D}_r$.

In order to balance any possible divergence between $\mathcal{L}_f$ and $\mathcal{L}_r$, we apply a gradient harmonization scheme before each weight update. In a similar manner to [30], we define the gradient harmonization process $\widetilde{g}_{r,f} := \mathtt{GH}_\alpha(g_r, g_f)$ as:

$$\widetilde{g}_{r,f} = \alpha \widetilde{g}_r + (1 - \alpha)\widetilde{g}_f, \tag{9}$$

$$\widetilde{g}_{r/f} = g_{r/f} - \mathbb{1}_{g_r^\top g_f < 0} \frac{g_r^\top g_f}{\|g_{f/r}\|_2} g_{f/r} \tag{10}$$

where $g_r = \nabla_{\theta^i}\mathcal{L}_r$ and $g_f = \nabla_{\theta^i}\mathcal{L}_f$ are the initial retain and forget gradients, respectively, and $\alpha > 0$ encodes the importance we convey to the projected retain gradient.

To better preserve the teacher's performance, particularly in the regime of large K_f, we additionally perform an EWC update with respect to the following objective:

$$\mathcal{L}_{\text{EWC}}(\theta^u, \theta^*, \lambda) = \lambda \|\theta^u - \theta^*\|_2^2, \tag{11}$$

where θ^* encodes some anchor w.r.t. which we want to stabilize the unlearned weights θ^u, and $\lambda \geq 0$ encodes the regularization strength. In our case, we define θ^* as the weights optimized during the previous unlearning step. This plays a role in θ^u's learning plasticity, ensuring smooth adaptation to the new forget concepts. We chose the ℓ_2 squared distance for efficiency reasons, arguing that more sophisticated mechanisms such as FIM, discussed in Sec. III-C, would impose a significant computational overhead in complex tasks such as generative modeling.

IV. Experiments

A. Setup

Inpainting. We evaluate our proposed algorithm on the task of inpainting, fixing the teacher model to the pretrained checkpoint of *Palette* [27]. We conduct our experiments on the Places365 dataset [46], containing 5000 train images and 100 validation images for each class. Training is performed using a random mix of irregular and rectangular masks.

Unlearning. We consider two unlearning scenarios: *single-class unlearning* – where we choose a single class to be part of our unlearning set $\mathcal{D}_f$ – and *multi-class unlearning* – where we randomly assign $K_f > 1$ classes to be part of $\mathcal{D}_f$. In both cases, the retain set $\mathcal{D}_r$ contains the same K_r classes.

During both evaluations, we present the teacher's performance as an anchor for measuring retaining and forgetting performance. We additionally consider the following baselines: i) *Composite Loss* [33] with $\alpha_{\text{forget}} = 0.25$; ii) *$\varepsilon$-Constrained* [34] with ε_{min} for single-class unlearning and ε_{max} for multi-class; and iii) *EraseDiff* [16] with $K = 1$. We evaluate all algorithms using 25% center crop inpainting masks. As performance metrics, we report separately for the retain and forget datasets: i) Frechét Inception Distance (FID) between the generated images and the ground truth [47]; and ii) Inception Score (IS) of the generated images [48]. All subsequently analyzed methods and models were trained from scratch to ensure a fair comparison within our evaluation context.

A great number of prior works in diffusion unlearning use class-conditional diffusion models; however, unlearning for unconditional generation represents a more difficult task, due to the lack of explicit class separability. These works often only assess the performance when forgetting a single class or concept [16], [31], [32], but do not verify their solution on subsequent unlearning steps. On the other hand, recent studies on image-to-image models [33], [34] choose to unlearn a large number of classes at a time, which is unrealistic and is prone to affect the retain performance.

TABLE I
Results on the scenario of single-class unlearning ($K_r = 50$, $K_f = 1$). For each class 100 samples were generated.

Method	Retain		Forget	
	IS↑	FID↓	IS↓	FID ↑
Teacher	$14.88_{\pm 0.45}$	12.61	$3.93_{\pm 0.32}$	118.78
Composite Loss	$14.60_{\pm 0.48}$	14.23	$3.69_{\pm 0.47}$	123.61
ε-Constrained	$14.52_{\pm 0.57}$	13.77	$3.71_{\pm 0.52}$	**161.66**
`EraseDiff`	**14.96**$_{\pm 0.46}$	13.39	$3.74_{\pm 0.57}$	125.40
DiF3CON	$14.78_{\pm 0.53}$	**12.74**	**3.68**$_{\pm 0.61}$	132.33

For multi-class unlearning we decided to simulate a scenario closer to reality, where data is requested for unlearning in smaller, incremental steps. Thus, we choose to apply our unlearning procedure on increments of 20% forget classes per step, for a total of K_f forget classes.

As discussed in Sec. III-D, for efficiency reasons we set buffers size to $p_r = p_f = 1\%$. For Algorithm 1 we chose $\alpha = 0.5$ and $\lambda = 0.1$. Ablation studies on these hyperparameters and on several choices for $\mathcal{L}_f$ are available in Sec. IV-D.

B. Single-Class Unlearning

Following [16], [31], [32], we first select the simple scenario of a single forget class ($K_r = 50$, $K_f = 1$) with the aim of creating a starting point for our subsequent evaluations. The evaluation w.r.t. FID is straightforward: we require similar (↓) feature distributions on $\mathcal{D}_r$ w.r.t. the teacher, and dissimilar (↑) on $\mathcal{D}_f$. In terms of IS, however, we believe some justifying words are in order: for retain images we wish to introduce as little distortion as possible, translating to low uncertainty for an out-of-the-box classifier, hence a higher IS; on the contrary, the distortion on forget data should lead to increased uncertainty, corresponding to the obstruction of discriminative image features, hence a lower IS is desired.

The performance of each method is listed in Tab. I. DiF3CON obtains consistent performance for both retaining and forgetting sets, leading to the most uncertain reconstructions for images in the forget class, while maintaining the closest performance on retain w.r.t the teacher model. Figure 3 presents visual examples showing that, while all methods reconstruct retained data similarly, our approach achieves superior obstruction in the inpainted area of the forgotten class.

C. Multi-Class Unlearning

Moving one step further, we validate our solution in a more challenging context of multiple forget classes ($K_r = 50$, $K_f = 25$). We argue that the unlearning process should be treated as a continuous problem, instead of a stand-alone task, since there is no limit on the number of possible consecutive erasure requests. The main objective of this continuous process is to maintain an appropriate performance on both the retain and forget datasets along all unlearning steps. Towards this goal, we designed DiF3CON to be plug-and-play, and we evaluate it in conjunction with each of the aforementioned baselines. Our

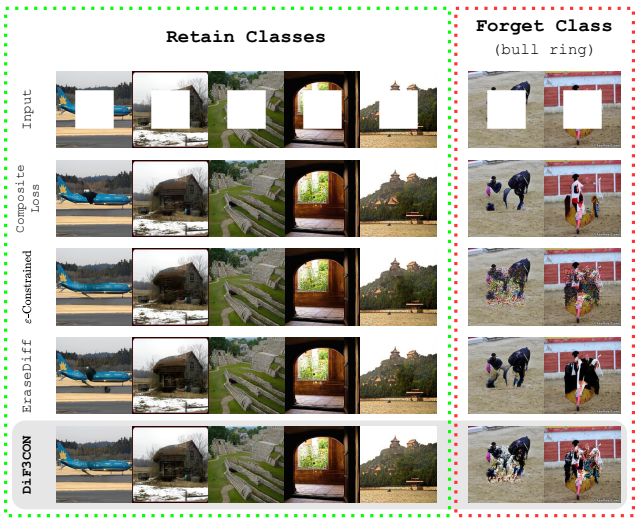


Fig. 3. Inpainted examples after 1-class unlearning. DiF3CON results in significant degradation on the forget class, while maintaining utility on the other concepts.

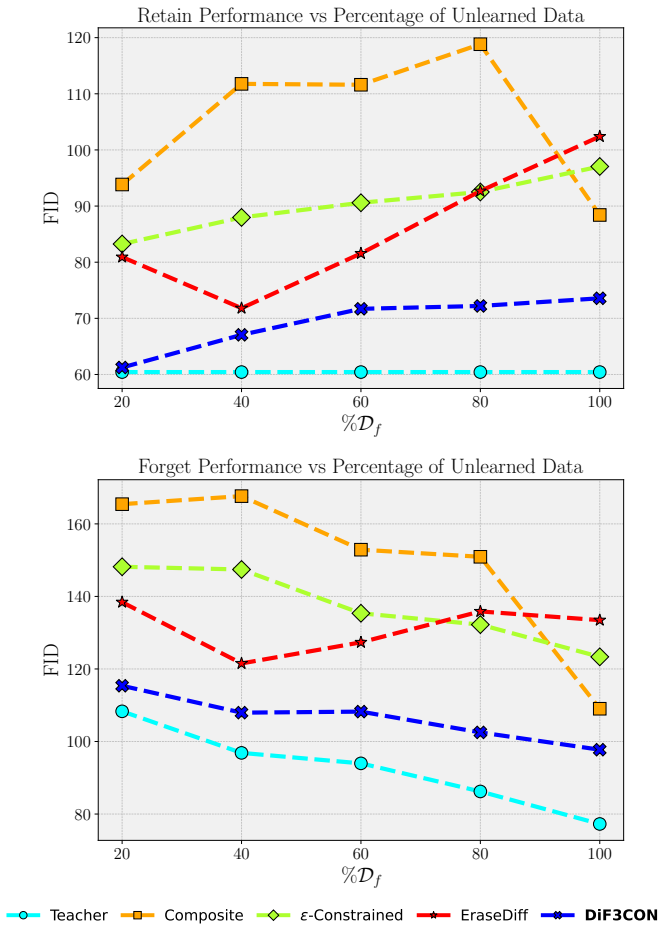
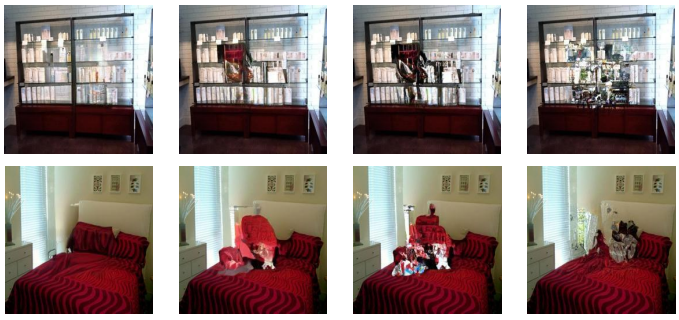


Fig. 4. Evolution of FID for retain and forget sets at different stages during continual unlearning ($K_r = 50$, $K_f = 25$).



(a) Teacher (b) 20%$\mathcal{D}_f$ (c) 60%$\mathcal{D}_f$ (d) 100%$\mathcal{D}_f$

Fig. 5. Evolution (from left to right) of sample reconstruction from classes ("beauty salon", "bedroom") unlearned in the first epoch of Algorithm 1, at different subsequent unlearning steps.

results are shown in Fig. 4, where we analyze the evolution of the retain and forget FID scores during the course of our algorithm. For computational efficiency, we evaluate all methods on reduced versions of the datasets, each containing 10 images per class. This setting is designed to highlight how performance scales under increasingly challenging yet realistic unlearning scenarios. All methods use the same retain and forget buffer size of 1% of the original data, with the exception of *Composite Loss*, which requires a 5% retain buffer for a meaningful comparison. We attribute this behavior to the fact that *Composite Loss* fine-tunes only the encoder, leading the fixed decoder to produce unstable reconstructions when trained on limited data. As unlearned data increases, we see a drop in performance on the retained set for each algorithm. This happens because, as more data is unlearned, the model observes fewer retained samples compared to forgotten ones. Additionally, with more data to unlearn, the forgetting per sample decreases, leading to overlap between retained and forgotten sets. This overlap allows useful information to leak, lowering the FID on $\mathcal{D}_f$.

Across all algorithms, DiF3CON best preserves model utility throughout unlearning, while simultaneously achieving competitive forgetting performance. Importantly, both retain and forget FID scores must be considered when assessing unlearning strategies: eliminating the influence of $\mathcal{D}_f$ at the expense of substantially degrading the model's generative capacity on $\mathcal{D}_r$ fundamentally defeats the purpose of unlearning. The examples in Fig. 2 illustrate that our method produces more natural obstructions for samples from $\mathcal{D}_f$, while optimally maintaining visual fidelity on $\mathcal{D}_r$.

D. Ablation Studies

Forget objective. Besides the previously discussed forget loss, we experiment with three other variants to empirically analyze their unlearning performance:

$$\mathcal{L}_f^{c_r} = \|\epsilon_{\theta^u}(x_t, t, c_f) - \epsilon_{\theta^0}(x_t, t, c_r)\|_2, \quad (12)$$

$$\mathcal{L}_f^{\max} = -\|\epsilon_{\theta^u}(x_t, t, c_f) - \epsilon_{\theta^0}(x_t, t, c_f)\|_2, \quad (13)$$

$$\mathcal{L}_f^{z} = \|\epsilon_{\theta^u}(x_t, t, c_f) - \epsilon_{\theta^0}(z, t, c_f)\|_2, \quad (14)$$

where $z \sim \mathcal{N}(0, \mathbf{I})$, and c_r, c_f are computed using elements within the same batch. Compared to the loss in (8), $\mathcal{L}_f^{c_r}$

TABLE II
ABLATION STUDIES FOR OUR METHOD. PERFORMANCE IS REPORTED AT BOTH INITIAL ($20\%\mathcal{D}_f$) AND FINAL ($100\%\mathcal{D}_f$) STAGES OF UNLEARNING, FOR 10 GENERATED IMAGES PER CLASS.

Training Config.		$20\%\mathcal{D}_f$				$100\%\mathcal{D}_f$			
		Retain		Forget		Retain		Forget	
		IS ↑	FID ↓	IS ↓	FID ↑	IS ↑	FID ↓	IS ↓	FID ↑
Forget Loss ($\lambda = 0.1, \alpha = 0.5$)	$\mathcal{L}_f$ (Eq. 8)	$10.21_{\pm 1.31}$	61.24	$2.81_{\pm 0.41}$	115.36	$9.67_{\pm 0.95}$	73.57	$7.83_{\pm 0.95}$	97.73
	$\mathcal{L}_f^{c_r}$ (Eq. 12)	$10.01_{\pm 1.24}$	70.91	$3.00_{\pm 0.40}$	132.90	$9.37_{\pm 1.23}$	88.47	$7.25_{\pm 0.95}$	116.10
	$\mathcal{L}_f^{\max}$ (Eq. 13)	$4.71_{\pm 0.50}$	169.89	$2.25_{\pm 0.48}$	242.52	$4.27_{\pm 0.56}$	180.52	$3.54_{\pm 0.28}$	205.68
	$\mathcal{L}_f^z$ (Eq. 14)	$10.25_{\pm 1.07}$	64.80	$2.79_{\pm 0.42}$	116.94	$8.36_{\pm 0.96}$	93.90	$7.03_{\pm 0.88}$	120.91
EWC	$\lambda = 0$	$8.20_{\pm 0.96}$	88.65	$2.87_{\pm 0.55}$	164.47	$8.54_{\pm 0.99}$	88.23	$6.78_{\pm 0.73}$	112.98
GH	$\alpha = 0.1$	$10.05_{\pm 1.33}$	62.19	$2.90_{\pm 0.44}$	115.70	$9.54_{\pm 0.97}$	75.56	$7.63_{\pm 0.78}$	101.08
	$\alpha = 0.9$	$10.17_{\pm 1.28}$	60.25	$2.94_{\pm 0.42}$	110.54	$9.92_{\pm 1.04}$	70.74	$8.08_{\pm 1.10}$	93.46
$p_r = p_f = 100\%$ (all data)		$9.80_{\pm 1.05}$	63.17	$2.76_{\pm 0.52}$	147.19	$8.79_{\pm 1.06}$	88.73	$6.24_{\pm 0.83}$	133.41

aims to enforce a reconstruction towards clean retain images by semantic information transfer. On the other hand, $\mathcal{L}_f^{\max}$ directly maximizes the error between teacher and ϵ_{θ^u} on $\mathcal{D}_f$, encoding a more aggressive unlearning direction. Lastly, $\mathcal{L}_f^z$ is similar to the Composite Loss [33] approach, although we account for the final output rather than latent space features.

Table II presents the results obtained for all ablation studies. We observe that although models trained using $\mathcal{L}_f^{c_r}$ and $\mathcal{L}_f^z$ have a closer performance to $\mathcal{L}_f$ on the first unlearning step, the final model notably diverged on the retain set, significantly dropping in performance. Furthermore, the maximizing loss $\mathcal{L}_f^{\max}$ completely degraded the model, causing complete loss of generative power regardless of $\mathcal{L}_r$.

EWC. To illustrate the efficiency of the additional objective in (11), we unlearn a model by setting EWC factor $\lambda = 0$. This yields a significant performance drop on retain set, poisoning the subsequent unlearning steps, which underlines the importance of imposing plasticity on the model weights w.r.t. previous unlearning steps.

Gradient Harmonization. We perform an analysis on the contribution of the gradient harmonization hyperparameter α on the performance of the model. A larger α drifts the resulting gradient more towards the retain gradient, which results in a less pronounced forgetting, and vice-versa. The gap in performance between the larger and smaller values widens as the unlearning process advances.

Training data. While we initially proposed using only a fraction of the retain and previously forgotten data for efficiency reasons, we additionally examine the performance of DiF3CON when trained on the entire datasets. The results shown in the last row of Tab. II are computed using $\mathcal{L}_f^z$. Using the complete training data enhances the model's performance over several unlearning iterations, but this improvement comes with an increase in training time. Thus, we show that efficient unlearning can also be achieved with minimal data usage. As for computational acceleration, using only $p_r = p_f = 1\%$ yields a $\times 13.6$ speedup on a single RTX A6000.

We now pose the following question: ***Is unlearning maintained for initially unlearned classes?*** Fig. 5 shows examples from $\mathcal{D}_f$ unlearned in the first epoch ($i = 1$), reconstructed at subsequent unlearning steps (using weights $\theta^i, i > 1$). It can be easily observed that the erasure of data influence persists for each unlearning step. This enforces the results shown in Fig. 4, proving the reliability of our method in continuous unlearning scenarios.

V. CONCLUSIONS

We present DiF3CON, a plug-and-play continual unlearning framework for inpainting diffusion models, acting towards a balance between inefficient unlearning and aggressive forgetting, ensuring a trade-off between plasticity and elasticity w.r.t. the teacher model. Extensive experiments on a comprehensive large-scale dataset revealed that our algorithm exceeds previous methods in both simple scenarios of single-concept unlearning, as well as in challenging incremental unlearning contexts of large amounts of data. In addition, we discuss and analyze our particular algorithm choices in order to offer a recipe for designing efficient unlearning paradigms.

ACKNOWLEDGEMENT

This work was supported by the Complex Bilateral Project Romania–Republic of Moldova, "Sensory Engagement Network for Shared Experiences (SENSE)", funded under UEFISCDI PN-IV-PCB-RO-MD-2024-0392 and ANCD 25.80013.0807.47ROMD.

REFERENCES

[1] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *NeurIPS*, vol. 33, pp. 6840–6851, 2020.

[2] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *CVPR*, 2022, pp. 10 684–10 695.

[3] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," in *ICML*, 2021, pp. 8821–8831.

[4] R. Heigl, "Generative artificial intelligence in creative contexts: a systematic review and future research agenda," *Management Review Quarterly*, pp. 1–38, 2025.

[5] A. H. Nobari, M. F. Rashad, and F. Ahmed, "Creativegan: Editing generative adversarial networks for creative design synthesis," *arXiv preprint arXiv:2103.06242*, 2021.

[6] X. Zeng, F. Wang, Y. Luo, S.-g. Kang, J. Tang, F. C. Lightstone, E. F. Fang, W. Cornell, R. Nussinov, and F. Cheng, "Deep generative molecular design reshapes drug discovery," *Cell Reports Medicine*, vol. 3, no. 12, 2022.

[7] C. K. Reddy and P. Shojaee, "Towards scientific discovery with generative ai: Progress, opportunities, and challenges," in *AAAI*, vol. 39, 2025, pp. 28 601–28 609.

[8] European Commission, "Proposal for a Regulation Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act)," https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206, 2021, cOM/2021/206 final, 21 April 2021.

[9] L. de la Torre, "A guide to the california consumer privacy act of 2018," *Available at SSRN 3275571*, 2018.

[10] S. Perlmutter, "Copyright and artificial intelligence. part 2: Copyrightability," Technical Report, United States Copyright Office, Tech. Rep., 2025.

[11] C. J. Hoofnagle, B. Van Der Sloot, and F. Z. Borgesius, "The European Union general data protection regulation: what it is and what it means," *Information & Communications Technology Law*, vol. 28, no. 1, pp. 65–98, 2019.

[12] L. Bourtoule, V. Chandrasekaran, C. A. Choquette-Choo, H. Jia, A. Travers, B. Zhang, D. Lie, and N. Papernot, "Machine unlearning," in *Symp. on Sec. and Priv.*, 2021, pp. 141–159.

[13] A. Halimi, S. Kadhe, A. Rawat, and N. Baracaldo, "Federated unlearning: How to efficiently erase a client in fl?" *arXiv preprint arXiv:2207.05521*, 2022.

[14] L. Graves, V. Nagisetty, and V. Ganesh, "Amnesiac machine learning," in *AAAI*, vol. 35, 2021, pp. 11 516–11 524.

[15] A. Heng and H. Soh, "Selective amnesia: A continual learning approach to forgetting in deep generative models," *NeurIPS*, vol. 36, pp. 17 170–17 194, 2023.

[16] J. Wu, T. Le, M. Hayat, and M. Harandi, "EraseDiff: Erasing data influence in diffusion models," *arXiv preprint arXiv:2401.05779*, 2024.

[17] R. Gandikota, J. Materzynska, J. Fiotto-Kaufman, and D. Bau, "Erasing concepts from diffusion models," in *CVPR*, 2023, pp. 2426–2436.

[18] G. Zhang, K. Wang, X. Xu, Z. Wang, and H. Shi, "Forget-me-not: Learning to forget in text-to-image diffusion models," in *CVPR*, 2024, pp. 1755–1764.

[19] N. George, K. N. Dasaraju, R. R. Chittepu, and K. R. Mopuri, "The Illusion of Unlearning: The Unstable Nature of Machine Unlearning in Text-to-Image Diffusion Models," in *CVPR*, 2025, pp. 13 393–13 402.

[20] C. Lu, Y. Zhou, F. Bao, J. Chen, C. Li, and J. Zhu, "Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps," *NeurIPS*, vol. 35, pp. 5775–5787, 2022.

[21] J. Choi, S. Kim, Y. Jeong, Y. Gwon, and S. Yoon, "ILVR: Conditioning Method for Denoising Diffusion Probabilistic Models," in *CVPR*, 2021, pp. 14 367–14 376.

[22] J. Ho and T. Salimans, "Classifier-free diffusion guidance," *arXiv preprint arXiv:2207.12598*, 2022.

[23] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, and the CompVis & Stability AI teams, "Stable Diffusion 2.1: High-Resolution Text-to-Image Latent Diffusion Model," 2022, version 2.1 release by Stability AI, December 2022. Available at https://github.com/Stability-AI/stablediffusion. [Online]. Available: https://huggingface.co/stabilityai/stable-diffusion-2-1

[24] H. Nam, G. Kwon, G. Y. Park, and J. C. Ye, "Contrastive denoising score for text-guided latent diffusion image editing," in *CVPR*, 2024, pp. 9192–9201.

[25] N. Giambi and G. Lisanti, "Conditioning diffusion models via attributes and semantic masks for face generation," *arXiv preprint arXiv:2306.00914*, 2023.

[26] C. Qin, S. Zhang, N. Yu, Y. Feng, X. Yang, Y. Zhou, H. Wang, J. C. Niebles, C. Xiong, S. Savarese *et al.*, "UniControl: A Unified Diffusion Model for Controllable Visual Generation In the Wild," *CoRR*, 2023.

[27] C. Saharia, W. Chan, H. Chang, C. Lee, J. Ho, T. Salimans, D. Fleet, and M. Norouzi, "Palette: Image-to-image diffusion models," in *ACM TOG*, 2022, pp. 1–10.

[28] M. H. Huang, L. G. Foo, and J. Liu, "Learning to unlearn for robust machine unlearning," in *ECCV*. Springer, 2024, pp. 202–219.

[29] S. Lu, Z. Wang, L. Li, Y. Liu, and A. W.-K. Kong, "Mace: Mass concept erasure in diffusion models," in *CVPR*, 2024, pp. 6430–6440.

[30] M. Ko, H. Li, Z. Wang, J. Patsenker, J. T. Wang, Q. Li, M. Jin, D. Song, and R. Jia, "Boosting alignment for post-unlearning text-to-image generative models," *NeurIPS*, vol. 37, pp. 85 131–85 154, 2024.

[31] X. Zhong, H. Luo, and C. Liu, "Dualoptim: Enhancing efficacy and stability in machine unlearning with dual optimizers," *arXiv preprint arXiv:2504.15827*, 2025.

[32] C. Fan, J. Liu, Y. Zhang, E. Wong, D. Wei, and S. Liu, "Salun: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation," *arXiv preprint arXiv:2310.12508*, 2023.

[33] G. Li, H. Hsu, C.-F. Chen, and R. Marculescu, "Machine Unlearning for Image-to-Image Generative Models," *CoRR*, 2024.

[34] X. Feng, Y. Li, C. Chen, L. Zhang, L. Li, J. Zhou, and X. Zheng, "Controllable Unlearning for Image-to-Image Generative Models via ε-Constrained Optimization," *arXiv preprint arXiv:2408.01689*, 2024.

[35] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska *et al.*, "Overcoming catastrophic forgetting in neural networks," *Proceedings of the national academy of sciences*, vol. 114, no. 13, pp. 3521–3526, 2017.

[36] S.-A. Rebuffi, A. Kolesnikov, G. Sperl, and C. H. Lampert, "ICARL: Incremental classifier and representation learning," in *CVPR*, 2017, pp. 2001–2010.

[37] L. Wang, X. Zhang, H. Su, and J. Zhu, "A comprehensive survey of continual learning: Theory, method and application," *IEEE TPAMI*, vol. 46, no. 8, pp. 5362–5383, 2024.

[38] J. Von Oswald, C. Henning, B. F. Grewe, and J. Sacramento, "Continual learning with hypernetworks," *arXiv preprint arXiv:1906.00695*, 2019.

[39] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Free-form image inpainting with gated convolution," in *ICCV*, 2019, pp. 4471–4480.

[40] S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Globally and locally consistent image completion," *ACM TOG*, vol. 36, no. 4, pp. 1–14, 2017.

[41] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *arXiv preprint arXiv:2010.02502*, 2020.

[42] Q. Zhang, M. Tao, and Y. Chen, "gDDIM: generalized denoising diffusion implicit models," in *ICLR*, 2023.

[43] Y. Wang, Q. Wang, F. Liu, W. Huang, Y. Du, X. Du, and B. Han, "Gru: Mitigating the trade-off between unlearning and retention for llms," *arXiv preprint arXiv:2503.09117*, 2025.

[44] T. Yu, S. Kumar, A. Gupta, S. Levine, K. Hausman, and C. Finn, "Gradient surgery for multi-task learning," *NeurIPS*, vol. 33, pp. 5824–5836, 2020.

[45] F. Meng, Z. Xiao, Y. Zhang, and J. Wang, "RI-PCGrad: Optimizing multi-task learning with rescaling and impartial projecting conflict gradients," *Applied Intelligence*, vol. 54, no. 22, pp. 12 009–12 019, 2024.

[46] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million Image Database for Scene Recognition," *IEEE TPAMI*, 2017.

[47] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," *NeurIPS*, vol. 30, 2017.

[48] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANs," *NeurIPS*, vol. 29, 2016.