

Continual Learning of Personalized Generative Face Models with Experience Replay

Annie N. Wang
UNC Chapel Hill
awang13@cs.unc.edu

Luchao Qi
UNC Chapel Hill
lqi@cs.unc.edu

Roni Sengupta
UNC Chapel Hill
ronisen@cs.unc.edu

Abstract

We introduce a novel continual learning problem: how to sequentially update the weights of a personalized 2D and 3D generative face model as new batches of photos in different appearances, styles, poses, and lighting are captured regularly. We observe that naive sequential fine-tuning of the model leads to catastrophic forgetting of past representations of the individual’s face. We then demonstrate that a simple random sampling-based experience replay method is effective at mitigating catastrophic forgetting when a relatively large number of images can be stored and replayed. However, for long-term deployment of these models with relatively smaller storage, this simple random sampling-based replay technique also forgets past representations. Thus, we introduce a novel experience replay algorithm that combines random sampling with StyleGAN’s latent space to represent the buffer as an optimal convex hull. We observe that our proposed convex hull-based experience replay is more effective in preventing forgetting than a random sampling baseline and the lower bound. We introduce continual learning datasets for five celebrities, along with the evaluation framework, metrics, and visualizations to examine this problem. See our [project page](#) for more details.

1. Introduction

Generative face models [6, 16–20, 26, 27, 40] have proven highly effective in learning global facial priors, enabling applications such as 3D face reconstruction, novel appearance synthesis, and attribute editing. While these models can generate realistic faces, they often fail to preserve identity when reconstructing, synthesizing, or editing images of specific individuals. Recent efforts have focused on improving inversion techniques [3, 14, 43] and developing personalized priors [28, 33] to address this challenge. Such personalized 2D [28] and 3D [33] models allow for an identity-preserving synthesis of novel appearances and semantic attribute editing, going beyond general inversion [38].

Personalization typically requires ~ 100 images per individual [28], encompassing diverse poses, styles, and lighting conditions to learn robust facial priors. However, accumulating such diverse photo collections can often take a significant amount of time for most users. Rather, users often capture images of themselves at the same place and time with limited stylistic variations, which also change over time due to different locations of capture or different styles of the user. Naively personalizing a generative model every time the user captures new images of themselves will lead to overfitting on a specific style and cause the model to forget previously seen facial variations resulting in poor generalization to new incoming data.

Continual learning has emerged as a promising solution to this challenge. A common technique in this field is experience replay, which involves storing the most informative past image samples in a replay buffer and integrating them with newly available image sets during model training. This method effectively reduces *catastrophic forgetting* [7, 23, 34], where the model loses previously learned knowledge due to overfitting on new data batches. Several experience replay algorithms [7, 23, 34] have been developed to select the most informative past samples. However, previous work has mainly concentrated on the task-incremental learning of conditional generative models [2, 41, 47, 49], where the generator adapts to represent different classes, or on domain-incremental learning for image classification models [22, 34, 46], a classifier learns to incrementally identify different classes. In this paper, to the best of our knowledge, we are the first to explore domain-incremental continual learning for unconditional generative models, specifically for personalized face generation.

We formulate a novel continual learning task, where at each timestamp the model is presented with a set of images of a similar style, extracted from a video recording, but the styles vary over time. We mainly explore the replay-buffer strategy for continual learning, where the model is fine-tuned at each timestamp on the newly captured image sets and the replay buffer. We propose two experience replay algorithms ('ER-Rand' and 'ER-Hull') to enhance

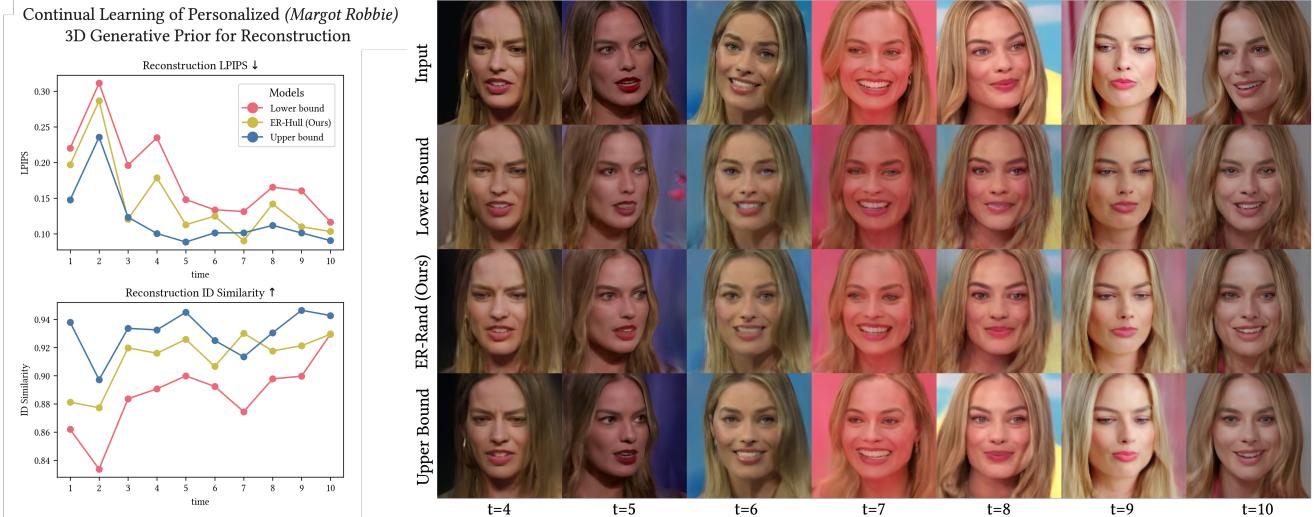


Figure 1. Open-world deployment and training of personalized generative models [28, 33] are challenging as images captured at each time has limited diversity in style, appearance, and lighting conditions. Naive finetuning of the model on each timestamp (Lower bound) leads to catastrophic forgetting, where the final model ($t=10$) performs poorly on test samples from previous timestamps. The Upper bound model is finetuned using all training images from all timestamps. We present an Experience Replay-based Continual Learning technique where we update a replay buffer to store the most informative images from the past as new images are captured at every timestamp. Our proposed technique, ER-Hull optimizes the buffer (size=3) as the most informative convex hull in StyleGAN’s latent space.

the preservation of the most informative samples in the replay buffer. We provide detailed quantitative and qualitative evaluations of personalized 2D [28] and 3D [33] generative models for reconstruction and synthesis tasks, with the following observations: (1) Compared to 2D, the problem of catastrophic forgetting is more pronounced for 3D generative models. (2) For a large replay buffer size to timestamps ratio of 100%, both random ('ER-Rand') and convex-hull optimization ('ER-Hull') based experience replay performs equally well and can match the upper bound. (3) For a small replay buffer size to timestamp ratio of 30%, 'ER-Hull' outperforms 'ER-Rand' by reducing forgetting of past images by $\sim 20\%$ for both 2D and 3D inversion, demonstrating its effectiveness in choosing the most informative samples to reduce forgetting.

We believe that for most practical applications a smaller buffer size compared to the number of timestamps is more storage efficient. For example, a daily update of the personalized generative model of a user for 5 years amounts to 1825 timestamps, which with 100% buffer size to timestamp ratio will lead to $\sim 9.1\text{GB}$ of storage, compared to $\sim 2.7\text{GB}$ for 30% buffer size to timestamp ratio (considering conservatively that each photo takes 5MB memory). However, due to limited computational resources, we limit the timestamp to $t = 10$ for most of our experiments.

Our key contributions include:

1. A new framework for continual learning of personalized face models, enabling open-world deployment

and training.

2. Two experience replay strategies: 'ER-Rand' and 'ER-Hull', with the latter optimizing a convex hull in StyleGAN’s latent space for improved performance for a smaller buffer-size to timestamp ratio.
3. A comprehensive evaluation framework using a diverse dataset of five individuals across 10 timestamps, demonstrating the effectiveness of our approach in mitigating catastrophic forgetting.

2. Related Work

Few-Shot Personalized Generative Prior Generative face models have significantly enhanced the realism of 2D facial generation [8, 13, 19, 20] and facilitated the creation of 3D facial models [6, 26, 27, 30, 40]. These advancements enable pretrained GANs to produce generalized image prior, which proves useful for tasks such as image enhancement [29, 31, 44] and semantic editing [29, 31, 44], where images are manipulated in the GAN’s latent space. However, these methods typically generate a prior across random faces, often leading to identity loss when images are projected into the latent space

Recent works, such as MyStyle [28], have tackled this issue by fine-tuning the generator using images of a single individual, employing an approach inspired by Pivotal Tuning [38] to establish a personalized prior. Similarly, for 3D generation, My3DGen [33] introduces a few-shot framework that applies techniques from MyStyle [28] to learn a per-

sonalized 3D prior from EG3D [6]. However, such methods operate in an offline setting, assuming that a comprehensive set of images of a person, captured across diverse poses, lighting conditions, styles, and environments, is available during training and does not require updates once training is done. This assumption may not hold in practical scenarios, where personal images typically arrive in small, consistent batches but vary over time. To address this limitation, we extend these methods to an online setting, continuously updating the model while mitigating catastrophic forgetting.

Continual Learning and Experience Replay Continual learning [37] involves sequentially receiving tasks or data, acquiring new knowledge while retaining previously learned information. For a detailed overview of continual learning, readers are encouraged to consult [48].

A major challenge in this domain is catastrophic forgetting [10, 11], where a model’s knowledge of earlier data deteriorates when trained on new information. Previous works [7, 23, 34] address forgetting through replay experience, where previously encountered training samples are stored in a replay buffer and used alongside new data during training. Various follow-up studies have proposed enhancements to experience replay through complex gradient manipulations [1, 4, 15, 23, 36, 45]. GDumb [32] presents a method for greedily storing samples in memory and subsequently training a model from scratch using all available samples. This approach demonstrates superior performance compared to previously proposed algorithms [15, 23, 36] in their respective experimental setups. Additionally, Buzzega et al. [5] show that simple modifications to traditional rehearsals techniques can achieve performance comparable to more sophisticated methods. These aforementioned works raise concerns about the commonly accepted assumptions, evaluation metrics, and the efficacy of various recently proposed algorithms for continual learning [32], particularly in generative models [2, 39, 41, 47, 49]. Furthermore, there is a notable gap in the literature regarding the application of these concepts to 3D GANs and personalized 3D models, especially within the context of domain-incremental continual learning. Therefore, we propose a new problem formulation targetting this application and introduce some simple experience-replay methods to solve this new problem.

3. Problem Setup and Methods

3.1. Motivation

With the advent of generative face modeling, the problem of personalizing pretrained generative models for a particular person in a practical few-shot nature has arisen as well. MyStyle [28] and My3DGen [33] tackle this problem in 2D and 3D respectively by tuning the pretrained model on a small region of the latent space and learning a personalized manifold that forms a strong personalized prior for

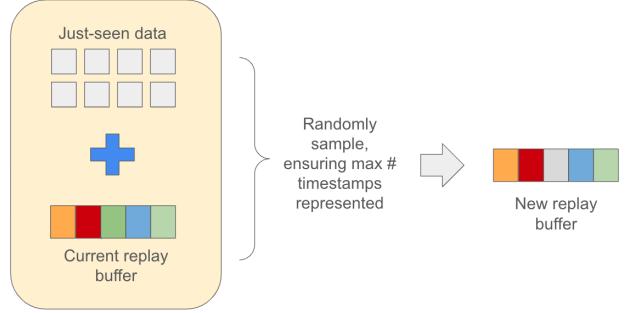


Figure 2. Diagram of the ER-Rand algorithm. We randomly sample the next replay buffer from the combination of the just-seen timestamp’s data plus the replay buffer, and we only consider samples where we have the maximum possible number of timestamps represented.

the generator. Although their training dataset size requirements are much lower than usual generative learning methods, they do require 50-200 images captured under diverse poses, appearances, styles, and lighting conditions. These methods face the problem that real-life data does not come all at once encompassing all possible variations in style, pose, and expression. Rather images of an individual often come in batches captured over time, where each batch of images has little diversity in appearance, style, and lighting since they are captured at the same place and time. Naively fine-tuning the generative model on the newly appearing image batches leads to catastrophic forgetting of previous appearances. Also, storing all image batches over time and re-training the personalized generative model on all of them at every timestamp is highly impractical. In this paper, we examine this realistic framework of continual learning for personalizing generative models and propose two experience replay strategies to alleviate catastrophic forgetting.

3.2. Problem Formulation

We consider a scenario in which we receive T batches of data sequentially of a specific individual, indexed by the timestamp $t = 1, 2, \dots, T$ when it was captured. Each batch contains images of the individual captured with a similar appearance, style and lighting at same place and time, e.g. captured during a zoom call in a particular environment and day. These T batches are captured at different timestamps in different environments, with varying appearance, style, and lighting across batches. Each batch is denoted as $X_t = \{x_t^1, x_t^2, \dots, x_t^n\}$, consisting of n training images x_t^i following \mathcal{D}_t , the data distribution for images from timestamp t . Our goal is to train the generative model sequentially on each data distribution such that we can best represent the combined data distribution across all timestamps.

We focus on experience replay, maintaining a small

fixed-size replay buffer that stores the most informative samples from the previously seen data. Let R_t denote the replay buffer at time t that stores the k most informative past images from $t = 1, \dots, t-1$ (We define $R_1 = \emptyset$ for convenience). The goal of experience replay algorithms is to use the current batch of images X_t to update the replay buffer R_{t-1} to create R_t and then train the generative model on $X_t \cup R_t$.

3.3. Training Protocol

To personalize a generator $G(\cdot; \theta)$ parameterized by θ in a continual learning fashion, we start from pretrained weights θ_0 at the beginning of training. We continuously update the weights with every incoming batch of data to obtain θ_t after training on X_t and the replay buffer R_t . For every timestamp t , we follow MyStyle [28] and My3DGen [33] and first invert images $x_t^i \sim \mathcal{D}_t$ into the latent space of the pretrained generator $G(\cdot; \theta_0)$, obtaining latent anchors $w_t^i \forall i = 1, \dots, n$. Inversion is performed using a pretrained encoder [35]. We then optimize the network weights θ , initialized at θ_{t-1} to obtain the updated weights θ_t . Optimization is performed by minimizing the reconstruction objective

$$\begin{aligned} \mathcal{L}_{rec}(G(\cdot; \theta), x_t^i, w_t^i) &= \mathcal{L}_{lpips}(G(w_t^i; \theta), x_t^i) \\ &\quad + \lambda_{L_2} \|G(w_t^i; \theta) - x_t^i\|_2 \end{aligned} \quad (1)$$

across both the incoming batch of images X_t and the replay buffer R_t . Formally, our loss across both sets is

$$\mathcal{L}_{rec}^t = \mathbb{E}_i[\mathcal{L}_{rec}(G, x_t^i, w_t^i)] + \lambda_R \mathbb{E}_j[\mathcal{L}_{rec}(G, x_t^j, w_t^j)] \quad (2)$$

where $x_t^j \in R_t \forall j = 1, \dots, k$ and λ_R is a hyperparameter.

Next, we propose two basic sampling strategies to populate the replay buffer R_t at every timestamp t .

3.4. ER-Random and ER-Hull

ER-Random. Our first algorithm is a modified version of balanced reservoir sampling, shown in Fig. 2. When possible, we randomly choose a combination of currently available images from $X_t \cup R_t$ satisfying the constraint that every previously seen batch of data is represented with at least one example. However, when $t > k$, i.e. the replay buffer is not large enough to satisfy this constraint, we randomly decide whether to replace a randomly chosen example from the buffer R_t with a new example from X_t .

This extremely simple method works well when the buffer size k is large relative to the number of observed timestamps T , but its performance is not always optimal when the buffer size is very small compared to T (for example, $T = 10, k = 3$). Thus, we propose a more tailored algorithm to further improve the quality of the replay buffer such that it can best capture all previously seen examples.

ER-Hull. Our next sampling algorithm ER-Random (Fig. 3) is based on the intuition from generative face models [20, 28] that the latent space of StyleGAN is disentangled

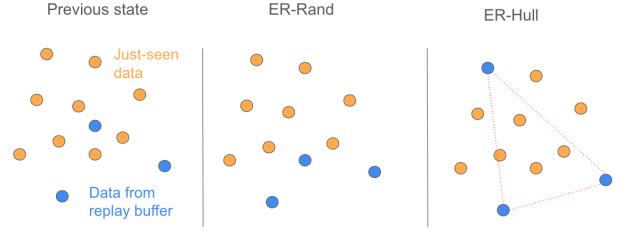


Figure 3. Illustration of ER-Hull. We perform RANSAC over many different possible replay buffers and choose the one that creates a convex hull that is "closest" to the other points, normalizing by timestamp.

and the convex hull of the anchor latent codes provides a well-behaved identity-preserving prior. We hypothesize that the best image to discard from the replay buffer is the one whose latent code is closest to the convex hull of the remaining latent anchors of the replay buffer. This means the current replay buffer will be able to best preserve the information of the discarded images.

Given the data batch X_t and the current replay buffer R_{t-1} , our goal is to choose R_t such that the average of the distance from each currently available data batch X_t^j to the convex hull $Hull(R_t)$ of the latent codes for the images in R_t is minimized. We define the distance as:

$$d(X_j, Hull(R_t)) = \sum_{i=1}^n \left(\mathbb{1}_{[x_j^i \in X_t \cup R_{t-1}]} \cdot \min_{p \in Hull(R_t)} \|p, x_j^i\|_2 \right), \quad (3)$$

where the indicator function in the summation ensures that we only calculate the distance for samples that we have available in X_t and R_t . Moreover, similar to ER-Random we additionally constrain the algorithm to only consider possible replay buffers that contain at least one sample from each previously seen batch of data when $k < t$ and no more than one example from each batch when $k \geq t$.

Brute-force searching over all such possible combinations of examples can be costly, so we modify the algorithm to use a RANSAC where we randomly sample at most N different combinations satisfying our constraints and find the optimal replay buffer from these N different options. Letting $S = \{R_t^{(\ell)} \forall \ell \in 1, 2, \dots, N\}$ denote our RANSAC-sampled set of possible buffers, we formally define our final selected replay buffer as

$$R_t^{final} = \arg \min_{\ell \in \{1, 2, \dots, N\}} \frac{\sum_{j=1}^t d(X_j, Hull(R_t))}{1 + \sum_{j=1}^{t-1} \mathbb{1}_{[X_j \cap R_{t-1} \neq \emptyset]}} \quad (4)$$

where the denominator is the number of all unique timestamps present in $X_t \cup R_{t-1}$.

Table 1. Continual Learning performance of personalized StyleGAN (MyStyle) in inverting an unseen test image, evaluated with Average Incremental Performance (AIP) measured with LPIPS (lower is better) and ID similarity (higher is better) as well as Forgetting of both metrics (lower is better), scaled by $\times 10$. ER-Rand and ER-Hull perform experience replay with simple random sampling and proposed convex hull optimization in StyleGAN latent space respectively.

| Buffer Size | Algorithm | Average over 5 celebrities | | | |
|-------------|-----------|----------------------------|-------------|-------------|-------------|
| | | AIP | | Forg. | |
| | | LPIPS | ID | LPIPS | ID |
| — | Lower | 1.17 | 8.97 | 0.64 | 1.09 |
| 3 | KMeans-3 | 1.07 | 9.25 | 0.44 | 0.68 |
| | ER-Rand | 1.04 | 9.28 | 0.40 | 0.55 |
| | ER-Hull | 0.99 | 9.34 | 0.30 | 0.47 |
| 5 | KMeans-5 | 1.00 | 9.36 | 0.29 | 0.40 |
| | ER-Rand | 0.98 | 9.38 | 0.25 | 0.39 |
| | ER-Hull | 0.98 | 9.39 | 0.25 | 0.33 |
| 10 | KMeans-10 | 0.92 | 9.44 | 0.18 | 0.26 |
| | ER-Rand | 0.91 | 9.46 | 0.15 | 0.23 |
| | ER-Hull | 0.91 | 9.45 | 0.16 | 0.22 |
| — | Upper | 0.89 | 9.45 | — | — |

3.5. Evaluation Metrics

We evaluate the performance of our method on both the reconstruction of test images as well as the synthesis of new images from the personalized prior. Let us define these tasks as follows:

Reconstruction We maintain a held-out test set for each cluster, X_t^{test} , whose test examples are drawn from \mathcal{D}_t . We evaluate how faithful our personalized prior is through the commonly-used projection-based approach [24, 25, 28, 33] of finding the best latent code in the personalized latent space that reconstructs the test image. This is done by freezing the generator and optimizing over the $W+$ latent space. We evaluate the projected images using LPIPS [50], DIST [9], PSNR, and ID score [28].

Synthesis Following MyStyle [28] and My3DGen [33], synthesis is conducted for each time t by sampling a latent code from the convex hull of X_t . For 3D synthesis, a random pose is selected as well to evaluate multiview synthesis quality. The quality and identity preservation of the synthesized image is measured through FID score [12] and ID score [28, 33]. Here, we take the maximum ID score between the synthesized image and any test image X_t^{test} .

For each task, we follow the standard practices of existing literature [7, 23, 48] of evaluating overall performance as well as forgetting. We adopt the following measures for

Table 2. Continual Learning performance of personalized StyleGAN (MyStyle) in synthesizing novel appearance, evaluated with Average Incremental Performance (AIP) measured with FID (lower is better) and ID similarity (higher is better) as well as Forgetting of both metrics (lower is better), scaled by $\times 10$. ER-Rand and ER-Hull perform experience replay with simple random sampling and proposed convex hull optimization in StyleGAN latent space respectively.

| Buffer Size | Algorithm | Average over 5 celebrities | | | |
|-------------|-----------|----------------------------|-------------|-------------|-------------|
| | | AIP | | Forg. | |
| | | FID | ID | FID | ID |
| — | Lower | 124.7 | 7.48 | 96.7 | 2.64 |
| 3 | KMeans-3 | 103.8 | 8.14 | 71.4 | 1.93 |
| | ER-Rand | 97.6 | 8.24 | 61.3 | 1.62 |
| | ER-Hull | 91.5 | 8.40 | 49.3 | 1.36 |
| 5 | KMeans-5 | 90.3 | 8.45 | 50.6 | 1.26 |
| | ER-Rand | 87.8 | 8.52 | 46.1 | 1.12 |
| | ER-Hull | 86.6 | 8.55 | 41.7 | 1.05 |
| 10 | KMeans-10 | 81.1 | 8.68 | 33.4 | 0.83 |
| | ER-Rand | 79.2 | 8.73 | 25.2 | 0.64 |
| | ER-Hull | 79.8 | 8.71 | 29.2 | 0.71 |
| — | Upper | 82.1 | 8.76 | — | — |

each aforementioned metric:

Average Incremental Performance (AIP) Let $a_{i,j}$ denote the performance of the model trained up until and including time i , evaluated for time j where $j \leq i$. That is, for reconstruction we evaluate on the test set X_j^{test} and for synthesis we use the convex hull X_j to sample new latent codes. Then the average performance at time t is $A_t = \frac{1}{t} \sum_{k=0}^t a_{t,k}$. To further measure the overall historical performance of the model across time, we take the average incremental performance $AIP = \frac{1}{T} \sum_{j=1}^T A_j$.

Forgetting Rather than measuring overall performance, this metric measures the memory stability of generalization to previous data distributions. Let f_j^i denote the forgetting on data cluster j after the model is trained on data cluster i :

$$f_j^i = \begin{cases} \max_{l \in \{0, 1, \dots, i-1\}} a_{l,j} - a_{i,j} & \text{if positive metric} \\ \max_{l \in \{0, 1, \dots, i-1\}} a_{i,j} - a_{l,j} & \text{if negative metric} \end{cases} \quad (5)$$

Then, we define the average forgetting to be $F = \frac{1}{T-1} \sum_{j=1}^{T-1} f_j^T$. Intuitively, forgetting measures the memory stability of the network while the AIP measures overall performance over time.

Table 3. Continual Learning performance of personalized EG3D (My3DGen) in reconstructing an unseen test image, evaluated with Average Incremental Performance (AIP) measured with LPIPS (lower is better) and ID similarity (higher is better) as well as Forgetting of both metrics (lower is better), scaled by $\times 10$. ER-Rand and ER-Hull perform experience replay with simple random sampling and proposed convex hull optimization in StyleGAN latent space respectively. Buffer size is 3.

| | Margot | | | | Harry | | | | IU | | | | Michael | | | | Average | | | |
|---------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| | AIP | | Forg. | | AIP | | Forg. | | AIP | | Forg. | | AIP | | Forg. | | AIP | | Forg. | |
| | LPIPS | ID | LPIPS | ID | LPIPS | ID | LPIPS | ID | LPIPS | ID | LPIPS | ID | LPIPS | ID | LPIPS | ID | LPIPS | ID | LPIPS | ID |
| Lower | 1.67 | 8.96 | 0.89 | 0.56 | 1.86 | 8.90 | 0.99 | 0.87 | 2.00 | 8.50 | 1.03 | 0.88 | 1.82 | 8.93 | 0.88 | 0.46 | 1.84 | 8.82 | 0.95 | 0.69 |
| ER-Rand | 1.39 | 9.23 | 0.62 | 0.39 | 1.54 | 9.14 | 1.43 | 0.94 | 1.78 | 8.76 | 0.67 | 0.62 | 1.64 | 9.17 | 1.22 | 0.39 | 1.59 | 9.08 | 0.98 | 0.59 |
| ER-Hull | 1.39 | 9.25 | 0.54 | 0.32 | 1.47 | 9.18 | 1.19 | 0.73 | 1.73 | 8.8 | 0.65 | 0.59 | 1.55 | 9.22 | 0.73 | 0.43 | 1.54 | 9.11 | 0.78 | 0.52 |
| Upper | 1.20 | 9.30 | — | — | 1.28 | 9.37 | — | — | 1.35 | 9.09 | — | — | 0.89 | 9.56 | — | — | 1.18 | 9.33 | — | — |

Table 4. Continual Learning performance of personalized EG3D (My3DGen) in synthesizing novel appearance, evaluated with Average Incremental Performance (AIP) measured with FID (lower is better) and ID similarity (higher is better) as well as Forgetting of both metrics (lower is better), scaled by $\times 10$. ER-Rand and ER-Hull perform experience replay with simple random sampling and proposed convex hull optimization in StyleGAN latent space respectively. Buffer size is 3.

| | Margot | | | | Harry | | | | IU | | | | Michael | | | | Average | | | |
|---------|--------------|-------------|-------------|-------------|--------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|--------------|-------------|-------------|-------------|--------------|-------------|-------------|-------------|
| | AIP | | Forg. | | AIP | | Forg. | | AIP | | Forg. | | AIP | | Forg. | | AIP | | Forg. | |
| | FID | ID | FID | ID | FID | ID | FID | ID | FID | ID | FID | ID | FID | ID | FID | ID | FID | ID | FID | ID |
| Lower | 122.1 | 5.40 | 38.0 | 0.67 | 217.0 | 4.42 | 108.3 | 0.97 | 154.1 | 4.89 | 138.2 | 1.62 | 193.2 | 4.98 | 90.1 | 1.07 | 171.6 | 4.92 | 93.6 | 1.08 |
| ER-Rand | 107.6 | 5.59 | 30.5 | 0.56 | 192.2 | 4.88 | 99.1 | 0.73 | 106.6 | 5.45 | 63.0 | 0.78 | 161.2 | 5.48 | 59.3 | 0.65 | 141.9 | 5.35 | 63.0 | 0.68 |
| ER-Hull | 106.2 | 5.55 | 31.2 | 0.52 | 183.1 | 4.99 | 84.6 | 0.60 | 97.6 | 5.47 | 50.6 | 0.91 | 153.6 | 5.48 | 53.8 | 0.60 | 135.1 | 5.37 | 55.0 | 0.66 |
| Upper | 68.6 | 5.83 | — | — | 124.7 | 5.42 | — | — | 58.0 | 5.74 | — | — | 93.2 | 6.11 | — | — | 86.1 | 5.78 | — | — |

4. Experiments

4.1. Experimental Setup

Data We introduce new personal face datasets of 5 celebrities (*Margot Robbie*, *Harry Styles*, *Sundar Pichai*, *Michael B. Jordan*, *IU*), each consisting of 10 batches of data (timestamps). Each batch contains 20 training images and 10 test images. Each batch contains images from video frames crawled from a single online video of the celebrity. These videos include interview videos and other short-form content uploaded online. The videos were chosen to have relatively consistent style and environment throughout, so that images from the same batch are captured with similar lighting and appearance but with variations in pose and expression, to model the real-world use case where an individual takes multiple photos of themselves in the same environment. The raw frames were automatically aligned and cropped [21] to size 512×512 , and filtered to only include faces of the specified identity [42]. We note that due to the inherent motion in videos and the large amount of cropping, some of the images are fairly low resolution and are of lower quality than *FFHQ* portrait images.

Fair use disclaimer. Our dataset is collected from YouTube videos and may contain copyrighted material. Such material is made available for research purposes only. This constitutes ‘fair use’ under Section 107 of the Copyright Act of 1976. All rights and credit go directly to its rightful owners. No copyright infringement is intended.

Training and Testing Details For the lower bound and both ER models, we train MyStyle for 1000 iterations (1 hour on 1 RTX A4500) and My3DGen for 500 iteration (5 hours on 4 RTX A6000 GPUs) for each timestamp. For the upper bound, we train MyStyle for 1000 iterations and My3DGen for 2000 iterations (because it is more difficult to converge than MyStyle). For our experience replay algorithms, we set $\lambda_R = 1$ and modify the training of MyStyle and My3DGen accordingly so that 50% of the training images at every iteration are from the replay buffer. For sampling, ER-Rand takes ~ 3 seconds while ER-Hull takes 40 minutes per timestamp for $N = 5000$ ransac iterations. We test replay buffer sizes of 3, 5, and 10 for 2D and buffer size 3 for 3D due to the higher compute requirement in 3D. We evaluate our results using the tasks described in section 3.6. Note that we did not use Pivotal Tuning [38] in our recon-

struction evaluation to focus on understanding the memory and generalization power of the model without per-image tuning. We also compare our methods ER-Rand and ER-Hull to the baseline experience replay method of K-Means clustering [7] on the 2D tasks, which both have computational requirements similar to that of ER-Rand, taking <5 seconds to sample per identity.

Computational complexity discussion We note that the training process for this experiment set up is very expensive, as for 2D, each timestamp requires 1 hour to train and 10 timestamps requires 10 hours to train. For 3D, each timestamp requires 5 hours to train and 10 timestamps requires 50 hours to train. For the actual experience replay sampling, ER-Rand takes 3 seconds for each timestamp and ER-Hull takes 40 minutes (for all buffer sizes). However, compared to the extensive training time especially for 3D experiments, the additional time required for ER-Hull is not so significant. For any given buffer size, the storage cost of both algorithms is $O(b)$ where b is the buffer size.

4.2. Evaluation on 2D Generative Models

We first investigate the role of continual learning in open-world deployment of the personalized 2D generative model, MyStyle [28]. We mainly focus on evaluating the inversion and synthesis capability of the MyStyle model at each timestamp on heldout test images of that and all previous timestamps. For inversion, we calculate LPIPS and ID similarity metrics. For synthesis we random sample latent codes in the convex hull defined by the buffer and the current batch of images and report FID and ID similarity metrics.

In Table 1 we report the Average Incremental Performance (AIP) and Forgetting for inversion tasks with LPIPS and ID similarity metric, averaged over all 5 celebrity datasets. We observe that for smaller buffer size of 3, ER-Hull improves AIP by 5% (0.99 vs 1.04) and Forgetting by 25% (0.3 vs 0.4) over ER-Rand in terms of LPIPS metric. For larger buffer size of 10, both ER-Rand and ER-Hull perform almost equal to the ideal Upper bound performance.

Similarly, in Table 5 we report AIP and Forgetting for synthesis tasks with FID and ID similarity metric, averaged over all 5 celebrity datasets. We observe that for smaller buffer size of 3, ER-Hull improves AIP by 6.3% (91.5 vs 97.6) and Forgetting by 19.6% (49.3 vs 61.3) over ER-Rand in terms of FID metric. For larger buffer size of 10, ER-Hull is slightly worse than ER-Rand and only slightly worse than the ideal Upper Bound performance.

In addition, in Fig 4a we visualize the performance deterioration of the final model trained at timestamp 10 on all previous timestamps, averaged over all 5 celebrity datasets, for inversion with LPIPS and ID similarity and for synthesis with FID and ID similarity metrics. We only show a buffer size of 3 since that is the most challenging and ideal for long-term deployment.

Analysis. In summary, we observe that a larger buffer size (10) with a 1:1 ratio with the number of timestamps makes the continual learning method easier and any reasonable experience replay-based technique can perform well and closely match the ideal Upper Bound performance. In many real-life scenarios for long-term deployment, timestamps can be in ~ 100 s or ~ 1000 s, where a 1:1 ratio between buffer size and timestamp can be extremely prohibitive. For example, storing 1000 images with JPEG compression in a buffer can lead to $1000 \times 5 \text{ MB} = 5 \text{ GB}$ of memory. This is where a smaller buffer size, e.g. in a 3:10 ratio, is more practical and presents a more challenging scenario to study continual learning. With a smaller buffer size, ER-Hull can significantly reduce forgetting compared to a simple algorithm like ER-Rand.

Even though in our experiments all of the buffer sizes are trivially small due to computational limitations, the difference between a 30% buffer size and a 50% buffer size can be huge when the number of timesteps scales up. We performed another experiment (see Tab. 5) where we extended the number of timestamps to 20 and used a buffer size of 6 (in accordance with the 30% buffer size of ER-Rand-3 and ER-Hull-3) on a single identity (*Michael B. Jordan*) due to limited resources, and found that ER-Hull still demonstrated much less forgetting than ER-Rand. This suggests that our method can be especially useful when the number of timesteps scales up significantly.

Table 5. Continual Learning performance of personalized StyleGAN (MyStyle) in synthesizing novel appearance for *Michael B. Jordan*, evaluated with Forgetting metrics measured with FID (lower is better) and ID similarity (higher is better, scaled by $\times 10$) over 20 timestamps

| Buffer Size | Algorithm | Forgetting | |
|-------------|-----------|------------|-----|
| | | FID | ID |
| 6 | ER-Rand | 50.98 | 1.5 |
| | ER-Hull | 29.80 | 1.0 |

4.3. Evaluation on 3D Generative Models

Next, we follow a similar investigation for continual learning of personalized 3D generative model (EG3D), My3DGen [33], in reconstructing an unseen test image and synthesizing novel appearance, using the same metrics as in Sec 4.2. In Table 3 and Table 1 we report AIP and Forgetting for reconstruction and synthesis tasks. We mainly focus on a buffer size of 3 since we observe that to be the most challenging scenario with large practical significance for open-world deployment for a longer time. We observe that ER-Hull is slightly better than ER-Rand for all 4 celebrities for both reconstruction and synthesis. ER-Hull improves Forgetting over ER-Rand by 20.4% (LPIPS: 0.78

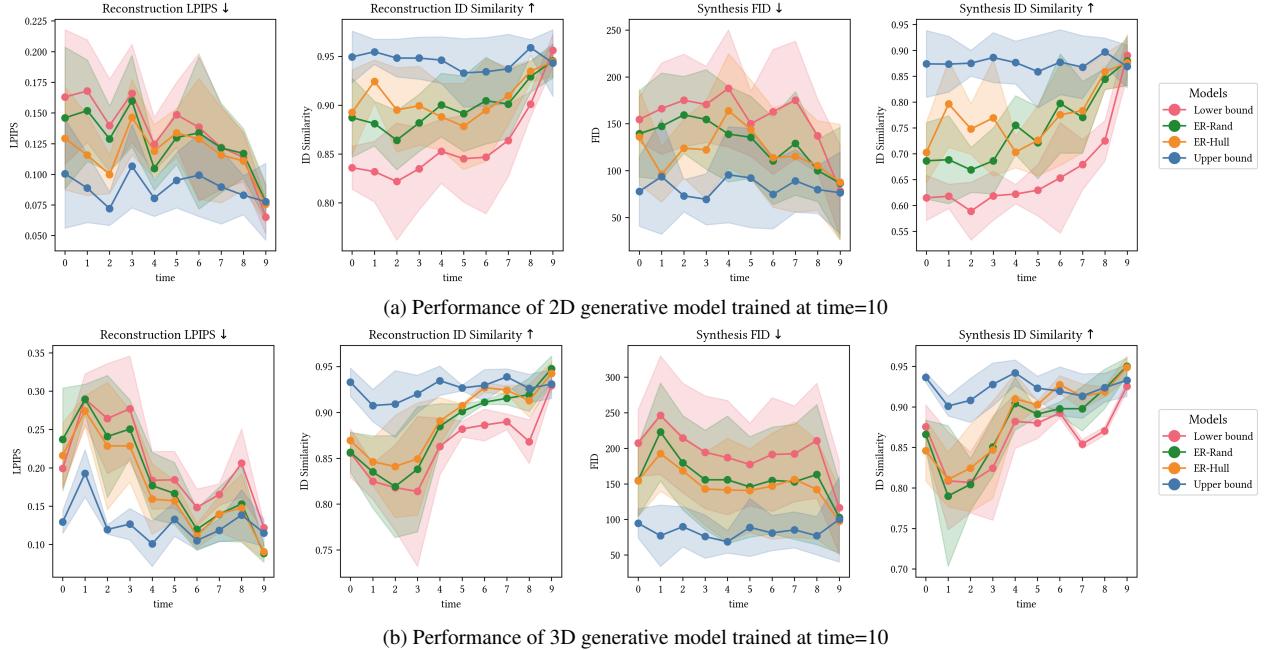


Figure 4. Performance deterioration of the final personalized (a) 2D generative model and (b) 3D generative model trained at $t=10$ on all previous time, averaged over 5 celebrities for 2D and 4 celebrities for 3D. ER-Hull outperforms ER-Rand on earlier timestamps proving its effectiveness in reducing forgetting.

vs 0.98) for reconstruction and by 12.7% (FID: 55 vs 63) for synthesis.

We also visualize the performance deterioration of the final model trained at timestamp 10 on all previous timestamps, averaged over all 4 celebrity datasets, for inversion with LPIPS and ID similarity and for synthesis with FID and ID similarity metrics in Fig. 4b. Qualitative evaluation of the final model for reconstruction and synthesis for ER-Rand and ER-Hull, compared to Lower and Upper bound performance is presented in the supplementary material.

Analysis. In summary, we observe that while ER-Hull is only slightly better than ER-Rand on average across all timestamps, it is significantly more effective in reducing forgetting, which is often the primary goal of a continual learning algorithm for open-world deployment. We also observe 3D tasks to be particularly more challenging than 2D tasks resulting in lower reconstruction and synthesis accuracy. This observation is supported by previous research on generative face models [6, 19, 28, 30, 33] where 3D generation was shown to be significantly more challenging than 2D and often requires larger data with more diversity for personalization.

5. Conclusion

Our work is the first to tackle the problem of open-world deployment of personalized generative models. We introduce a novel problem formulation, dataset, experimental

framework, evaluation metrics and visualizations to examine this problem. We introduce two experience replay-based continual learning techniques; a simple random sampling-based solution (ER-Rand) that works well for larger buffer sizes, and a more advanced one that optimizes convex hull maximization in StyleGAN latent space (ER-Hull) and works better for smaller buffer sizes. ER-Hull improves over the lower bound and closely matches the upper bound for large buffer sizes.

Limitations and Future Work. While ER-Hull is better than ER-Rand for small buffer sizes, the performance with respect to the upper bound is still poor, highlighting the need for future research. Additionally, we test our model on 10 timestamps with 20 images each due to limitations in resources. This problem can be extended to significantly more timestamps (~ 100) to closely match real-world use cases. Lastly, the quality of the generated images dropped due to reliance on publicly available TV interviews or content videos which are often limited by resolution, motion blur, and low SNR.

Ethical Considerations. Access to personalized generative models trained with continual learning has the potential to reveal training images of the individual and generate unintended manipulated images. Recent research on securing deployment of ML models and detecting deep fake images will help to mitigate this risk.

References

- [1] Victoria Fernández Abrevaya, Adnane Boukhayma, Stefanie Wuhrer, and Edmond Boyer. A decoupled 3d facial shape model by adversarial training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9419–9428, 2019. 3
- [2] Sergey Bartunov and Dmitry Vetrov. Few-shot generative modelling with generative matching networks. In *International Conference on Artificial Intelligence and Statistics*, pages 670–678. PMLR, 2018. 1, 3
- [3] Ananta R Bhattacharai, Matthias Nießner, and Artem Sevestopolsky. Triplanenet: An encoder for eg3d inversion. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3055–3065, 2024. 1
- [4] Zalán Borsos, Mojmir Mutny, and Andreas Krause. Coresets via bilevel optimization for continual learning and streaming. *Advances in neural information processing systems*, 33:14879–14890, 2020. 3
- [5] Pietro Buzzega, Matteo Boschini, Angelo Porrello, and Simone Calderara. Rethinking experience replay: a bag of tricks for continual learning. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 2180–2187. IEEE, 2021. 3
- [6] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16123–16133, 2022. 1, 2, 3, 8
- [7] Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, Puneet K Dokania, Philip HS Torr, and Marc’Aurelio Ranzato. On tiny episodic memories in continual learning. *arXiv preprint arXiv:1902.10486*, 2019. 1, 3, 5, 7
- [8] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 2
- [9] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE transactions on pattern analysis and machine intelligence*, 44(5):2567–2581, 2020. 5
- [10] Robert M French. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3(4):128–135, 1999. 3
- [11] Ian J Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. An empirical investigation of catastrophic forgetting in gradient-based neural networks. *arXiv preprint arXiv:1312.6211*, 2013. 3
- [12] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 5
- [13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2
- [14] Ali Jahanian, Lucy Chai, and Phillip Isola. On the “steerability” of generative adversarial networks. *arXiv preprint arXiv:1907.07171*, 2019. 1
- [15] Xisen Jin, Arka Sadhu, Junyi Du, and Xiang Ren. Gradient-based editing of memory examples for online task-free continual learning. In *Neural Information Processing Systems*, 2020. 3
- [16] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 1
- [17] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *Advances in neural information processing systems*, 33:12104–12114, 2020. 1
- [18] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in neural information processing systems*, 34:852–863, 2021. 1
- [19] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 1, 2, 8
- [20] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 1, 2, 4
- [21] Vahid Kazemi and Josephine Sullivan. One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1867–1874, 2014. 6
- [22] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017. 1
- [23] David Lopez-Paz and Marc’Aurelio Ranzato. Gradient episodic memory for continual learning. *Advances in neural information processing systems*, 30, 2017. 1, 3, 5
- [24] Xuan Luo, Xuaner Zhang, Paul Yoo, Ricardo Martin-Brualla, Jason Lawrence, and Steven M Seitz. Time-travel rephotography. *ACM Transactions on Graphics (TOG)*, 40(6):1–12, 2021. 5
- [25] Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 2437–2445, 2020. 5
- [26] Thu Nguyen-Phuoc, Chuan Li, Lucas Theis, Christian Richardt, and Yong-Liang Yang. Hologan: Unsupervised learning of 3d representations from natural images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7588–7597, 2019. 1, 2
- [27] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields.

- In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11453–11464, 2021. [1](#) [2](#)
- [28] Yotam Nitzan, Kfir Aberman, Qiurui He, Orly Liba, Michal Yarom, Yossi Gandelsman, Inbar Mosseri, Yael Pritch, and Daniel Cohen-Or. Mystyle: A personalized generative prior. *ACM Transactions on Graphics (TOG)*, 41(6):1–10, 2022. [1](#), [2](#), [3](#), [4](#), [5](#), [7](#), [8](#)
- [29] Yotam Nitzan, Amit Bermano, Yangyan Li, and Daniel Cohen-Or. Face identity disentanglement via latent space mapping. *arXiv preprint arXiv:2005.07728*, 2020. [2](#)
- [30] Roy Or-El, Xuan Luo, Mengyi Shan, Eli Shechtman, Jeong Joon Park, and Ira Kemelmacher-Shlizerman. Stylesdf: High-resolution 3d-consistent image and geometry generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13503–13513, 2022. [2](#), [8](#)
- [31] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2085–2094, 2021. [2](#)
- [32] Ameya Prabhu, Philip Torr, and Puneet Dokania. Gdumb: A simple approach that questions our progress in continual learning. In *The European Conference on Computer Vision (ECCV)*, August 2020. [3](#)
- [33] Luchao Qi, Jiaye Wu, Annie N. Wang, Shengze Wang, and Roni Sengupta. My3dgen: A scalable personalized 3d generative model, 2024. [1](#), [2](#), [3](#), [4](#), [5](#), [7](#), [8](#)
- [34] Sylvestre-Alvise Rebiffé, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2001–2010, 2017. [1](#), [3](#)
- [35] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2287–2296, 2021. [4](#)
- [36] Matthew Riemer, Ignacio Cases, Robert Ajemian, Miao Liu, Irina Rish, Yuhai Tu, and Gerald Tesauro. Learning to learn without forgetting by maximizing transfer and minimizing interference. *arXiv preprint arXiv:1810.11910*, 2018. [3](#)
- [37] Mark Bishop Ring. *Continual learning in reinforcement environments*. The University of Texas at Austin, 1994. [3](#)
- [38] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *ACM Transactions on graphics (TOG)*, 42(1):1–13, 2022. [1](#), [2](#), [6](#)
- [39] Michał Sadowski, Karol J Piczak, Przemysław Spurek, and Tomasz Trzcinski. Continual learning of 3d point cloud generators. In *Neural Information Processing: 28th International Conference, ICONIP 2021, Sanur, Bali, Indonesia, December 8–12, 2021, Proceedings, Part I 28*, pages 330–341. Springer, 2021. [3](#)
- [40] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. Graf: Generative radiance fields for 3d-aware image synthesis. *Advances in Neural Information Processing Systems*, 33:20154–20166, 2020. [1](#), [2](#)
- [41] Ari Seff, Alex Beatson, Daniel Suo, and Han Liu. Continual learning in generative adversarial nets. *arXiv preprint arXiv:1705.08395*, 2017. [1](#), [3](#)
- [42] Sefik Serengil and Alper Ozpinar. A benchmark of facial recognition pipelines and co-usability performances of modules. *Bilisim Teknolojileri Dergisi*, 17(2):95–107, 2024. [6](#)
- [43] Yujun Shen, Jinjin Gu, Xiaou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9243–9252, 2020. [1](#)
- [44] Yujun Shen, Ceyuan Yang, Xiaou Tang, and Bolei Zhou. Interfacegan: Interpreting the disentangled face representation learned by gans. *IEEE transactions on pattern analysis and machine intelligence*, 44(4):2004–2018, 2020. [2](#)
- [45] Dongsub Shim, Zheda Mai, Jihwan Jeong, Scott Sanner, Hyunwoo Kim, and Jongseong Jang. Online class-incremental continual learning with adversarial shapley value. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 9630–9638, 2021. [3](#)
- [46] Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. *Advances in neural information processing systems*, 30, 2017. [1](#)
- [47] Xiaoyu Tao, Xiaopeng Hong, Xinyuan Chang, Songlin Dong, Xing Wei, and Yihong Gong. Few-shot class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12183–12192, 2020. [1](#), [3](#)
- [48] Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: Theory, method and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. [3](#), [5](#)
- [49] Mengyao Zhai, Lei Chen, Frederick Tung, Jiawei He, Megha Nawhal, and Greg Mori. Lifelong gan: Continual learning for conditional image generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2759–2768, 2019. [1](#), [3](#)
- [50] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. [5](#)