frontiers in Imaging

TYPE Original Research
PUBLISHED 27 May 2026
DOI 10.3389/fimag.2026.1817515

OPEN ACCESS

EDITED BY
Christian Kraetzer,
Otto-von-Guericke University, Germany

REVIEWED BY
Nuno Goncalves,
University of Coimbra, Portugal
Marija Ivanovska,
University of Ljubljana, Slovenia

*CORRESPONDENCE
Guido Borghi
guido.borghi@unimore.it

RECEIVED 25 February 2026
REVISED 14 April 2026
ACCEPTED 20 April 2026
PUBLISHED 27 May 2026

CITATION
Pellegrini L, Borghi G, Franco A and Maltoni D (2026) Adaptive-LwF: continual training of morphing attack detector without forgetting. *Front. Imaging* 5:1817515. doi: 10.3389/fimag.2026.1817515

Adaptive-LwF: continual training of morphing attack detector without forgetting

Lorenzo Pellegrini[1], Guido Borghi[2]*, Annalisa Franco[1] and Davide Maltoni[1]

[1]Department of Computer Science and Engineering, University of Bologna, Cesena, Italy, [2]Department of Education and Humanities, University of Modena and Reggio Emilia, Reggio Emilia, Italy

In Biometrics, the presence of privacy restrictions on personal data transfer and storage poses significant challenges in creating a sufficiently comprehensive and varied dataset by leveraging various data sources for traditional batch-based training procedures. This is particularly true in the Morphing Attack Detection (MAD) task, in which data involves facial images and a limited number of public datasets of well-controlled images are available. In this context, MAD systems generally suffer from limited generalization capabilities, with low performance on new and unseen data. Therefore, in this paper, we propose Adaptive-LwF, adopting the recent paradigm of Continual Learning (CL) as a viable solution to enable incremental training across multiple sites. Indeed, CL assumes that once a model has been trained, previous data cannot be utilized in subsequent training iterations and can be deleted. In particular, we investigate the performance of different methods in this new scenario, where a model is updated each time a new chunk of data, of variable size, becomes available. We focus our attention on the well-known Learning without Forgetting (LwF) algorithm, proposing a novel adaptive approach able to automatically fine-tune its parameters in relation to the variable size of the specific input chunks. Experimental results confirm that our approach is capable of mitigating the catastrophic forgetting effects, and the superior performance of the Adaptive-LwF algorithm with respect to alternative solutions.

KEYWORDS
continual learning, face recognition, morphing attack, morphing attack detection, presentation attack

1 Introduction

Privacy issues severely limit the acquisition, storage, and sharing of data, especially when personal information, such as facial images or sex, age, and similar are included. In particular, this peculiar condition arises in the development of biometric applications, and specifically with Morphing Attack Detection (MAD) systems (Raja et al., 2020), i.e., key tools to contrast the Morphing Attack (Ferrara et al., 2014), a recent security threat through which a criminal can elude the automatic controls based on face verification. More precisely, the morphing attack consists of visually combining into a single facial image two different identities, belonging to a criminal and an accomplice (see Figure 1), making a legal document shareable by two subjects. In practice, an electronic Machine Readable Travel Document (eMRTD) could be used by a criminal to pass face verification-based systems (Scherhag et al., 2017; Ferrara and Franco, 2022), such as the Automated Board Control (ABC) systems in international airports.

Therefore, the development and deployment of effective MAD algorithms represent a critical imperative for both private and public institutions operating within the domains of national and international security. Nevertheless, compliance with privacy regulations frequently constrains research laboratories to develop MAD solutions relying exclusively on internally sourced data, which is often characterized by a limited quantity and diversity of morphed samples (Borghi et al., 2022). As a consequence, MAD methods, especially if based on data-hungry deep learning architectures, frequently present good performance on internal and private data, but at the same time exhibit limited accuracy on new unseen data, due to hampered generalization capabilities.

A partial solution is represented by the availability of public web-based platforms, i.e., NIST FRVT MORPH[1] and FVC-onGoing,[2] that offer the possibility to test the MAD systems on sequestered datasets, i.e., data never publicly released, creating challenging cross-dataset evaluations and making it clearer the real contribution of newly proposed approaches in the literature. If, on the one hand, these public benchmarks allow to objectively assess algorithms' performance in a strongly supervised approach, on the other hand, little work has been done so far in order to strengthen the model training procedure.

We believe that a suitable solution consists of incrementally training the same MAD model on multiple sites, each belonging, for instance, to a different research institution, through a privacy regulations-compliant approach not based on the transfer of any personal data. In this way, training data are increased in terms of amount and variety, whenever new data are collected or become available. Therefore, in this paper, we propose to leverage an incremental training strategy based on the Continual Learning (CL) (Parisi et al., 2019) paradigm. Our assumption is to enable lifelong learning MAD models, i.e., MAD systems that are able to continuously accumulate knowledge and learn during their deployment phase. In this manner, a MAD model is continuously trained on different sites consisting of one or more private datasets stored in a temporary or permanent manner and not shared with third parties. In practice, this is the case, for instance, in which new images are acquired at the ABC gates in airports and are available, only for a limited time, to increase the performance of the MAD model. Another example is the case in which a research laboratory collects a new private dataset that can be used to improve the performance of a global MAD model. A visual example of the proposed scenario is depicted in Figure 2.

Unfortunately, this incremental training strategy introduces challenges generally not present in the traditional Machine Learning scenario, in which the training phase is usually conducted once on the whole training dataset, before the deployment of the trained model, relying on the assumption that all training data are available before the training procedure and well represent the testing domain. In particular, the main issue is represented by the so-called catastrophic forgetting (McCloskey and Cohen, 1989), i.e., the tendency to abruptly forget the previous knowledge when the model is updated on new unseen data.

We introduce Adaptive-LwF, a solution based on the well-known Learn without Forgetting (LwF) (Li and Hoiem, 2017) CL method, to tackle the introduced continual MAD scenario, consisting of an algorithm able to automatically adjust the LwF parametrization in relation to the size of new training data provided in input. Experimental results show that our Adaptive-LwF method, through a proper value weight of the distillation loss (λ), is an enabling and suitable algorithm to incrementally train new MAD systems, achieving interesting accuracy, mitigating the catastrophic forgetting and, at the same time, respecting privacy issues.

1.1 Summary of contributions

Summarizing, the contributions are the following:

- **Adaptive-LwF**: we propose the use of the LwF algorithm for the MAD task, introducing a method to properly set its main parameter, specifically the distillation loss's weight (λ), in relation to the amount of the input data of the learning experience. Indeed, the value of λ is critical for LwF performance and, to the best of our knowledge, this is one of the first investigations about this element in the recent literature.
- **Incremental MAD training**: we investigate the incremental training MAD systems, enabling the update of the model whenever new data become available. This is a novel and interesting scenario to comply with privacy limitations and issues, often present with personal data (in our case, facial images used in official documents).
- **Extensive validation**: we conducted an extensive experimental validation, based on different morphed-based datasets organized in novel benchmarks, conceived to represent real-world training case studies in the MAD scenario. In addition, we introduce and discuss new metrics aiming to assess MAD model's performance on the proposed incremental training scenario.

This paper resumes and further extends the initial work described in Pellegrini et al. (2023).

2 Problem statement

2.1 Face morphing

Face morphing is an image processing technique that makes it possible to gradually transform one image into another with a seamless transition. In the context of face and electronic documents, this technique—originally described in Ferrara et al. (2014)—can be effectively used for the creation of human faces with a hybrid identity, i.e., a face that hosts a sort of double identity as shown in Figure 1. As previously mentioned, face

1 https://pages.nist.gov/frvt/html/frvt_morph.html

2 https://biolab.csr.unibo.it/fvcongoing

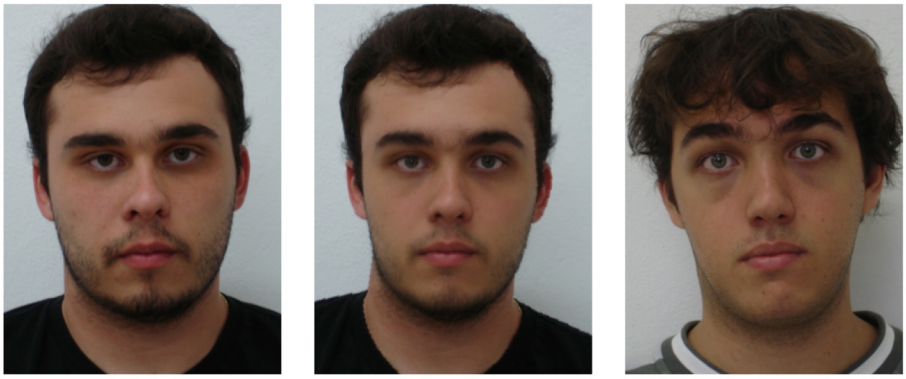

FIGURE 1
An example of a morphed face (central), created starting from two subjects, i.e., the accomplice (Subject 1)—a person without criminal records—and the criminal (Subject 2). Experimental studies (Scherhag et al., 2017; Ferrara and Franco, 2022) have revealed that the morphed identity can effectively deceive both human examiners and automatic face verification-based systems. **(a)** Subject 1. **(b)** Morphed. **(c)** Subject 2. Images 1a and 1b belong to the FEI dataset, while image 1b is from the FEI Morph dataset.

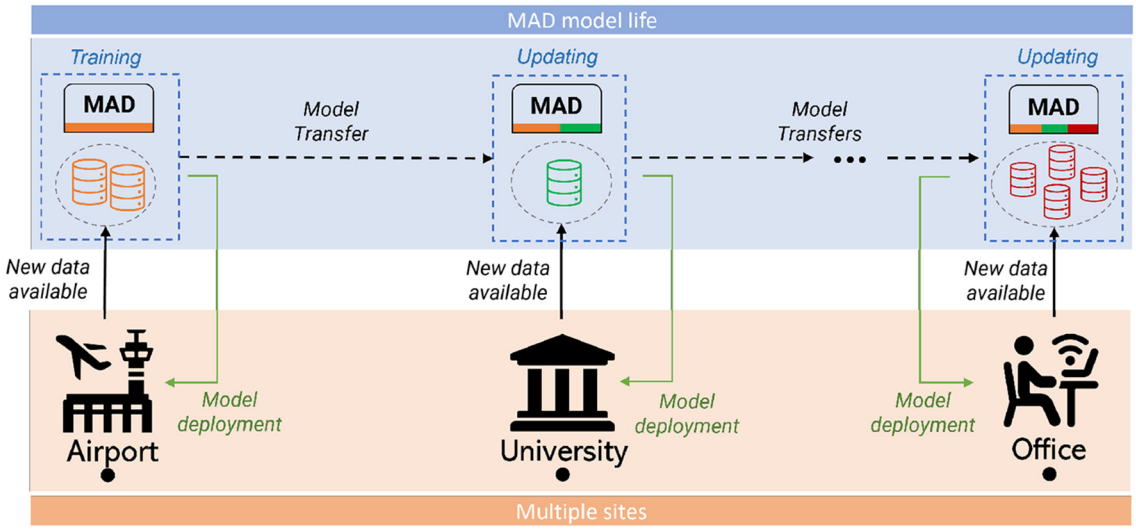

FIGURE 2
Visualization of the proposed continual training scenario investigated in this paper. Specifically, a MAD model is updated whenever a new portion of data becomes temporarily accessible on various sites. This scenario is privacy-compliant since it eliminates the need for data transfer. We propose the use of a Continual Learning algorithm, namely Adaptive-LwF, that is eligible to effectively leverage the available data for incremental model training.

morphing has recently arisen as a security threat, since it has been proven that a morphed face can fool police officers and automatic vision-based face verification systems (Scherhag et al., 2017; Raja et al., 2020). This condition is exacerbated by the fact that, alongside the traditional public methods of morphing based on facial landmarks, new ones based on GAN have emerged (Zhang et al., 2021; Venkatesh et al., 2020), thus simplifying the work of a potential criminal. Furthermore, the image quality can be further improved through manual or even automatic retouching that can remove visible or not visible artifacts (Borghi et al., 2021a; Di Domenico et al., 2024). Therefore, private and public institutions strongly need the development of new Morphing Attach Detector (MAD) systems, as automatic tools able to detect the presence of the morphing procedure in document images.

2.2 Morphing attack detection

Generally, two types of MAD methods are investigated in the literature: Single-image MAD (S-MAD) and Differential MAD (D-MAD). A D-MAD system receives as input two different images, i.e., the live acquisition and the photo, possibly morphed, stored in the eMRTD. Then, D-MAD systems operate on the assumption that one of the two inputs is acquired in a trusted manner, e.g., through the camera installed at ABC gate, or through an acquisition procedure monitored by a police officer. Conversely, S-MAD systems are based on single input images present in the documents. A third category of MAD methods has been recently proposed in the literature, and it is referred to as Video MAD (V-MAD) (Borghi et al., 2024), in which the detection systems receive as input the document image and a video, usually acquired by

face commercial recognition frameworks. In this manner, D-MAD and S-MAD detection techniques can be combined, in addition to aggregation or selection methods that can be used to identify the most useful frames in the video sequence.

Of the three possible approaches, in this paper, we focus on the D-MAD task.

Several D-MAD methods have been proposed in the literature. Broadly, these approaches can be categorized based on the types of features they extract from the two input images: artifact-based, identity-based, demorphing-based, and hybrid methods, commonly referred to as mixed feature-based approaches.

Artifact-based D-MAD methods (Scherhag et al., 2018; Chaudhary et al., 2021) focus on extracting general-purpose features from the input images using both traditional computer vision techniques and deep learning architectures. These methods rely on the hypothesis that feature comparison can reveal the subtle anomalies—i.e., artifacts—introduced during the morphing process. Commonly used hand-crafted features in this category include BSIF (Kannala and Rahtu, 2012), LBP (Ojala et al., 1994), and HOG (Dalal and Triggs, 2005), typically combined with machine learning classifiers. While these techniques have demonstrated some success, their performance remains suboptimal in certain scenarios. Other more effective methods are based on deep learning features and architectures (Medvedev et al., 2024; Shiqerukaj et al., 2022).

In contrast, identity-based D-MAD methods (Scherhag et al., 2020; Kessler et al., 2023) aim to extract features specifically tied to the identity of the subjects depicted in the input images. The underlying assumption is that a morphing attack can be detected by employing a neural network trained for the face recognition task to analyze discrepancies in identity-related features. In the demorphing-based category (Ferrara et al., 2017; Long et al., 2024), the primary objective is to reverse the morphing procedure to reconstruct the identity of the legitimate document holder. By comparing the morphed image with the restored version, it becomes possible to detect discrepancies that indicate the presence of a morphing attack. Finally, hybrid D-MAD methods (Borghi et al., 2021b) integrate both identity-related and artifact-based features. This combination aims to leverage the strengths of each feature type to enhance the detection of morphing attacks.

2.3 Incremental training in biometrics

Incremental training, or rather the adaptation of biometrics systems on new unseen data, has been investigated in the literature, but very few works are based on the recent paradigm of Continual Learning. From a general point of view, in biometrics systems, there is a need for updates when the biometric reference no longer represents the acquired biometric feature (Pisani et al., 2019). In particular, this happens when: (i) enrollment conditions are changed, due to environmental variations (e.g., changes in illumination) or the use of a different acquisition device. In this context, a limited number of samples available during the enrollment stage could also limit the coverage of all possible conditions present in future phases; (ii) physical and behavioral biometric modalities are influenced by the passage of time (*template aging*) due to, for instance, injuries, speckles, and wrinkles. In our scenario, we focus on the first point, i.e., on the possibility of incrementally updating our MAD model on newly available data since, as mentioned, privacy issues severely limit the availability of publicly released datasets. For example, the ability to leverage realistic daily acquired data of an ABC gate, which cannot be stored except than temporarily, can lead to the development of particularly robust MAD systems.

Unfortunately, several adaptive or incremental learning methods in the biometric literature rely on replaying the old data buffer for retraining (Kang et al., 2007; Pisani et al., 2019). We exclude from our analysis these methods due to privacy regulations for our case. Thus, we focus our attention on the Continual Learning paradigm since it has demonstrated good performance even without the use of stored data, usually referred to as replay memory. It should be noted that in the literature, there are similar approaches aimed at attempting to adapt already-trained models, even if only partially. For the sake of completeness, we briefly analyze them in the following.

Recently, Federated Learning (FL) (McMahan et al., 2017) has been proposed as an effective way to train a collaborative model without the need to centralize data. Indeed, the training procedure is conducted on several different clients with their own data, exchanging training information (e.g., the gradient or the weights) with a centralized server, which usually performs the role of a supervisor and organizer of that training procedure. We observe that this approach does not fully meet the requirements of our task based on several training procedures whenever new training data becomes available (Banabilah et al., 2022). Indeed, these clients, organized in an infrastructure that guarantees low latency, must be simultaneously available at the same time. Furthermore, a copy of the training data must be stored in case of a new update of the model, but this storage can be limited by privacy issues. However, several publicly released frameworks that handle the complexity of FL algorithms have been released, and despite the above-mentioned challenges, this seems to be a suitable solution for distributed learning in which each client is allowed to store its own data for a certain amount of time.

A possible paradigm to adapt a learning model is the Domain Adaptation (DA) (Patel et al., 2015) task, i.e., a set of strategies to adapt a source model on new target data different from the training ones. We observe that this approach is more focused, as the name suggests, on adapting instead of incrementally updating a model on new data. In other words, DA methods are focused on improving the performance of the new domain data, while in our case is essential to preserve model performance both on new and past data. In addition, DA solutions usually require a representative sample of the target domain during the adaptation procedure, but this implies the transfer of face images across different sites, not complying with possible privacy regulations. Finally, DA solutions often require a challenging parametrization that could limit their generalization capabilities and hamper their diffusion, and usage (Singhal et al., 2023).

2.4 Continual learning

Continual Learning, also known as lifelong learning, refers to the capacity to continually acquire, refine, and transfer knowledge over time (Parisi et al., 2019). While this ability is inherent in humans and animals, it is not naturally present in artificial learning systems, particularly those based on deep learning architectures. Unlike computational systems that are typically trained on fixed batches of data, they encounter difficulties in assimilating new incremental knowledge from non-stationary data distributions, leading to the phenomenon known as catastrophic forgetting (McCloskey and Cohen, 1989). The Continual Learning paradigm significantly diverges from the traditional Machine Learning approach, which involves two distinct phases: learning and deployment. In the Machine Learning paradigm, training data collected solely before the learning phase are expected to unrealistically represent all the subtleties of future test data (Graffieti et al., 2022). In this context, the fine-tuning (Friederich, 2017) procedure is a simple approach to adapt a model to data with a different distribution from the training phase. This procedure is based on a careful selection of training parameters (e.g., the initial learning rate, the optimizer) that can compromise the final model, leading to a dramatic loss of performance (Chhabra et al., 2019) and catastrophic forgetting.

Therefore, several approaches have been proposed in the CL literature to mitigate or counteract the problem of forgetting. These approaches span from regularization methods (Ahn et al., 2019), which incorporate constraints on the weight updates of neural networks, to dynamic architectures (Douillard et al., 2022), which introduce changes in architectures to accommodate new information, and memory replay methods (Van de Ven et al., 2020), which involve storing past data for use in current training procedures.

In particular, the method proposed in this work falls under the regularization-based approaches. These algorithms are designed to constrain updates to the model parameters during learning to avoid catastrophic forgetting. Regularization approaches can be classified into two main sub-categories: (i) per-weight plasticity control: selectively adjusts the per-weight plasticity depending on their importance to retain knowledge related to previously encountered data; (ii) loss function regularization: constrain the model loss to encourage the retention of past knowledge. This is usually achieved by using distillation and/or teacher-student approaches.

For a good starting point for exploring approaches belonging to the first subcategory, we refer to Elastic Weight Consolidation (EWC) (Kirkpatrick et al., 2017) and Synaptic Intelligence (SI) (Zenke et al., 2017). These and similar methods aim to prevent catastrophic forgetting by protecting critical weights, but they can struggle with large-scale, non-stationary data distributions. In some cases, they may cause the model's plasticity to collapse, effectively freezing the model and preventing it from learning from a long stream of small, diverse batches of data.

In contrast, the approach proposed in this work belongs to the second subcategory, as it uses a distillation-driven learning paradigm. Unlike methods that depend on the specifics of past experiences, distillation-based approaches allow learning from a long stream of data. Those approached also exhibit good resistance to data drifts, which, in our scenario, may be necessary to learn to detect novel morphing techniques without forgetting previous ones. However, while distillation-based methods are more flexible in handling a continuous stream of experiences, they are typically not designed to adapt to variations in the size of each batch. As demonstrated in the experimental section (see Section 4), this can be problematic in scenarios where the number of training examples available per experience cannot be predicted in advance, or when data availability is highly unpredictable.

Dynamic architecture methods, also known as structural methods, aim to reduce forgetting by evolving the structure of the network. This can be achieved through various strategies, such as expanding the model's capacity by adding new layers, duplicating modules, or even replicating the entire model. Alternatively, some methods may mask certain weights or activations to maintain previously learned knowledge. For reference, well-known approaches include Progressive Neural Networks (Rusu et al., 2016), Dynamically Expandable Networks (Yoon et al., 2018) PathNet (Fernando et al., 2017), and Piggyback (Mallya et al., 2018).

Replay mechanisms offer another effective strategy by storing examples of previously encountered data in a finite memory buffer. These examples, which can be stored as raw data or in a compressed form, are managed by algorithms that insert and replace stored examples based on various criteria such as class, task, diversity, or training loss. At each training step, these stored samples are replayed alongside new data, helping the model retain previously learned information. While these methods are efficient at reducing forgetting, they are often not suitable for use in scenarios in which the data to be handled is critical for security, privacy, and regulatory reasons. For an overview of the main replay approaches, we refer to Bagus and Gepperth (2021) and Buzzega et al. (2021).

2.5 Preliminary investigation

We conduct a preliminary investigation to assess the challenges of the proposed continual training scenario.

To this aim, we train a MAD system in two different ways. In the first case—referred to as *joint*—all training data are available at the same time: this is the traditional setting of Machine Learning, in which a single training procedure is conducted on a single, and possibly large, dataset, followed by the deployment phase. In the second case—referred to as *naive*—the MAD model is trained in an incremental manner, i.e., exploiting training data available at each step. Training data, coming from four different datasets (see Section 3.3), are sequentially exploited for model updating in order to simulate the continual MAD training scenario. The error metrics (lower values are better) are computed after each model update on an external testing dataset (see Section 3.5). For the MAD system, we adopt an MLP classifier that receives as input the embeddings produced by a ResNet-50 network trained through the ArcFace loss (Deng et al., 2019) function for the face recognition task, following the well-known MAD approach proposed in Scherhag et al. (2020).

Experimental results are reported in Figure 3 and show that the joint case represents an upper bound of performance, represented by an unvaried value since there is only a single training phase.

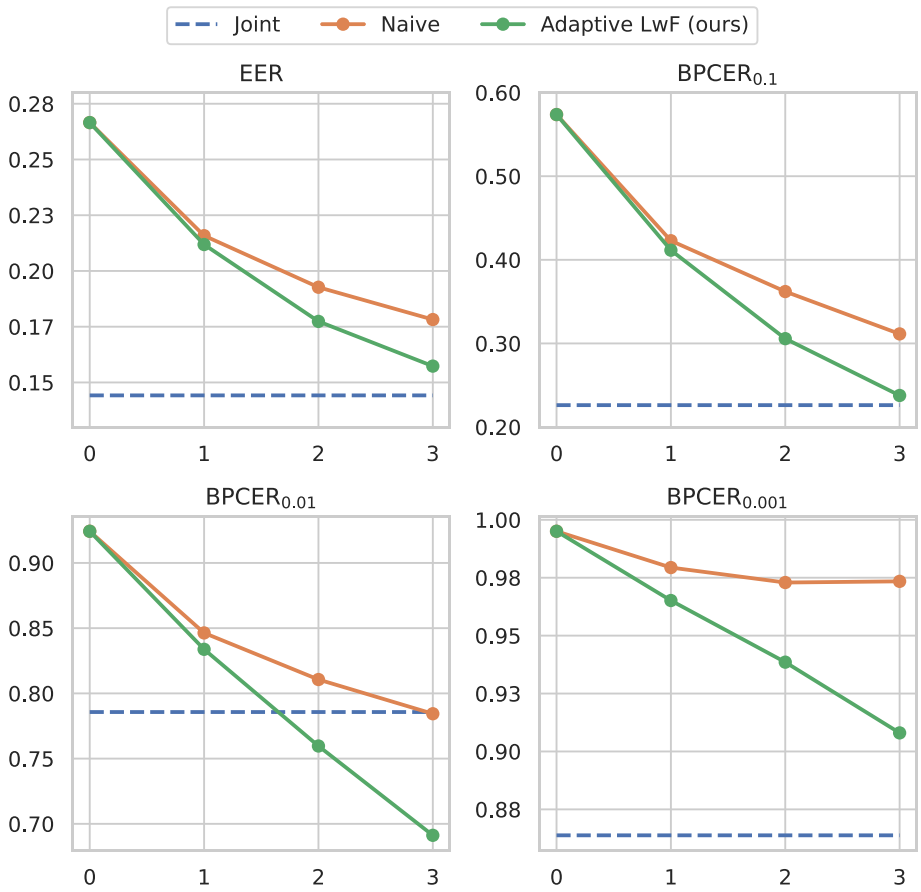

FIGURE 3
Comparison between the *joint* and *naive* approaches. The *joint* method generally represents the upper bound in terms of performance. The *naive* approach learns across different learning experiences, but a significant gap is noticeable. The proposed Adaptive-LwF is able to reduce the gap, achieving final performance similar to the joint approach.

This approach generally represents the best solution when the whole training dataset is available for a single batch-based training phase. We observe that the naive approach, whose error rates are initially high due to the limited amount of available training data, is able to improve the knowledge across different steps, limiting the negative influence of catastrophic forgetting. However, it still presents a consistent gap in performance, especially in the EER and $\text{BPCER}_{0.001}$ metrics, with respect to the joint approach due to the lack, in this simple approach, of any mechanism aimed at preventing catastrophic forgetting. Finally, the graph also includes the results of the proposed Adaptive-LwF, which exhibits better performance, achieving a lower error with respect to the naive approach across all the evaluation steps, with a final error closer to (or even better than) the joint method.

To conclude, this preliminary investigation confirms that: (i) when possible, the training on a single large dataset is preferable with respect to incremental training on different portions of data; (ii) there is a gap in performance between *joint* and *naive* approaches. This gap is reduced thanks to Adaptive-LwF, an approach specifically designed to deal with the continual MAD task, as described in the rest of the paper.

3 Proposed scenario

3.1 Notation

For the sake of paper readability, taking inspiration from the Continual Learning field and the previous work of Pellegrini et al. (2023), we introduce the following notation used in the rest of the paper.

In all the reported experiments, we simulate the incremental training scenario collecting several datasets $D = \{d_i, i = 1, ..., J\}$ of bona fide and morphed images available in the literature and then creating C, i.e., the set of available training data divided into several chunks $C = \{c_i, i = 1, ..., N\}$ and E, the testing dataset. The test set E is fixed to ensure a consistent and fair comparison of the results obtained across different testing experiences in our continual learning setting. No training or adaptation is performed on E, which is used exclusively for model evaluation.

Two main elements characterize the proposed MAD training scenario:

- **Learning experience**: a learning experience l is a training procedure in which the MAD model M is updated taking into account new data. The amount of new training data is

variable and not known in advance, i.e., the amount of training data may vary across different learning experiences. Then, a learning experience is formally defined as a pair:

$$l_i = (M_k, c_i),\ 1 \leq i \leq N \tag{1}$$

where M_k represents the MAD model trained at the l_k that is being updated with new data chunks $c_i \in C$ of variable size. $N = |C|$ is the total number of data chunks available.

- **Testing experience**: during a testing experience t the model M_k is tested on the set E. $N = |C|$, the total number of data as defined before. Then, from a formal point of view:

$$t_i = (M_k, E),\ 1 \leq i, i \leq N \tag{2}$$

Testing experiences are particularly important to monitor the performance of the MAD model across the incremental training, i.e., after the different learning experiences l_i. In our experiment, we conduct a single testing experience after each learning experience: therefore, the proposed incremental training scenario is formally defined as :

$$B = (l_i, t_i),\ i = 1, ..., N \tag{3}$$

or rather an ordered set of different training experiences l_i, each followed by a testing experience t_i. Following the previous considerations, during each l_i past and future chunks of data are not available for the training procedure.

3.2 Benchmarks

The key idea in the proposed benchmarks is to complement the scenario typically employed in continual learning applications, which usually involves the use of fixed-sized data chunks, with a contextually diverse scenario inspired by real-world MAD applications, where data may become available in variable quantities. Therefore, we model the size of data chunks that become available across the learning experiences l_i in five different ways.

1. **Fixed**: all data chunks c_i used in B have the same size s, obtained by splitting all the training datasets, after a shuffling procedure, in equal parts.

$$|c_i| = s,\ i = 1, ..., N \tag{4}$$

We observe this is the typical setting of the Continual Learning paradigm, in which a certain amount of new training data becomes available, and it is used for model updating. This configuration is useful in order to investigate the effects of the variation of the weight loss with respect to the chunk size.

2. **Variable-dataset**: in this case each single training dataset d_i corresponds to a single chunk c_i, then resulting in $J = |D|$ different learning and testing experiences.

$$|c_i| = |d_i|,\ i = 1, ..., J \tag{5}$$

From a practical point of view, this corresponds, for instance, to the case where a new dataset is acquired and becomes available in the literature, and then the MAD model is updated accordingly.

3. **Variable-uniform**: each data chunk c_i has the same probability to have a size in the range $[a, b]$. More formally:

$$|c_i| = U(a, b),\ i = 1, ..., N \tag{6}$$

where $U(a, b)$ represents the uniform distribution in the range $[a, b]$ given by the function:

$$f(x) = \begin{cases} \frac{1}{b-a}, & a \leq x \leq b \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

where $a = 50$ and $b = 500$. In other words, across different learning experiences, the chunk size may vary with the same probability, not favoring any specific size within the defined range. In a real-world scenario, this corresponds to the case in which several new data chunks of varying sizes become sequentially available.

4. **Variable-small**: as in the previous scenario, we fix the range $[a, b] = [50, 500]$ of the possible size of a chunk, and at training time the size of each chunk is sampled from this range following Z, i.e., the *Zipf*'s law (Li, 2002):

$$|c_i| = Z(a, b),\ i = 1, ..., N \tag{8}$$

where Z is defined as the function:

$$f(x) = \frac{p}{|x - a| + \epsilon} \tag{9}$$

where x is the size corresponding to the i-th position in the above-defined range, and p is the normalization parameter equal to the reciprocal of the sum of all probabilities. In this benchmark, small chunk sizes are more likely to be selected, focusing on small sets of new data that become available, for instance, at airports, for a limited time range.

5. **Variable-large**: similar to the previous case, but the range is $[500, 50]$, and then $|c_i| = Z(b, a)$. In this case, large chunk sizes are more likely to be selected, as follows by Equation 9. Conversely, in this scenario, we focus on large sets of new training data that, for instance, become available after specific acquisition procedures operated by institutions.

3.3 Datasets

In this section, we describe the datasets from which data chunks are created. A summary of the dataset and morphing algorithms used in the experimental evaluation is reported in Table 1.

3.3.1 Progressive morphing database

The Progressive Morphing Database (PMDB) (Ferrara et al., 2017) is a dataset containing morphed images created with the algorithm described in (Ferrara et al., 2017), starting from the original images of AR (Martinez and Benavente, 1998), Color Feret (Phillips et al., 1998), and FRGC (Phillips et al., 2005) datasets. A total of 280 subjects (146 females and 134 males) are present. Some visible artifacts are visible in the morphed faces, while the background is automatically replaced and then it does not include any artifacts.

TABLE 1 Overview of the morphing techniques and datasets adopted in the experimental analysis.

Morphing algorithm	Dataset	Data source	#Morphed	Quality
UBO (Ferrara et al., 2017)	PMDB	AR - FRGC - Feret	711 - 199 - 198	Medium
OpenCV	Idiap Morph	FRGC - FRLL - Feret	964 - 1,222 - 529	Low
FaceMorpher	Idiap Morph	FRGC - FRLL - Feret	964 - 1,222 - 529	Low
StyleGAN (Karras et al., 2020)	Idiap Morph	FRGC - FRLL - Feret	964 - 1222 - 529	Low
WebMorph (DeBruine and Jones, 2017)	Idiap Morph	FRLL	1,221	Low
AMSL (Neubert et al., 2018)	Idiap Morph	FRLL	2,175	Low
Sqirlz morph	MorphDB	FRGC - Feret	50 - 50	High

For each technique, the corresponding dataset and the origin of the images employed in the morphing process are specified. Additionally, the number of generated morphed images is provided for each data source, while the final column indicates the visual quality of the resulting morphs.

3.3.2 Idiap morph

It is a public collection of different datasets, i.e., a set of morphed images created using different morphing algorithms, i.e., FaceMorpher,[3] OpenCV,[4] StyleGAN (Karras et al., 2020), WebMorph (DeBruine and Jones, 2017), and AMSL (Neubert et al., 2018), taking the original face images from the Color Feret (Phillips et al., 1998), the Face Research Lab London Set (DeBruine and Jones, 2017), and FRGC (Phillips et al., 2005). The overall quality of morphed images varies since some visible artifacts are present in the background or the foreground, especially in the areas of the eyes, nose, and mouth. A total of about 11.5k morphed images are available.

3.3.3 MorphDB

In this dataset (Ferrara et al., 2017), morphing images are created starting from the Color Feret (Phillips et al., 1998) and FRGC (Phillips et al., 2005) databases. The dataset consists of 100 high-quality morphed images, belonging to 50 male and 50 female subjects, obtained through the Sqirlz Morph algorithm.[5] Due to the presence of a manual retouch to remove visible artifacts, this dataset represents a challenging set for the MAD task. MorphDB is not publicly released, but it is available on the FVC-onGoing (see text footnote 2) web platform as a sequestered dataset.

3.4 Experimental setup

The training set D consists of the images taken from the Feret, FRGC, and FRLL datasets using the StyleGAN, OpenCV, and FaceMorpher morphing algorithms, respectively. The testing set E is composed of the MorphDB dataset, which, despite its limited size, represents a challenging dataset due to the presence of manual retouching on morphed images, and considering that training data do not have the same visual quality (i.e., morphed images present visible artifacts).

Each data chunk is obtained by dividing the training datasets and then simulating incremental training across different sites. Since the order of the training experiences, and then the order of the data chunks that belong to different datasets, may influence the final result, we average the reported performance on all possible $4! = 24$ orders.

In all experiments, we address the Differential MAD (D-MAD) task, in which then the input is represented by two different images, i.e., the document and the trusted live acquisition. Besides, we include pairs that include both the criminal and the accomplice: from a practical point of view, the pairs with the accomplice represent the case in which the accomplice presents the morphed image during, for instance, the enrollment procedure.

3.5 Metrics

To measure the performance of MAD model at each testing experience t_i, we exploit the common error-based metrics used for the MAD task. The Bona Fide Presentation Classification Error Rate (BPCER) indicates the percentage of bona fide images wrongly classified as morphed:

$$BPCER(\sigma) = \frac{1}{N}\sum_{i=1}^{N} H(b_i - \sigma) \tag{10}$$

where σ is the threshold on which the detection score of bona fide b_i scores are compared, and the step function $H(x) = 1, \text{if} x > 0, 0$ otherwise. Conversely, the Attack Presentation Classification Error Rate (APCER) indicates the percentage of morphed images wrongly classified as bona fide:

$$APCER(\sigma) = 1 - \left[\frac{1}{M}\sum_{i=1}^{M} H(m_i - \sigma)\right] \tag{11}$$

where σ is the threshold on which the detection score of morphed m_i scores are compared, and the step function is the same.

APCER and BPCER values are usually computed at different thresholds, in particular, the BPCER is reported with respect to a defined value of APCER. In our experiments, we consider $BPCER_{0.1}$, $BPCER_{0.01}$ and $BPCER_{0.001}$, that represent the lowest BPCER with APCER $\leq$ 10%, $\leq$ 1%, and $\leq$ 0.1%. We observe that the last point represents a challenging case and it is the

3 https://github.com/alyssaq/face_morpher

4 https://learnopencv.com/face-morph-using-opencv-cpp-python

5 https://sqirlz-morph.it.uptodown.com/windows

typical working point of face verification-based systems in reality. In addition, we also report the point at which the BPCER is equal to the APCER, i.e., the Equal Error Rate (EER), useful to compare different MAD approaches through a single and general value.

In order to summarize the results across different testing experiences t_i, we compute the Temporal-Area Under the Curve (T-AuC) metric, obtained through the trapezoidal rule and averaged by the number of training experiences l_i. In other words, the traditional AuC metric is here adapted to consider the temporal evolution of the previously defined error metrics, i.e., the variation across different testing experiences.

Finally, we hereby present an additional metric referred to as *Borda Ranking over Time* (BRoT), that aims to measure the performance throughout the entire learning process. Specifically, BRoT is calculated over a collection of algorithms denoted as $\mathcal{A}$, and it is predicated upon the concept of bestowing recognition to algorithms that exhibit superior performance at each testing iteration. Let $r(a_j, t_i)$ be the ranking of algorithm $a_j \in \mathcal{A}$ at the testing experience t_i; ranking is here established according to the $\text{BPCER}_{0.001}$ for MAD. At each t_i, Borda count (Lippman, 2013) is applied to score the tested algorithms, i.e., a decreasing number of points $p(r(a_j, t_i))$ is assigned to each algorithm based on the corresponding ranking, with $p(i) = |\mathcal{A}| - i$. For each algorithm a_j, the points are accumulated over the different learning experiences, and the total score is finally normalized by the maximum theoretical score:

$$\text{BRoT}(a_j) = \frac{\sum_{i=1}^{N} p(\mathcal{R}(a_j, t_i))}{|\mathcal{A}| \times N} \tag{12}$$

where N is the total number of testing experiences.

The T-AuC metric serves as a valuable tool for comprehending the overall performance of a method throughout several testing experiences, emphasizing the highest attained accuracy. Conversely, the BRoT metric offers insights into which algorithm is most likely to exhibit consistently high accuracy throughout the entire learning process, regardless of potential disparities in absolute accuracy between different methods.

4 Adaptive-Lwf

4.1 Model

To tackle the D-MAD task, we adopt a pipeline inspired by the state-of-the-art work reported in Scherhag et al. (2020): two different feature embeddings of size 512 are extracted from the two input images through a ResNet backbone trained on different facial datasets (i.e., VGGFace2, CASIA, and MS1MV2) with the ArcFace loss (Deng et al., 2019). This choice is motivated by the good performance this approach obtained on the FVC-onGoing and FRVT MORPH NIST benchmarks. These embeddings are then combined through subtraction and then fed in input to a Multi-Layer Perceptron (MLP) consisting of 5 hidden layers of size 512, 250, 125, 64 (differently from the original work in which the classifier is an SVM). Finally, in output, there are 2 neurons that emit the probability of the input images belonging to the class "morphed" or "bona fide." The MLP classifier has been selected in order to have the possibility to apply CL algorithms, possible only on deep learning-based architectures. The network is trained with the Categorical Cross Entropy (CCE) loss and the ReLU is used as the activation function. The SGD is used as an optimizer, with a learning rate of $10^{-}2$ and a momentum of 0.9. No weight decay is applied.

4.1.1 Analysis of continual learning strategies

Then, we move our analysis to Continual Learning methods that can be applied in the proposed scenarios. This analysis is motivated by the fact that Continual Learning methods are typically designed to operate on streams of similarly-sized experiences, as commonly assumed in Class- and Task-Incremental settings. Considering that these scenarios are characterized by different experience size distributions, the goal of this preliminary investigation is to decide which method can better manage learning in these more challenging setups under the default hyperparameter configurations. We evaluate these representative approaches on the two opposite-sized scenarios, Small-Variable and Large-Variable. In particular, we consider several mainstream replay-free Continual Learning methods, such as Synaptic Intelligence (SI) (Zenke et al., 2017), Elastic Weight Consolidation (EWC) (Kirkpatrick et al., 2017), Deep Streaming Linear Discriminant Analysis (SLDA) (Hayes et al., 2019), and Learning without Forgetting (LwF) (Li and Hoiem, 2017). SI exploits a quadratic regularization term to preserve the model's weights that mostly contribute to the performance of past learning experiences. EwC tackles the catastrophic forgetting problem through a penalty loss that constrains the changes in the model parameters and weights deemed most important to preserve past knowledge. Different from the other approaches, SLDA implements a covariance matrix in the final classification stage, relying on features extracted from a pre-trained learning model. Finally, the LwF approach uses a distillation strategy, consisting of an old and a current model, to prevent forgetting. In addition to these methods, we also ran experiments with the *naive* approach, that is, incremental learning without adopting any forgetting-contrasting strategy, and *joint* approach, which serves as the reference upper bound. The results, reported in Table 2, show that LwF is the best-performing approach, achieving the lowest error rate computed relative to the joint training upper bound on both scenarios. For this reason, we choose LwF as the basis for our proposed adaptive approach.

4.2 Proposed algorithm

This section describes our Adaptive-LwF algorithm to tackle the challenges introduced in the incremental training scenario. In particular, also following the findings described in Pellegrini et al. (2023), we base our method on the Learning without Forgetting (LwF) algorithm (Li and Hoiem, 2017), originally introduced in the Continual Learning field. One interesting feature of this algorithm is that it is not based on the use of a replay memory, which temporarily stores part of the data previously used for training, thus

not raising privacy-related concerns typically present in the MAD literature.

From a technical point of view, LwF contrasts catastrophic forgetting with a distillation-based approach, balancing and properly combining two different models: one old model, which maintains the knowledge of past learning experiences, and a current one, which tries to learn from new samples.

In this context, the weight of the distillation loss, usually referred to with the symbol λ, represents the weight assigned to the old model when learning from the current experience. For this reason, λ is a key parameter to be set in order to improve the performance of the CL approach. In particular, higher values of λ are associated with a conservative approach that aims to preserve past knowledge at the expense of a slower adaptation to new data. On the contrary, lower values determine a faster adaptation to the incoming data, with more substantial modifications of the old weights.

Unfortunately, in the literature, it is not clear how to properly set the value of λ, especially in relation to different input data sizes, but, on the other hand, this parameter strongly influences the performance of the algorithm. We prove this through the experiment reported in Figure 4, in which we show how different values of the distillation loss positively or negatively influence the final MAD classification (up to a difference of $\sim$ 5% in the EER, and $\sim$ 15% in $\text{BPCER}_{0.01}$) adopting the Variable-Dataset benchmark. The results of this experiment suggest that it is possible to identify a proper λ value set in order to maximize the performance. Indeed, $\lambda = 100$ present always the best performance, while, for instance, $\lambda = 1,000$ represents the worst choice in this specific benchmark. Therefore, further investigations about how to choose the proper value for the distillation loss are needed.

In order to further analyze the impact of λ on the MAD performance, we run a set of experiments in the Fixed benchmark, whose results are reported in Figure 5; in particular, the colored matrices visually represent the T-AuC and BRoT metrics previously introduced (see Section 3.5) computed from different error indicators, as a function of the value of λ (reported in the *x*-axis) and the chunk sizes (reported in the *y*-axis): darker color denotes better performance (low error). As shown, diagonal dark regions suggest two observations: (i) it is important to define the proper range of λ since specific values strongly impact the performance in a negative way. For instance, for $\lambda = 100$ or

TABLE 2 Comparison of different CL methods under both Small-Variable and Large-Variable scenarios.

Exp. size	MAD ↓	
	Small	Large
Naive	+14%	+16%
EWC (Kirkpatrick et al., 2017)	+14%	+16%
SI (Zenke et al., 2017)	+21%	+24%
SLDA (Hayes et al., 2019)	+27%	
LwF (Li and Hoiem, 2017)	**+13%**	**+12%**

Results are reported as percentage variations in T-AuC with respect to the ideal *joint* training case. Positive values indicate higher error. Lower values are better. Bold values are the best scores obtained.

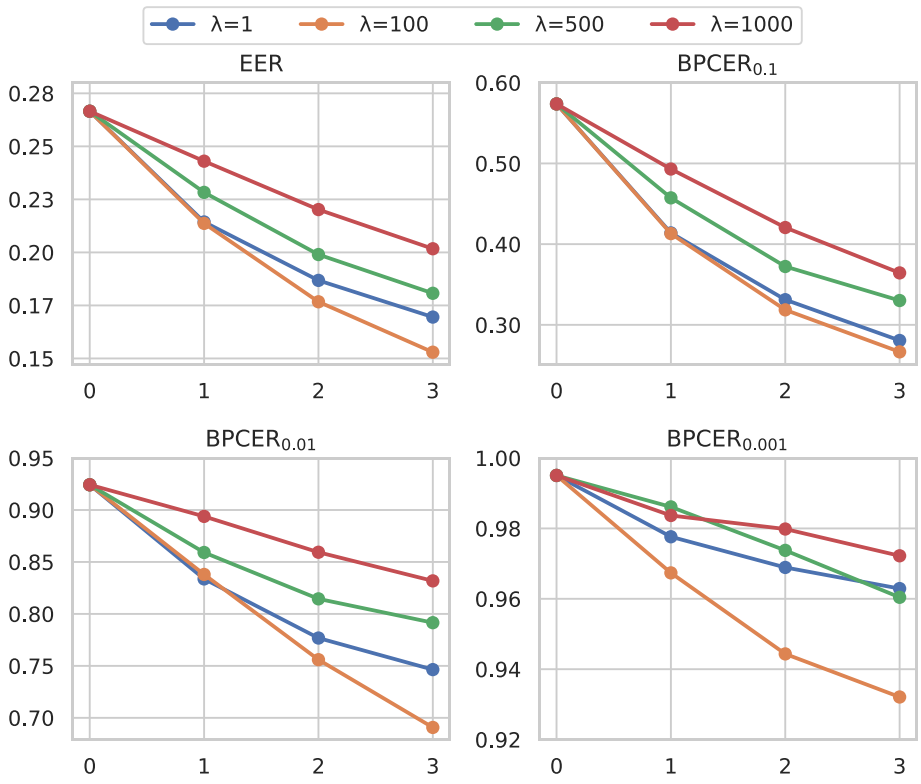

FIGURE 4
Effects of varying the distillation loss (λ) values in the Variable-Dataset benchmark (see Section 3.2). Results suggest that for a given chunk size, a proper λ value can be found.

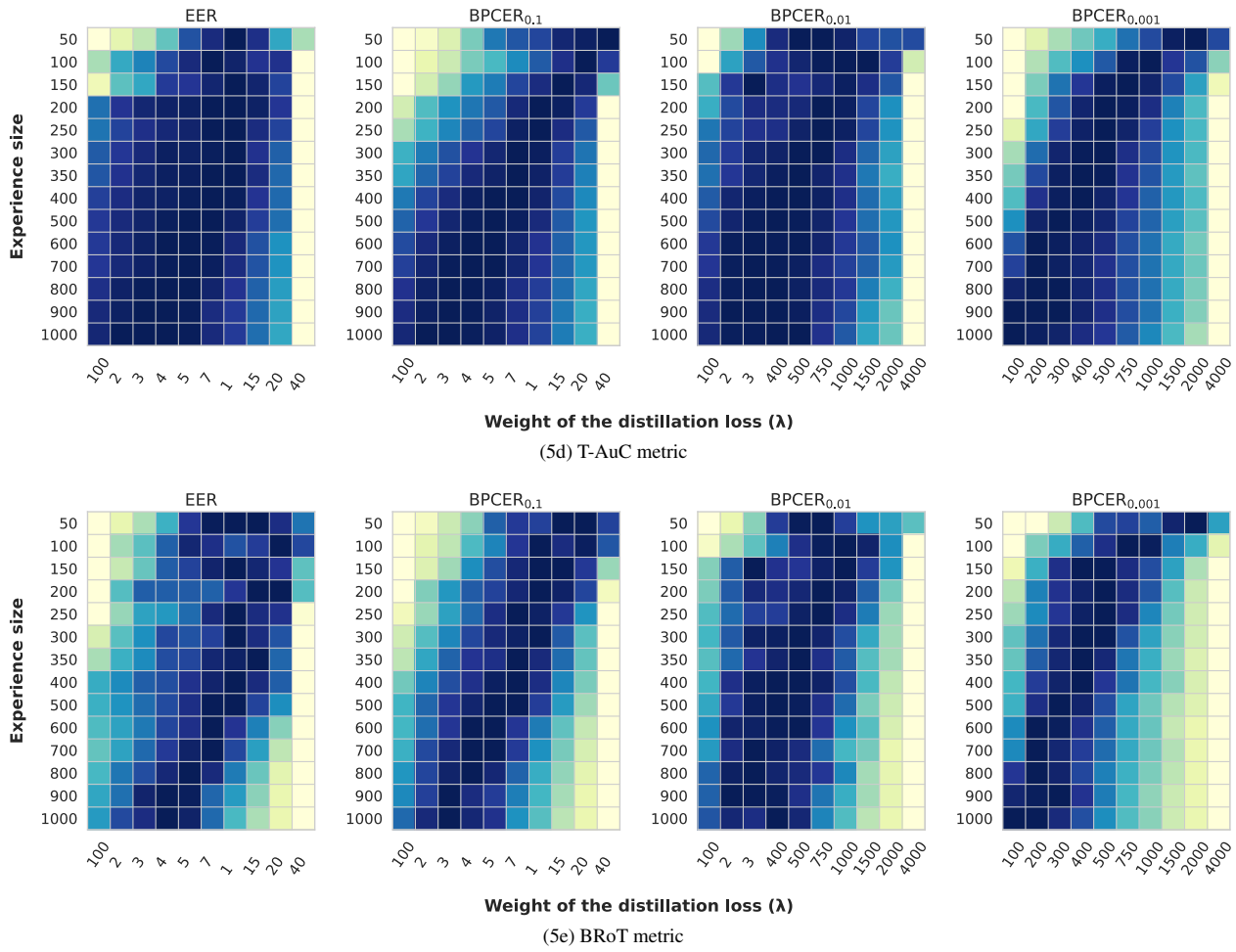

FIGURE 5
Performance of the LwF (Li and Hoiem, 2017) algorithm on the MAD task analyzed in relation to varying training experience sizes (y-axis) and distillation loss (λ) values (x-axis). In the first line, results are condensed through the T-AuC metric, while in the second one the BRoT metric is expressed. The intensity of color represents the quality of performance, with darker shades indicating better results. The results indicate that different λ values have a significant impact on the performance of the LwF algorithm, suggesting that larger values are preferable for smaller experience sizes.

4,000, the model performance is limited. (ii) it is possible to reduce the error by setting λ inversely proportional with respect to the size of the input chunks. This is particularly visible in the $\text{BPCER}_{0.001}$ metric, a point that represents the typical working point of real-world applications (e.g., ABC gates). Both BroT and T-AuC metrics, exhibit the same behavior, revealing that the proper choice of λ is needed to enhance the probability of achieving the best performance on the whole continual training procedure. Aimed at finding an analytic expression for the optimal lambda value, we reported in the graph of Figure 6 the optimal λ value obtained experimentally for each chunk size. The selected optimal λ value is obtained by adding the T-AuC of the two most significant error indicators, i.e., the EER and the $\mathbf{B}_{0.001}$. The reported trend clearly shows that the exponential function well approximates the data and can therefore be used to properly estimate the distillation loss given the chunk size.

By fitting a regression function over the values observed in our experimental analysis (Figure 6) we empirically derived a function to compute an approximation of the optimal λ value: $\lambda = ax^{-b}$, where $a = 17,403$, $b = 0.636$ and x is the input chunk size. The direct use of this formula eliminates the need for repeated hyperparameter tuning (e.g., grid search or manual selection of λ). Such tuning procedures typically require multiple training runs per experience (especially in the case of variable-size chunks), which can significantly increase the computational cost.

To validate our empirical function for the estimation of λ, we apply this Adaptive-LwF mechanism in three different Variable benchmarks (Uniform, Large, and Small), with the aim of assessing the validity of the proposed approach under different data distributions. The results, reported in Figure 7, compare the proposed Adaptive-LwF approach to three other approaches. The first one is the naive approach, i.e., the simple MAD model updating without any mechanism to control catastrophic forgetting. Moreover, we test two different LwF algorithms with a static value of the distillation loss, i.e., the same value of λ for all the learning experiences. In particular, we adopt $\lambda = 1$ and $\lambda = 1,500$, in order to highlight the contribution of small and large values of the loss itself. The graphs clearly illustrate that the naive approach fails to effectively leverage the newly

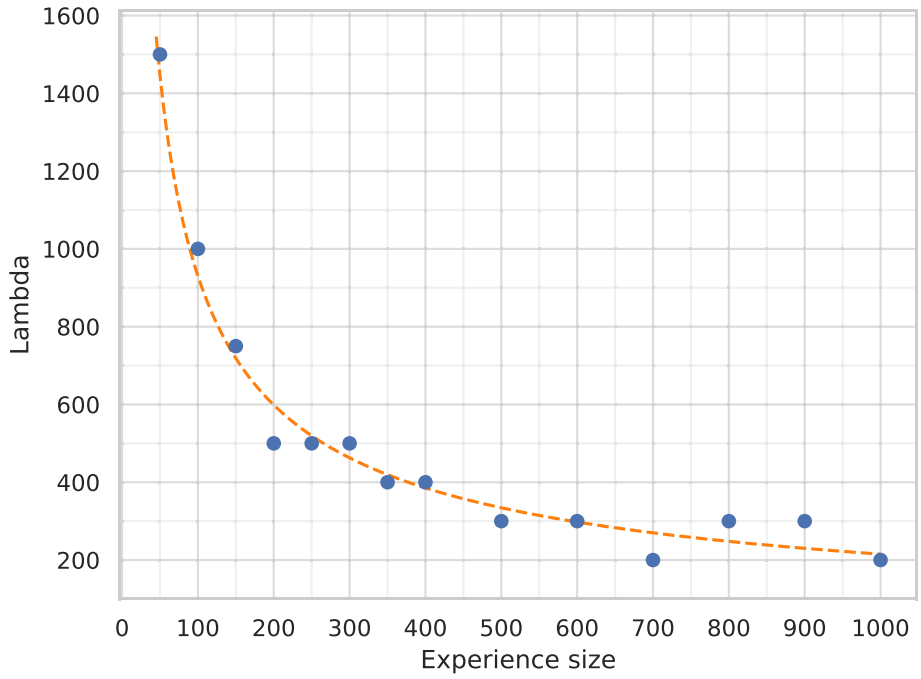

FIGURE 6
Optimal lambda values for each data chunk size reported as blue dots. As shown, an exponential function can be used to interpolate the points and to compute the proper value of λ with respect to the input size.

incoming data, resulting in only minimal error reduction across the learning experiences. Similarly, the LwF algorithm exhibits comparable behavior when a low value for the distillation loss parameter is used ($\lambda = 1$), demonstrating that such configurations do not adequately support robust learning. In contrast, the LwF algorithm with a fixed high distillation loss value ($\lambda = 1,000$) generally produces noteworthy results, indicating that higher λ values can lead to more stable learning outcomes. However, the proposed Adaptive-LwF approach consistently outperforms both the naive and fixed-parameter LwF strategies. This performance improvement is accompanied by a significant practical advantage: the elimination of the need to experimentally determine an optimal parameter value. This dual benefit—the enhanced learning performance and the simplification of hyperparameter tuning—reinforces the value of our proposed methodology.

4.3 Results and limitations

The findings of the experimental evaluation can be summarized as follows. Adaptive tuning dynamically adjusts the distillation loss weight (λ) based on the size of new input data, enabling effective integration of new information while preserving prior knowledge. This approach mitigates catastrophic forgetting, where previously acquired knowledge is lost during adaptation to new data, by tailoring the influence of the new data to prevent overwriting essential features. Proper parameter adjustment ensures that the model maintains a balance between learning new information and retaining performance on older tasks, supporting robust generalization across learning experiences. Predefining optimal parameter values for diverse scenarios is impractical and time-consuming: in this context, the Adaptive-LwF algorithm automates this process using mathematical functions (i.e., the exponential scaling of λ), optimizing computational efficiency while maintaining high accuracy.

The introduced Adaptive-LwF exhibits some limitations. In particular, the algorithm relies on precise tuning of the distillation loss parameter, which is essential for balancing old and new knowledge. An analytic function is proposed in this work and its effectiveness is confirmed by the experimental evaluation; however, the generalization of this function to different testing scenarios should be assessed. In addition, as with incremental learning methods in general (not limited to the proposed approach), it should be noted that the performance of Adaptive-LwF may remain below that of traditional batch learning approaches, due to the inherent constraints of learning from sequentially available data rather than from the full training set at once.

The benchmarks used for evaluation, such as Fixed, Variable-Dataset, and Variable-Small, simulate limited real-world scenarios and may only partially reflect the complexities of dynamic or adversarial environments. Then, the algorithm's robustness against heavily imbalanced or noisy datasets remains to be tested. Privacy constraints further limit Adaptive-LwF by preventing access to old data, excluding replay mechanisms and relying solely on knowledge distillation, which may falter with highly divergent tasks.

Dependency on pre-trained ResNet embeddings introduces an additional limitation, as the performance is tied to the generalization quality of these features. Moreover, while the method performs well for tasks with clear boundaries, it may struggle in scenarios in which the data updates are not well-defined.

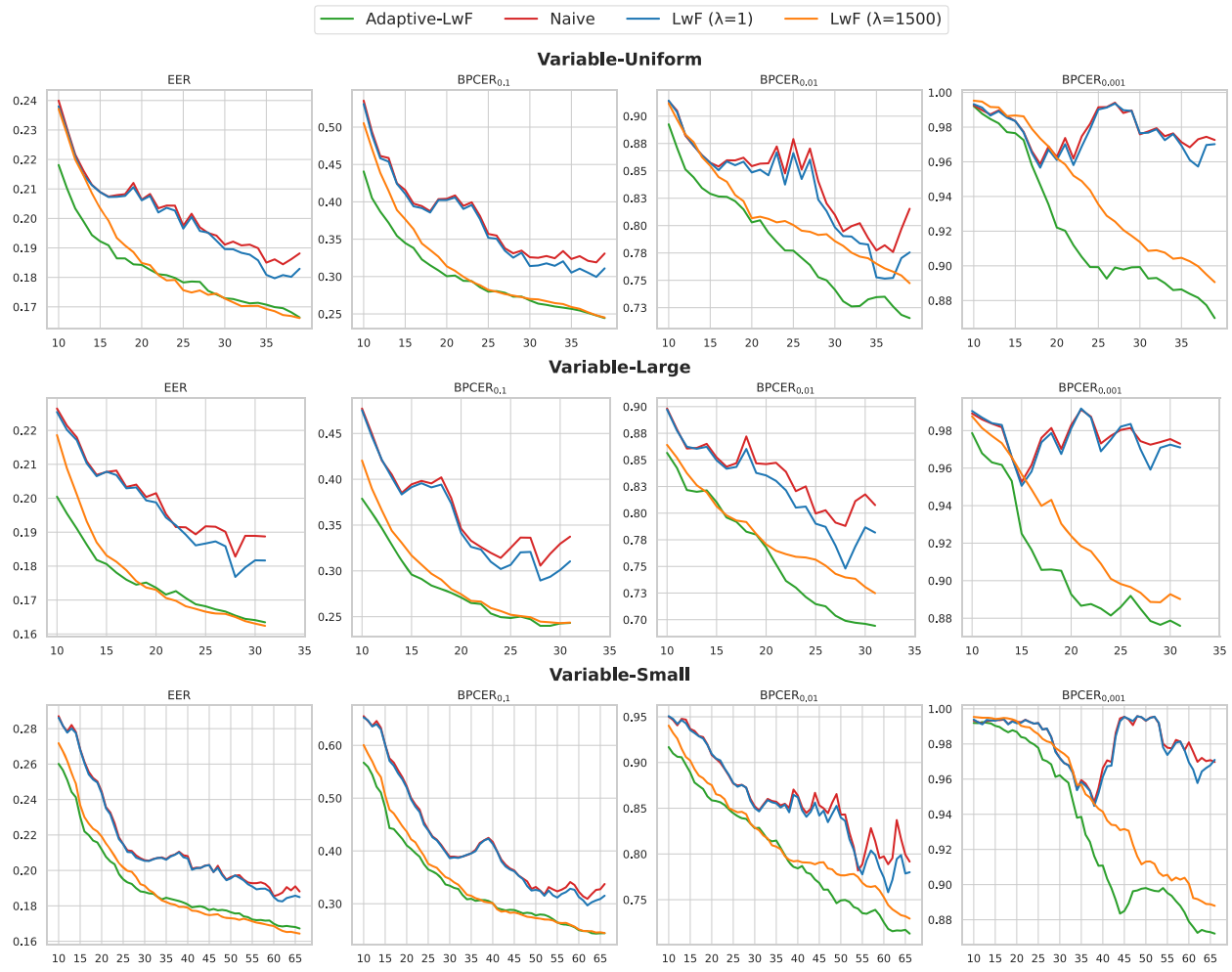

FIGURE 7
Performance of the proposed Adaptive-LwF on the benchmarks introduced in Section 3.2. In particular, our Adaptive-LwF (green) is compared with LwF (Li and Hoiem, 2017) with a fixed $\lambda - \lambda = 1$ (blue) $\lambda = 1,500$ (orange), and the *naive* approach (red), implemented as described in Section 2.5. In all the reported benchmarks, Adaptive-LwF represents the best solution in reducing each final error metric.

5 Conclusion

In this paper, we introduce Adaptive-Lwf, a Continual Learning algorithm that addresses the Differential Morphing Attack Detection (D-MAD) task in a continual training scenario. In scenarios where data cannot be transferred between sites due to privacy concerns, this paradigm proves valuable by enabling model transfer instead of data transfer. In particular, considering that data chunks available in each training experience may have varying sizes in practical usage, we prove that it is possible to select the appropriate parameter of the distillation loss (λ) in the LwF approach in relation to the learning experience size.

The outcomes of our paper can be summarized as follows: (i) experimental results confirm the effectiveness of the Adaptive-LwF approach for training a D-MAD detector in an incremental manner to overcome privacy concerns; (ii) the introduced benchmarks pave the way to future research direction in the MAD field, since they represent a challenging scenario that require specific solutions; (iii) further research investigations are required to improve the overall accuracy of MAD models trained in the proposed scenario, as their performance falls short in comparison to models trained through batch-based procedures on the whole single dataset.

We believe that our findings will yield significant value for forthcoming research endeavors in the field of Morphing Attack Detection systems. This includes facilitating incremental training, addressing privacy concerns, and training models on expansive and heterogeneous datasets. Furthermore, our findings make a valuable contribution to the Continual Learning task by introducing a way to properly determine the distillation loss parameter of LwF in relation to the size of the input training data.

Data availability statement

The data analyzed in this study is subject to the following licenses/restrictions: NDA. Requests to access these datasets should be directed to guido.borghi@unibo.it.

Author contributions

LP: Software, Writing – review & editing, Writing – original draft, Data curation. GB: Writing – original draft, Conceptualization, Writing – review & editing, Investigation. AF: Writing – original draft, Investigation, Writing – review & editing, Conceptualization. DM: Writing – original draft, Funding acquisition, Project administration, Writing – review & editing.

Funding

The author(s) declared that financial support was received for this work and/or its publication. This project received funding from the European Union's Horizon 2020 research and innovation program under Grant Agreement No. 883356.

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declared that generative AI was not used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

Author disclaimer

This text reflects only the author's views, and the Commission is not liable for any use that may be made of the information contained therein.

References

Ahn, H., Cha, S., Lee, D., and Moon, T. (2019). "Uncertainty-based continual learning with adaptive regularization," in *Advances in Neural Information Processing Systems*, 32.

Bagus, B., and Gepperth, A. (2021). "An investigation of replay-based approaches for continual learning," in *2021 International Joint Conference on Neural Networks (IJCNN)*, 1–9. doi: 10.1109/IJCNN52387.2021.9533862

Banabilah, S., Aloqaily, M., Alsayed, E., Malik, N., and Jararweh, Y. (2022). Federated learning review: fundamentals, enabling technologies, and future applications. *Inf. Proc. Manag.* 59:103061. doi: 10.1016/j.ipm.2022.103061

Borghi, G., Franco, A., Domenico, N. D., Ferrara, M., and Maltoni, D. (2024). "V-mad: video-based morphing attack detection in operational scenarios," in *Proceedings of the International Joint Conference on Biometrics (IJCB24)* (Buffalo, NY, USA). doi: 10.1109/IJCB62174.2024.10744469

Borghi, G., Franco, A., Graffieti, G., and Maltoni, D. (2021a). Automated artifact retouching in morphed images with attention maps. *IEEE Access* 9, 136561–136579. doi: 10.1109/ACCESS.2021.3117718

Borghi, G., Graffieti, G., Franco, A., and Maltoni, D. (2022). "Incremental training of face morphing detectors," in *2022 26th International Conference on Pattern Recognition (ICPR)* (IEEE), 914–921. doi: 10.1109/ICPR56361.2022.9956395

Borghi, G., Pancisi, E., Ferrara, M., and Maltoni, D. (2021b). A double siamese framework for differential morphing attack detection. *Sensors* 21:3466. doi: 10.3390/s21103466

Buzzega, P., Boschini, M., Porrello, A., and Calderara, S. (2021). "Rethinking experience replay: a bag of tricks for continual learning," in *2020 25th International Conference on Pattern Recognition (ICPR)*, 2180–2187. doi: 10.1109/ICPR48806.2021.9412614

Chaudhary, B., Aghdaie, P., Soleymani, S., Dawson, J., and Nasrabadi, N. M. (2021). "Differential morph face detection using discriminative wavelet sub-bands," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1425–1434. doi: 10.1109/CVPRW53098.2021.00158

Chhabra, S., Majumdar, P., Vatsa, M., and Singh, R. (2019). "Data fine-tuning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 8223–8230. doi: 10.1609/aaai.v33i01.33018223

Dalal, N., and Triggs, B. (2005). "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)* (IEEE), 886–893. doi: 10.1109/CVPR.2005.177

DeBruine, L., and Jones, B. (2017). Face research lab London set. *Psychol. Methodol. Des. Anal.*

Deng, J., Guo, J., Xue, N., and Zafeiriou, S. (2019). "Arcface: additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4690–4699. doi: 10.1109/CVPR.2019.00482

Di Domenico, N., Borghi, G., Franco, A., and Maltoni, D. (2024). "Face restoration for morphed images retouching," in *2024 12th International Workshop on Biometrics and Forensics (IWBF)* (IEEE), 1–6. doi: 10.1109/IWBF62628.2024.10593948

Douillard, A., Ramé, A., Couairon, G., and Cord, M. (2022). "Dytox: Transformers for continual learning with dynamic token expansion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. doi: 10.1109/CVPR52688.2022.00907*

Fernando, C., Banarse, D., Blundell, C., Zwols, Y., Ha, D., Rusu, A. A., et al. (2017). Pathnet: Evolution channels gradient descent in super neural networks. *arXiv preprint arXiv:1701.08734*.

Ferrara, M., Annalisa, F., and Davide, M. (2014). "The magic passport," in *IEEE International Joint Conference on Biometrics (IJCB–14)*, 1–7. doi: 10.1109/BTAS.2014.6996240

Ferrara, M., and Franco, A. (2022). "Morph creation and vulnerability of face recognition systems to morphing," in *Handbook of Digital Face Manipulation and Detection: From DeepFakes to Morphing Attacks* (Cham: Springer International Publishing), 117–137. doi: 10.1007/978-3-030-87664-7_6

Ferrara, M., Franco, A., and Maltoni, D. (2017). Face demorphing. *IEEE Trans. Inf. Forens. Secur.* 13, 1008–1017. doi: 10.1109/TIFS.2017.2777340

Friederich, S. (2017). Fine-tuning.

Graffieti, G., Borghi, G., and Maltoni, D. (2022). Continual learning in real-life applications. *IEEE Robot. Autom. Lett.* 7, 6195–6202. doi: 10.1109/LRA.2022.3167736

Hayes, T. L., Cahill, N. D., and Kanan, C. (2019). "Memory efficient experience replay for streaming learning," in *2019 International Conference on Robotics and Automation (ICRA)*, 9769–9776. doi: 10.1109/ICRA.2019.8793982

Kang, P., Hwang, S.-S., and Cho, S. (2007). "Continual retraining of keystroke dynamics based authenticator," in *Advances in Biometrics: International Conference, ICB 2007, Seoul, Korea, August 27–29, 2007* (Proceedings. Springer).

Kannala, J., and Rahtu, E. (2012). "Bsif: binarized statistical image features," in *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)* (IEEE), 1363–1366.

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., and Aila, T. (2020). "Analyzing and improving the image quality of stylegan," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. doi: 10.1109/CVPR42600.2020.00813*

Kessler, R., Raja, K., Tapia, J., and Busch, C. (2023). Towards minimizing efforts for morphing attacks-deep embeddings for morphing pair selection and improved morphing attack detection. *arXiv preprint arXiv:2305.18216*. doi: 10.1371/journal.pone.0304610

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., et al. (2017). Overcoming catastrophic forgetting in neural networks. *Proc. Nat. Acad. Sci.* 114, 3521–3526. doi: 10.1073/pnas.1611835114

Li, W. (2002). Zipf's law everywhere. *Glottometrics* 5, 14–21.

Li, Z., and Hoiem, D. (2017). Learning without forgetting. *IEEE Trans. Pattern Anal. Mach. Intell.* 40, 2935–2947. doi: 10.1109/TPAMI.2017.2773081

Lippman, D. (2013). *Voting theory*. Creative Commons BYSA.

Long, M., Yao, Q., Zhang, L.-B., and Peng, F. (2024). Face de-morphing based on diffusion autoencoders. *IEEE Trans. Inf. Forens. Secur.* 19, 3051–3063. doi: 10.1109/TIFS.2024.3359029

Mallya, A., Davis, D., and Lazebnik, S. (2018). "Piggyback: adapting a single network to multiple tasks by learning to mask weights," in *Computer Vision-ECCV 2018*, eds. V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss (Cham: Springer International Publishing), 72–88. doi: 10.1007/978-3-030-01225-0_5

Martinez, A., and Benavente, R. (1998). *The ar face database: Cvc Technical Report 24*. Research Report.

McCloskey, M., and Cohen, N. J. (1989). "Catastrophic interference in connectionist networks: The sequential learning problem," in *Psychology of Learning and Motivation* (Elsevier), 109–165. doi: 10.1016/S0079-7421(08)60536-8

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data," in *Artificial Intelligence and Statistics* (PMLR), 1273–1282.

Medvedev, I., Pimenta, J. A., and Gonçalves, N. (2024). "Fused classification for differential face morphing detection," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1043–1050. doi: 10.1109/WACVW60836.2024.00114

Neubert, T., Makrushin, A., Hildebrandt, M., Kraetzer, C., and Dittmann, J. (2018). Extended stirtrace benchmarking of biometric and forensic qualities of morphed face images. *IET Biometr.* 7, 325–332. doi: 10.1049/iet-bmt.2017.0147

Ojala, T., Pietikainen, M., and Harwood, D. (1994). "Performance evaluation of texture measures with classification based on kullback discrimination of distributions," in *Proceedings of 12th International Conference on Pattern Recognition* (IEEE), 582–585. doi: 10.1109/ICPR.1994.576366

Parisi, G. I., Kemker, R., Part, J. L., Kanan, C., and Wermter, S. (2019). Continual lifelong learning with neural networks: a review. *Neural Netw.* 113, 54–71. doi: 10.1016/j.neunet.2019.01.012

Patel, V. M., Gopalan, R., Li, R., and Chellappa, R. (2015). Visual domain adaptation: a survey of recent advances. *IEEE Signal Process. Mag.* 32, 53–69. doi: 10.1109/MSP.2014.2347059

Pellegrini, L., Borghi, G., Franco, A., and Maltoni, D. (2023). "Detecting morphing attacks via continual incremental training," in *Proceedings of the IEEE International Joint Conference on Biometrics (IJCB 2023)* (Ljubljana, Slovenia). doi: 10.1109/IJCB57857.2023.10449306

Phillips, P. J., Flynn, P. J., Scruggs, T., Bowyer, K. W., Chang, J., Hoffman, K., et al. (2005). "Overview of the face recognition grand challenge," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)* (IEEE), 947–954. doi: 10.1109/CVPR.2005.268

Phillips, P. J., Wechsler, H., Huang, J., and Rauss, P. J. (1998). The feret database and evaluation procedure for face-recognition algorithms. *Image Vis. Comput.* 16, 295–306. doi: 10.1016/S0262-8856(97)00070-X

Pisani, P. H., Mhenni, A., Giot, R., Cherrier, E., Poh, N., Ferreira de Carvalho, A. C. P., et al. (2019). Adaptive biometric systems: Review and perspectives. *ACM Comput. Surv.* 52, 1–38. doi: 10.1145/3344255

Raja, K., Ferrara, M., Franco, A., Spreeuwers, L., Batskos, I., de Wit, F., et al. (2020). Morphing attack detection-database, evaluation platform, and benchmarking. *IEEE Trans. Inf. Forens. Secur.* 16, 4336–4351. doi: 10.1109/TIFS.2020.3035252

Rusu, A. A., Rabinowitz, N. C., Desjardins, G., Soyer, H., Kirkpatrick, J., Kavukcuoglu, K., et al. (2016). Progressive neural networks. *arXiv preprint arXiv:1606.04671*.

Scherhag, U., Nautsch, A., Rathgeb, C., Gomez-Barrero, M., Veldhuis, R. N., Spreeuwers, L., et al. (2017). "Biometric systems under morphing attacks: assessment of morphing techniques and vulnerability reporting," in *2017 International Conference of the Biometrics Special Interest Group (BIOSIG). doi: 10.23919/BIOSIG.2017.8053499*

Scherhag, U., Rathgeb, C., and Busch, C. (2018). "Towards detection of morphed face images in electronic travel documents," in *2018 13th IAPR International Workshop on Document Analysis Systems (DAS)* (IEEE), 187–192. doi: 10.1109/DAS.2018.11

Scherhag, U., Rathgeb, C., Merkle, J., and Busch, C. (2020). Deep face representations for differential morphing attack detection. *IEEE Trans. Inf. Forens. Secur.* 15, 3625–3639. doi: 10.1109/TIFS.2020.2994750

Shiqerukaj, E., Rathgeb, C., Merkle, J., Drozdowski, P., and Tams, B. (2022). "Fusion of face demorphing and deep face representations for differential morphing attack detection," in *2022 International Conference of the Biometrics Special Interest Group (BIOSIG)* (IEEE), 1–5. doi: 10.1109/BIOSIG55365.2022.9897023

Singhal, P., Walambe, R., Ramanna, S., and Kotecha, K. (2023). Domain adaptation: challenges, methods, datasets, and applications. *IEEE Access* 11, 6973–7020. doi: 10.1109/ACCESS.2023.3237025

Van de Ven, G. M., Siegelmann, H. T., and Tolias, A. S. (2020). Brain-inspired replay for continual learning with artificial neural networks. *Nat. Commun.* 11:4069. doi: 10.1038/s41467-020-17866-2

Venkatesh, S., Zhang, H., Ramachandra, R., Raja, K., Damer, N., and Busch, C. (2020). "Can gan generated morphs threaten face recognition systems equally as landmark based morphs?-vulnerability and detection," in *2020 8th International Workshop on Biometrics and Forensics (IWBF)* (IEEE), 1–6. doi: 10.1109/IWBF49977.2020.9107970

Yoon, J., Yang, E., Lee, J., and Hwang, S. J. (2018). "Lifelong learning with dynamically expandable networks," in *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings* (OpenReview.net).

Zenke, F., Poole, B., and Ganguli, S. (2017). "Continual learning through synaptic intelligence," in *Proceedings of the 34th International Conference on Machine Learning - Volume 70, ICML'17* (JMLR.org), 3987–3995.

Zhang, H., Venkatesh, S., Ramachandra, R., Raja, K., Damer, N., and Busch, C. (2021). Mipgan—generating strong and high quality morphing attacks using identity prior driven gan. *IEEE Trans. Biometr. Behav. Identity Sci.* 3, 365–383. doi: 10.1109/TBIOM.2021.3072349