Rehearsal-Free Domain Continual Face Anti-Spoofing: Generalize More and Forget Less

Rizhao Cai[1] Yawen Cui[2] Zhi Li[3] Zitong Yu[4,5*]
Haoliang Li[6] Yongjian Hu[7] Alex Kot[1]
[1]Rapid-Rich Object Search (ROSE) Lab, NTU-PKU Joint Research Institute, Nanyang Technological University
[2]University of Oulu [3]Bytedance Ltd. [4]School of Computing and Information Technology, Great Bay University
[5] Great Bay Institute for Advanced Study [6]City University of Hong Kong [7]South China University of Technology
rzcai@ntu.edu.sg, yawen.cui@oulu.fi, zhi.li.2023@bytedance.com, yuzitong@gbu.edu.cn
haoliang.li@cityu.edu.hk, eeyjhu@scut.edu.cn, eackot@ntu.edu.sg

Abstract

Face Anti-Spoofing (FAS) is recently studied under the continual learning setting, where the FAS models are expected to evolve after encountering data from new domains. However, existing methods need extra replay buffers to store previous data for rehearsal, which becomes infeasible when previous data is unavailable because of privacy issues. In this paper, we propose the first rehearsal-free method for Domain Continual Learning (DCL) of FAS, which deals with catastrophic forgetting and unseen domain generalization problems simultaneously. For better generalization to unseen domains, we design the Dynamic Central Difference Convolutional Adapter (DCDCA) to adapt Vision Transformer (ViT) models during the continual learning sessions. To alleviate the forgetting of previous domains without using previous data, we propose the Proxy Prototype Contrastive Regularization (PPCR) to constrain the continual learning with previous domain knowledge from the proxy prototypes. Simulating practical DCL scenarios, we devise two new protocols which evaluate both generalization and anti-forgetting performance. Extensive experimental results show that our proposed method can improve the generalization performance in unseen domains and alleviate the catastrophic forgetting of previous knowledge. The code and protocol files are released on https://github.com/RizhaoCai/DCL-FAS-ICCV2023.

1. Introduction

Face recognition (FR) has been widely used in identity authentication because of its convenience. However, face recognition-based authentication systems are threatened by face spoofing attacks [48, 22, 63]. To protect FR systems from spoofing attacks, various Face Anti-Spoofing (FAS) techniques are deployed to detect spoofing faces and reject malicious attempts [24, 23]. Although recent FAS methods based on deep learning and neural networks achieve exquisite accuracy in intra-domain testing, the performance of existing methods heavily relies on the diversity of the training data and degrades severely if there are domain shifts between the training and testing data domains [26]. The cross-domain problem hence becomes the most challenging issue of state-of-the-art FAS research.

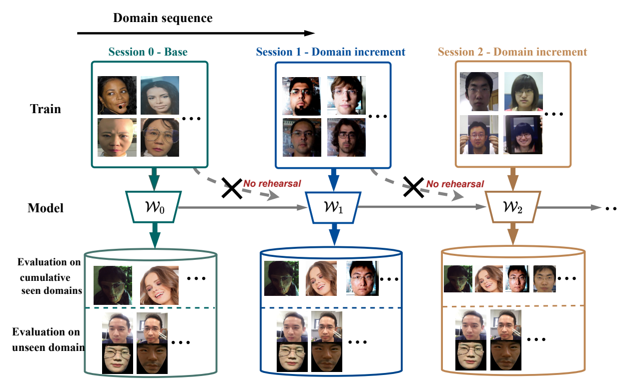

Figure 1. The rehearsal-free DCL consists of a sequence of learning sessions. The FAS model is initially trained on a large-scale base domain and then continually adapts to new domains in the following continual sessions. For each continual session, only a few data of new domain is available for training and previous data is NOT available. After the DCL, the model is tested on all previous domains and extra unseen domains.

To tackle the cross-domain problem, domain generalization [50, 16, 61] and adaption [26, 13] techniques for FAS have been extensively studied in recent years. Domain generalization-based methods aim to develop a generalized FAS model with training data from multiple source domains. Despite improving the generalization to some ex-

*Corresponding author

tent, they are still not satisfactory in unseen domains. Besides, domain adaptation-based methods utilize target domain data for model adaptation. Although the target domain performance can be significantly improved, the benefit is at the cost of expensive target data collection. Moreover, it is even impractical to collect sufficient data at a static point in time since the domain shifts are caused by constantly changing factors like illuminations and attack types.

In real-world scenarios, the deployed FAS systems constantly encounter new data from various domains. The new data will be collected and become available for model training gradually. Completely retraining a model from scratch with old and new data has both efficiency and privacy issues. Although fine-tuning the base model with the new data only is more efficient, the past knowledge will be forgotten after fine-tuning, and the performance on previous data decreases dramatically, *i.e.* catastrophic forgetting [44]. To adapt models efficiently, continual learning methods for FAS have been proposed in recent works [44, 49]. To alleviate the catastrophic forgetting, both methods [44, 49] utilize replay buffers to store previous data for rehearsal while fine-tuning with new data. However, the use of replay buffers causes extra storage burdens. Even worse, the previous data is not always available for storage and transfer since face data contains identity information.

In this work, we tackle the FAS problem under the rehearsal-free Domain Continual Learning (DCL) setting. Unlike existing work [49], where the FAS model is expected to learn sequentially from data of novel attack types, the aim of our work is to enable the FAS model to continually evolve with the data from constantly varying domains. Due to efficiency and privacy issues, previous data is not allowed to be stored and accessed for rehearsal, and only a few (low-shot) new data are available for continual learning, which are different from previous works [44, 49]. We first evaluate the baseline method under the DCL setting and obtain the below interesting observations from experiments: catastrophic forgetting usually occurs when the new coming data dataset has large domain gaps from previous ones, and a model with better performance of unseen domain generalization usually forgets less previous domain knowledge. Motivated by above the observations, we propose to address the DCL-FAS problem from the aspect of generalization.

During continual sessions, using a small amount of data may lead to overfitting, bringing poor generalization performance and catastrophic forgetting. To update models continually and efficiently, we introduce the Efficient Parameter Transfer Learning (EPTL) paradigm for the DCL-FAS and utilize Adapters [12, 13] for Vision Transformer (ViT) [9]. By using the adapters, ViT models can be efficiently adapted even with low-shot training data. However, we find that vanilla adapters consisting of linear layers cannot satisfy the need to extract fine-grained features for the FAS task. Hence, we replace the vanilla Linear Adapter with our proposed Dynamic Central Difference Convolutional Adapter (DCDCA), which empowers ViT with image-specific inductive bias by convolution and extracts fine-grain features with adaptive central difference information [65]. Unlike [65] where the ratio of central difference information is fixed for all layers, the ratio in our designed DCDCA is self-adaptive to new data domains, which is more suitable in the DCL setting. Besides, to further improve the generalization performance, we optimize DCDCA with contrastive regularization and reduce forgetting during optimization by our proposed Proxy Prototype Contrastive Regularization (PPCR). Without access to previous data, our PPCR utilizes previous data knowledge extracted from the class centroids of previous tasks, which are approximated by model weights of the fully-connected layers, instead of previous data.

Our contributions include: ***1)*** We formulate and tackle the FAS problem in a more practical scenario: low-shot and rehearsal-free Domain Continual Learning (DCL). In each continual learning session, only a few new data are available for training and no previous data is accessible; ***2)*** We design the Dynamic Central Difference Convolutional Adapter (DCDCA) to efficiently adapt ViT-based models in continual domains and capture intrinsic live/spoof cues; ***3)*** We propose the Proxy Prototype Contrastive Regularization (PPCR) to further improve the generalization and alleviate the forgetting of FAS models during rehearsal-free DCL; ***4)*** We design two practical protocols to evaluate both anti-forgetting and generalization capacities of FAS models under DCL settings, with up to 15 public datasets covering both 2D and 3D attacks. We find that the proposed DCDCA and PPCR can significantly improve generalization while forgetting less over baselines on these two DCL protocols.

2. Related works

2.1. Cross-Domain Face Anti-Spoofing

Due to the powerful representation ability of deep neural networks, deep learning based FAS methods gradually surpass and replace the traditional methods based on handcrafted features [48, 63, 22, 60]. Recently, plenty of domain generalization and adaptation techniques have been proposed to improve cross-domain FAS performance.

Domain generalization-based methods [25, 54, 34] aim to improve the generalization ability of FAS models by learning generalized feature representations with training data of multiple data domains. Li *et al.* [25] proposed to learn feature representations via domain alignment, which minimizes the distance between feature distributions of different data domains. To improve the generalization ability, disentangled representation learning techniques have been used to extract domain-invariant features [54, 56]. To deal with the biased distribution of different data do-

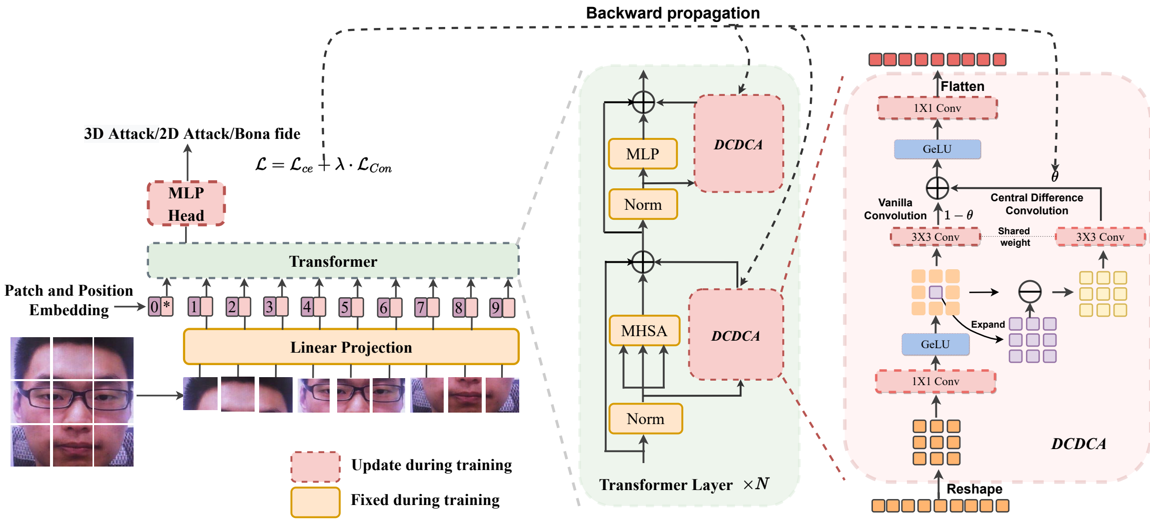

Figure 2. The architecture of the proposed ViT-DCDCA. The Dynamic Central Difference Convolutional Adapter (DCDCA) is able to extract the fine-grained central difference information for intrinsic live/spoof representation. Only the DCDCA and the MLP head are updated during training. 'MHSA' and 'MLP' denote multi-head self-attention and multi-layer perceptron, respectively.

mains, Liu *et al.* [34] proposed to re-weight the relative importance between different samples. Besides, meta-learning concepts have been extensively used in FAS to improve the generalization performance via elaborate meta-tasks [47, 51, 33, 45, 3]. Considering that manually partitioning training data into different domains is expensive and sub-optimal, Chen *et al.* [4] proposed a method for dividing training data from a mixture of data domains automatically.

Domain adaptation [26, 55, 54, 47, 13, 46, 27] based methods help further improve the performance of FAS models with target domain data for adaptation. To avoid the expense of target domain data annotation, unsupervised domain adaptation methods have been proposed to improve the cross-domain performance of FAS with some unlabelled target domain data [26, 55, 54, 36, 38]. Considering the expense of target data collection, the few-shot learning concept has been incorporated into the domain adaptation methods [47, 13]. Since the data collection of genuine face samples is easier and cheaper than spoofing ones, recent works [46, 27] proposed methods that only use genuine face samples of the target domain for one-class domain adaptation. However, these existing methods require using source domain data during the adaptation, and the forgetting of the source domain after adaptation is not examined.

2.2. Continual Learning

Due to uncontrollable factors such as the changes in the environment and the emergence of novel attack data, a FAS system will always encounter testing examples that are from unseen data distributions. The FAS systems are thereby expected to have the ability to continually learn from newly collected data and adapt themselves. Naively fine-tuning neural network models with new data usually overwrites the knowledge learned from previous data [19, 42]. Continual learning [43, 7, 6, 8] aims to alleviate catastrophic forgetting with elaborate regularization [28, 21] or memory replay [52, 18]. Recently, some continual learning frameworks for FAS have been proposed [44, 49], which alleviate the forgetting of FAS models with the help of replay buffers. However, the use of a replay buffer causes storage inefficiency and privacy issues since facial images, as biometric information, are sensitive for storage and transfer. Different from existing works [44, 49], we propose the first rehearsal-free DCL framework for FAS without using replay buffers to store previous data. Most similar works that address continual learning under the rehearsal-free setting are using prompts for ViT [57]. However, we experimentally find that the prompts for ViT can hardly forget previous or seen knowledge but are unable to generalize to new and unseen domains. Our work designs task-aware adapter modules for ViT under the continual learning framework, which addresses the catastrophic forgetting and unseen domain generalization problems at the same time. Besides, another related work based on prototypes under the continual learning setting is the PASS method [69]. However, PASS [69] utilizes self-supervised learning, but the augmentation-based proxy task (rotation prediction) in [69] is improper for the FAS task and thus has limited benefits.

3. Methodology

3.1. Dynamic Central Difference Convolutional Adapter

Given the observation from experiments that a model that generalizes well can usually have less catastrophic forgetting (see Sec. 4.2 and 4.3), to achieve the goals of DCL-FAS: generalize more and forget less, we propose Dynamic Central Difference Convolutional Adapter (DCDCA), which adapts ViT with dynamic central difference information during the continual learning.

Fine-tuning ViT with ViT-Adapter. ViT [9] consists of a stack of transformer blocks, and each block comprises a Multi-Head Self Attention (MHSA) layer and Multi-Layer Perceptron (MLP) layers to extract features. By ignoring the skip connection and Norm layer, the inference procedure can be expressed as

$$out = \text{MLP}_{\mathcal{W}}(\text{MHSA}_{\mathcal{W}}(x)), \tag{1}$$

where x is the input token, and $\mathcal{W}$ represents the parameters of the transformer, out is the output token. Although ViT has strong feature representation capability, there are a large number of parameters to update when fine-tuning the ViT to a downstream task. It usually requires a large amount of data and training time. Recent studies of parameter-efficient transfer learning (PETL) on transformers [13, 14, 17] show that inserting adapter layers is an efficient way to fine-tune ViT (ViT-Adapter). Vanilla adapters for ViT usually have extra linear layers $\mathcal{A}$ inserted into transformer layers. As such, the inference of a ViT-Adapter is expressed as

$$out = \mathcal{A}(\text{MLP}_{\mathcal{W}}(\mathcal{A}(\text{MHSA}_{\mathcal{W}}(x))). \tag{2}$$

When using a ViT model for a new downstream task, parameters of the pre-trained ViT backbone ($\mathcal{W}$) are fixed, and only the parameters of the inserted adapter layers ($\mathcal{A}$) are updated. As $\mathcal{A}$ takes up a small ratio of parameters compared to the entire ViT, PETL requires only a small amount of training data and applies to DCL-FAS where a limited new domain data is available in the continual sessions.

Fine-tuning ViT with DCDCA. Since local and fine-grain features are important for generalized FAS [65, 64], we propose the Dynamic Central Difference Convolution Adapter (DCDCA) to introduce the locality inductive bias for ViT and extract fine-grained information on top of CDC [65]. As illustrated in Figure 2, the DCDCA is embedded in the ViT backbone as a residual bottleneck connection [17, 11]. During the continual learning, only the DCDCA and the classification MLP head are updated, while the other pre-trained layers are fixed.

Specifically, 2D convolutional layers inside the DCDCA are utilized to provide the locality inductive bias. To reconstruct spatial information, the 1D flattened image token from the ViT backbone is reshaped back to a 2D structure for processing. Then, the reshaped 2D token is forwarded to a stack of convolutional layers for feature extraction. To extract features for subtle live/spoof discrimination, we use CDC to extract fine-grained contextual info from neighbor visual tokens. The output $y(p_0)$ is defined as

$$\begin{aligned} y(p_0) =& \underbrace{\theta \sum_{p_n \in \mathcal{R}} \omega(p_n) \cdot (x(p_0 + p_n) - x(p_0))}_{\text{central difference convolution}} + \\ & \underbrace{(1-\theta) \sum_{p_n \in \mathcal{R}} \omega(p_n) \cdot x(p_0 + p_n)}_{\text{vanilla convolution}}, \end{aligned} \tag{3}$$

where ω is the convolutional kernel, p_0 is the center token of a 2D token map and $\mathcal{R}$ denotes the neighbor tokens around the token p_0. θ is the ratio of central difference information, which is empirically set as 0.7 for all layers in [65]. However, using a united and fixed θ in all CDC layers is sub-optimal for DCL-FAS from two perspectives. First, the proportion of central difference information in features should be layer-specific because the semantic information and grain fineness of features are different among hierarchical layers. Second, in the continual learning scenario, the data domains change dynamically thus the contribution of central difference cues should be dynamically adapted as well. Therefore, we parameterize the θ of DCDCA as learnable variables that are self-adaptable to different layers and continual learning sessions.

To increase generalization capability, we propose to treat Eq. 3 as a type of feature transformation [53], which transforms vanilla convolution feature with a scaling factor Θ sampled from a learnable Gaussian Distribution $\text{N}(\mu, \sigma^2)$. Since the sampling would stop the gradient backward propagation, we utilize the re-parameterization skill of Gaussian distribution that

$$\Theta \sim \text{N}(\mu, \sigma^2) \iff \Theta = \mu + \sigma \cdot \epsilon, \epsilon \sim \text{N}(0, 1). \tag{4}$$

During training, μ and σ are updated to sample Θ, and we use $\theta = \text{Sigmoid}(\Theta)$ in Eq 3, where Sigmoid is used to constrain the output in $[0, 1]$. During testing, randomness is removed, and $\Theta = \mu$. We also compare our domain-aware dynamic θ estimation method with other learnable θ strategies in *Appendix*.

3.2. Proxy Prototype Contrastive Regularization

To learn more generalized models, we adopt supervised contrastive loss [20] for network optimization. Considering that the distributions of real face samples are relatively similar to each other than spoofing ones [16], all real face samples are regarded as one cluster while spoofing face samples are divided into 2D attacks and 3D mask attacks. Therefore, the loss for optimization is expressed as

$$\begin{aligned} \mathcal{L}_{Con} &= \mathcal{L}_{Con}^{\mathcal{C}^1} + \mathcal{L}_{Con}^{\mathcal{C}^2} + \mathcal{L}_{Con}^{\mathcal{C}^3}, \\ \mathcal{L}_{Con}^{\mathcal{C}^1} &= \sum_{i \in \mathcal{C}^1} \frac{-1}{|\mathcal{C}^1|} \sum_{j \in \mathcal{C}^1, j \neq i} log \frac{\exp(z_i \cdot z_j)}{\sum_{a \in \mathcal{C}^1 \cup \mathcal{C}^2 \cup \mathcal{C}^3} \exp(z_i, z_a)}, \\ \mathcal{L}_{Con}^{\mathcal{C}^2} &= \sum_{i \in \mathcal{C}^2} \frac{-1}{|\mathcal{C}^2|} \sum_{j \in \mathcal{C}^2, j \neq i} log \frac{\exp(z_i \cdot z_j)}{\sum_{a \in \mathcal{C}^1 \cup \mathcal{C}^2 \cup \mathcal{C}^3} \exp(z_i, z_a)}, \\ \mathcal{L}_{Con}^{\mathcal{C}^3} &= \sum_{i \in \mathcal{C}^3} \frac{-1}{|\mathcal{C}^3|} \sum_{j \in \mathcal{C}^3, j \neq i} log \frac{\exp(z_i \cdot z_j)}{\sum_{a \in \mathcal{C}^1 \cup \mathcal{C}^2 \cup \mathcal{C}^3} \exp(z_i, z_a)} \end{aligned} \tag{5}$$

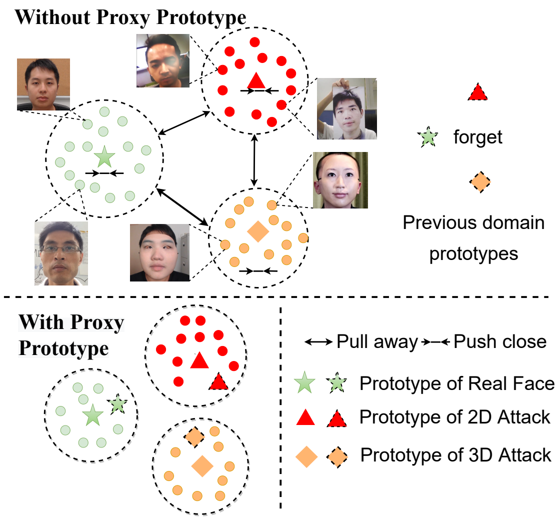

Figure 3. Illustration of Proxy Prototype Contrastive Regularization (PPCR). The top shows when learning in a new domain without prototypes, the new features might shift away from the previous prototypes, and previous knowledge is forgotten. The bottom shows our PPCR regularizes the new features to be clustered near the previous prototypes and forget less previous knowledge.

Algorithm 1: Domain Continual Face Anti-Spoofing with PPCR

An ImageNet pre-trained ViT backbone $\mathcal{W} = \{\mathcal{W}_b, \mho\}$;
Insert DCDCA modules $\mathcal{A}$ to the backbone;
Train the network on the base dataset, and only $\mathcal{A}$and $\mho$ are updated;
for $t = 1$ *to* T **do**
  Clone and detach f^1, f^2, and f^3 from $\mho$;
  while *Session t not finished* **do**
    Sample a batch of data X^b
    Conduct inference on X^b and sort out the features of samples into $\mathcal{C}^1$, $\mathcal{C}^2$, and $\mathcal{C}^3$;
    Include proxy prototypes: $\mathcal{C}^1 = \mathcal{C}^1 \cup f^1$, $\mathcal{C}^2 = \mathcal{C}^2 \cup f^2, \mathcal{C}^3 = \mathcal{C}^3 \cup f^3$;
    Calculate loss based on Eq. 6;
    Conduct backward propagation to update $\mathcal{A}$ and $\mho$
  end
end
Output: the optimized $\mathcal{A}$ and $\mho$

where $\mathcal{C}^1$, $\mathcal{C}^2$, and $\mathcal{C}^3$ denote the set of sample indices of real face, 2D attack and 3D attack examples respectively, $|\mathcal{C}^k|$ denotes the number of samples in $\mathcal{C}^k$, z_i denotes the feature from the last transformer layer of a sample i.

After the features of the same class are aligned and clustered, the clusters' centroids are set as prototypes. When continually learning with a new data domain, the FAS model would forget the previous knowledge if the features of new domain data are far away from the old prototypes, which is illustrated in the upper part of Figure 3. Recent research of source-free model transfer [29] shows that model weight can provide knowledge of the source training data and the linear classifier $\mho$ of a model is equivalent to the prototype in supervised contrastive learning [35]. Therefore, we propose the Proxy Prototype Contrastive Regularization (PPCR) to reduce forgetting during continual learning without accessing previous data. We set proxy prototypes $\mho$ as the anchors in contrastive training and regularize clustering with previous prototypes to make the previous knowledge less forgotten, as illustrated in the bottom part of Figure 3. We define the linear classifier weight as $\mho = \{f^1, f^2, f^3\}$, where f^1, f^2, and f^3 are the weights and the proxy prototypes of the classes of real face, 2D attack, and 3D attack, respectively. Then, we define the final loss for optimization as

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda \mathcal{L}_{Con}, \tag{6}$$

where $\mathcal{L}_{CE}$ is the cross-entropy loss, and λ is a constant scaling factor to balance two terms. Finally, the overall algorithm for the DCL-FAS with our proposed PPCR is described in Algorithm 1.

4. Experiment

4.1. Protocols for DCL-FAS

To establish the DCL-FAS setting, we construct two practical continual learning protocols based on the RGB data of 15 publicly available datasets: IDIAP REPLAY-ATTACK [5], CASIA-FASD [68], MSU MFSD [58], HKBU MARsV2[32], OULU-NPU[2], CSMAD[1], CASIA-SURF[66], WFFD[15], WMCA[10], CASIA-SURF 3DMASK (CASIA-3DMASK) [64], ROSE-YOUTU [26], CASIA-SURF CeFA (CeFA) [30], CelebA-Spoof [67], CASIA-SURF HiFiMask [31] and SiW [37]. We are aware that there are scenarios of using multiple modalities for the FAS task [59, 62], and our work's scope is only using the RGB modality because the RGB modality is the most commonly accessible. The details about the above datasets can be found in the *Appendix*.

Shared configuration of the starting session. We first describe the shared configuration for Protocol-1 and Protocol-2. In practical development, to train a base model (base session, $t = 0$), large-scale data including various 2D and 3D attack samples is often collected as the base dataset. As such, we combine SiW, Celeba-Spoof, and HiFiMask datasets as the base database since they contain a large amount of data. Protocol-1 and Protocol-2 share the same

Table 1. Illustration of Proctocol-1 and Protocol-2. † denotes the dataset contains 2D attack and ‡ means the dataset contains 3D attack.

Base	Celeba-Spoof†, SiW†, HiFI Mask‡									
SessionID	1	2	3	4	5	6	7	8	9	10
Protocol 1	REPLAY-ATTACK†	CASIA-FASD†	MSU MFSD†	HKBUMarV2‡	OULU-NPU†	CSMAD‡	CASIA-SURF†	WFFD‡	WMCA†‡	CASIA-3DMASK ‡
Protocol2	CASIA-3DMASk‡	WMCA†‡	WFFD‡	CASIA-SURF†	CSMAD‡	OULU-NPU†	HKBUMarV2‡	MSU MFSD†	CASIA-FASD†	REPLAY-ATTACK†
Unseen	ROSE-YOUTU†, CeFA†‡									

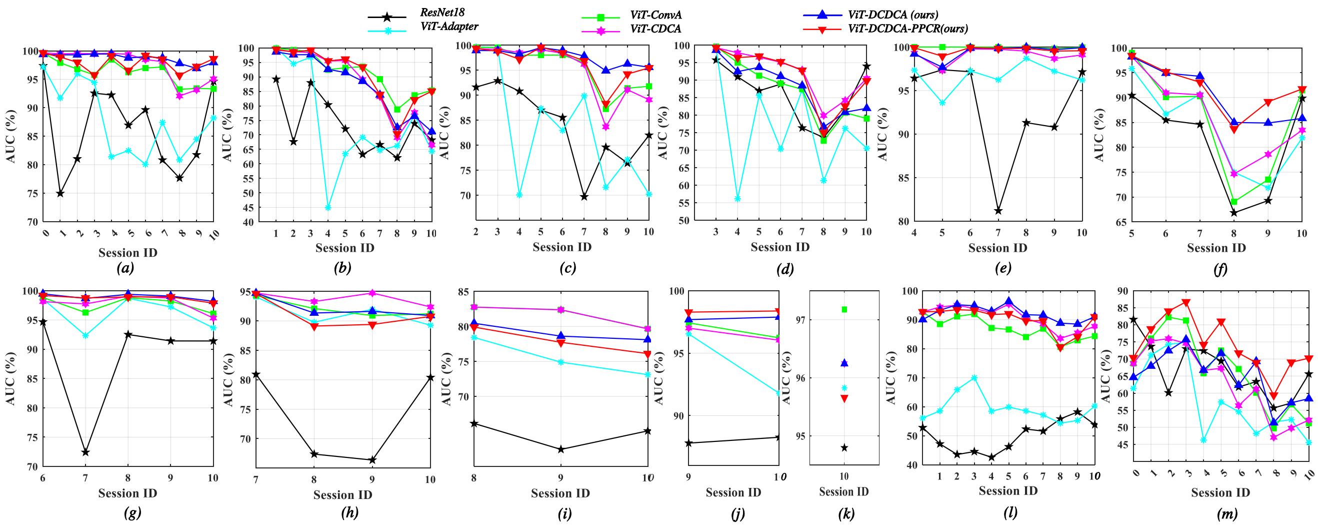

Figure 4. Performance of proposed Protocol-1 with the used architectures. (a)-(k) show the models performance on $\mathcal{D}_0$ to $\mathcal{D}_{10}$ in different sessions. (l) and (m) show the testing performance of unseen data domains ROSE-YOUTU and CeFA in different sessions.

datasets for base model training. When adapting the base model for a new domain, we consider that the model should be adapted efficiently to a small collection of data. Hence, we set up a low-shot condition in continual session $t > 0$ by randomly extracting 50 frames of real face examples and 50 frames of spoofing attack examples from the training partition of the new dataset. At the testing stage, all samples of the testing partition of the dataset are used for model evaluation.

Protocol-1 simulates the scenario where a base model is adapted as new data domains emerge. As such, from Session 1 to 10, we arrange the incoming domain sequence according to the release years of the public datasets by the ascending orders (from old to new), as shown in Table 1. **Protocol-2** simulates the scenario where a base model needs to be compatible with data from old devices. As such, Protocol-2 is set up by reverting Protocol-1 (from old to new), as shown in Table 1. For both protocols, we use ROSE-YOUTU and CeFA datasets as unseen datasets because ROSE-YOUTU has diverse 2D attack samples and CeFA includes diverse 2D and 3D attack samples. The model's unseen domain generalization performance is evaluated on these two datasets after the training of each session.

Evaluation metrics. We first introduce the notations we used before describing the evaluation metrics. The session ID is denoted by t, and $t = 0$ is the base session, and there are $T + 1$ sessions with T incremental sessions. In Session t, the training dataset is denoted as $\mathcal{D}_t$. After the training in Session t, the model status is denoted as $\mathcal{W}_t$. The model $\mathcal{W}_t$ is evaluated on the testing data of previously seen datasets $\mathcal{D}_e (e \leq t)$, and the corresponding performance is denoted as $R_{t,e}$.

Motivated by [39], we use Area Under the Receiver operating characteristic Curve (AUC) to define mean Average AUC (mAA), mean Accumulative Backward Transfer of AUC ($mABT$), and Average unseen domain Generalization AUC ($mAGA$) as

$$
\begin{aligned}
mAA &\triangleq \frac{1}{T+1}\sum_{t=0}^{T}\Big(\frac{1}{T-t+1}\sum_{i=t}^{T} R_{t,i}\Big), \\
mABT &\triangleq \frac{1}{T}\sum_{t=0}^{T-1}\Big(\frac{1}{T-t}\sum_{i=t+1}^{T} R_{i,t} - R_{t,t}\Big), \\
AGA_u &\triangleq \frac{1}{T+1}\sum_{t=0}^{T} R_t^u, \\
mAGA &\triangleq \frac{1}{2}(AGA_1 + AGA_2),
\end{aligned} \tag{7}
$$

where R_t^u means testing $\mathcal{W}_t$ on a unseen domain u, AGA_1 and AGA_2 denotes the performance on ROSE-YOUTU and CeFA respectively. The mAA (the higher, the better) measures the average intra-domain AUC performance, and $mAGA$ (the higher, the better) evaluates the average cross-domain performance in the unseen domains. Our designed $mABT$ is similar to Backward Trans-

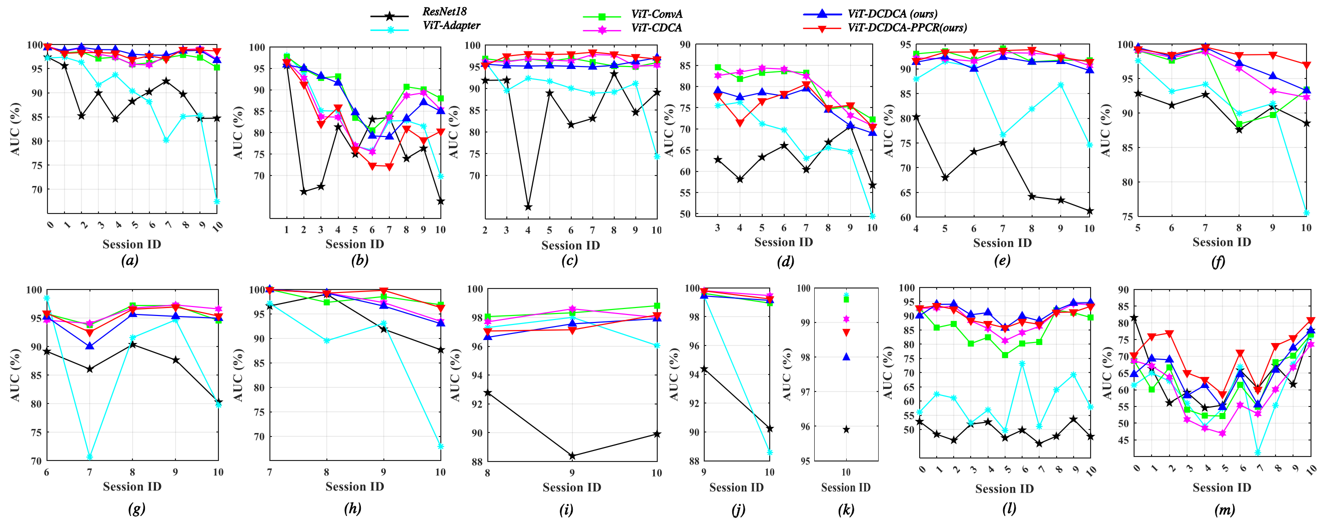

Figure 5. Performance of proposed Protocol-2 with the used architectures. (a)-(k) show the models performance on $\mathcal{D}_0$ to $\mathcal{D}_{10}$ in different sessions. (l) and (m) show the testing performance of unseen data domains ROSE-YOUTU and CeFA in different sessions.

Table 2. Compare different architectures on Protocol-1 and Protocol-2. Better performance (%) in comparison is in **bold**.

Architecture	Protocol-1			Prococol-2		
	mAA	$mABT$	$mAGA$	mAA	$mABT$	$mAGA$
ResNet18	83.02	-7.19	58.29	83.28	-7.36	56.76
ViT-Adapter	86.72	-9.87	58.72	87.93	-7.74	59.53
ViT-ConvA	**93.29**	-4.24	76.73	**94.65**	-2.28	73.79
ViT-CDCA (ours)	93.06	-4.08	77.01	94.42	-1.89	74.48
ViT-DCDCA(ours)	93.23	**-3.59**	**78.70**	93.79	**-1.82**	**78.09**

fer (BWT) [39], and high negative values of $mABT$ mean more forgetting. While BWT merely considers the final session to measure forgetting, our $mABT$ considers all intermediate sessions due to practical concerns. In practical development, it is undetermined when a new domain session $t+1$ comes and the model should be deployed after a session t. Therefore, our $mABT$ involves the results of intermediate sessions to measure the average forgetting.

Implementation details. We conduct experiments with ResNet18. Besides, we use 'vit_base' [9] as the backbone for different adapters in our work. ViT-Adapter denotes inserting vanilla Linear layers after the MHSA and MLP blocks of the ViT backbone. We insert our proposed DCDCA to ViT and derive ViT-DCDCA, as shown in Figure 2. If we fix $\theta = 0.7$, ViT-DCDCA degrades to ViT-CDCA (Central Difference Convolutional Adapter). If we fix $\theta = 0.0$, there is no central difference and ViT-CDCA becomes ViT-ConvA (Convolutional Adapter). In the training of the base model ($t = 0$), the number of the epoch is 10, the batch size is 64. In each continual session ($t > 0$), the model is trained for 100 epochs and the batch size is 20. $\lambda = 0.05$ is used for the PPCR. For all training, the Adam optimizer is used with an initial learning rate of 0.0001.

4.2. Qualitative Analysis

Figure 4 and Figure 5 show the results of Protocol-1 and Protocol-2, respectively, and all the detailed numbers can be found in the *Appendix*. In Figure 4 and Figure 5, ViT-DCDCA-PPCR means the model is trained by our PPCR algorithm, while the others are trained by using finetuning with $\mathcal{L}_{CE}$. For both figures, the sub-figures (a)-(k) show testing models on the data domains from $\mathcal{D}_0$ to $\mathcal{D}_{10}$. For example, Figure 4(a) show the results of testing models $\mathcal{W}_i$ on the base dataset $\mathcal{D}_0$, and $i \in [0, 1]$ corresponds to the x-axis of Figure 4(a). similarly, Figure 4(c) shows the results of testing models $\mathcal{W}_i$ on $\mathcal{D}_2$, and $i \in [2, 10]$. Figure 4(l) and Figure 4(m) shows the performance of testing $\mathcal{W}_i$ ($i \in [0, 10]$) on the ROSE-YOUTU and CeFA datasets respectively. Figure 5 is also organized in the same way to show the results of Protocol-2. Through experiments as shown in Figure 4 and Figure 5, we obtain the below observations.

Observation 1: Significant catastrophic forgetting usually occurs when there is a significant domain gap between a new domain and previous domains. If a new domain has shared knowledge with previous domains, previous knowledge can be recalled. To analyze the significant catastrophic forgetting, we use ViT-Adapter in Figure 4(b), (c), and (d) as examples. In Protocol-1, $\mathcal{D}_1$ (REPLAY-ATTACK), $\mathcal{D}_2$ (CASIA-FASD), and $\mathcal{D}_3$ (MSU MFSD) are all 2D attack (Photo, Replay) datasets. In Session 4, the new coming domain dataset HKBU MarV2 ($\mathcal{D}_4$) only contains 3D Mask attacks, which is significantly different from the 2D attack datasets. Therefore, if we observe the curve of vanilla ViT-Adapter in Figure 4(b), (c), and (d), after the training on $\mathcal{D}_4$(HKBU) in Session 4, the previous knowledge about REPLAY-ATTACK, CASIA-FASD, and MSU-MFSD are significantly forgotten. Therefore, in Session 4 of Figure 4(b), (c), and (d), the AUC performance dramatically drops from Session 3. Besides forgetting, previous knowledge can be recalled when the new data domain has shared

Table 3. Results of finetuning (FT) and our PPCR in optimizing models in our DCL-FAS setting. Better performance (%) in comparison is in **bold**.

Metric	Method	Protocol-1			Protocol-2		
		ViT-ConvA	ViT-CDCA	ViT-DCDCA	ViT-ConvA	ViT-CDCA	ViT-DCDCA
mAA (%)	FT	**93.29**	**93.06**	93.23	**94.65**	**94.42**	93.79
	PPCR (ours)	93.04	92.12	**93.54**	93.91	94.26	**94.03**
$mABT$ (%)	FT	-4.24	-4.08	-3.59	-2.28	-1.89	-1.82
	PPCR (ours)	**-3.08**	**-3.92**	**-3.29**	**-1.76**	**-1.25**	**-1.72**
$mAGA$ (%)	FT	76.73	77.01	78.70	73.79	74.48	78.09
	PPCR(ours)	**80.86**	**77.17**	**82.06**	**79.14**	**75.73**	**80.08**

knowledge with the data in the previous domain. For example, in Figure 4(h), the new coming dataset $\mathcal{D}_{i}$ is the CASIA-SURF dataset [66]. After Sessions 8 and 9, AUC performance (black curve) obtained by ResNet18 drops significantly. In Session 10, the AUC performance experiences a downtrend first because of the forgetting issue, and then this performance increases to that of Session 8. By analyzing the datasets $\mathcal{D}_7$, $\mathcal{D}_8$, $\mathcal{D}_9$, and $\mathcal{D}_{10}$ in Protocol-1, we find that $\mathcal{D}_7$ (CASIA-SURF [66]) and $\mathcal{D}_{10}$ (CASIA-SURF 3D Mask) mainly contain Asian subjects. In contrast, the WFFD [15] and WMCA [10] mainly contain Caucasian subjects. Thus, there is less domain shift between CASIA-SURF and CASIA-SURF 3DMask. In Session 10, the model learns on CASIA-SURF 3DMask and recalls the knowledge of CASIA-SURF. Similarly, in Protocol-2, $\mathcal{D}_1$, $\mathcal{D}_2$, $\mathcal{D}_3$, and $\mathcal{D}_4$ are the CASIA-SURF 3DMask, WMCA, WFFD and CASIA-SURF datasets respectively. As shown in Figure 5(b), the ResNet18 also forgets previous knowledge of $\mathcal{D}_1$ in Session 2 and 3, but recalls it in Session 4. Similar situations can be observed with other backbones.

Observation 2: A model that generalizes better to unseen domains usually forgets less previous domain knowledge. From Figure 4(l) and Figure 5(l), we can see the ResNet18 has much lower generalization performance on the ROSE-YOUTU dataset than our ViT-CDCA and ViT-DCDCA. Also, Figure 4(a)-(k) and Figure 5(a)-(k) show that ResNet18's performance fluctuates more than our ViT-CDCA and ViT-DCDCA, indicating that ResNet18 often forgets previous knowledge in continual learning. By contrast, the performance curve of our ViT-CDCA and ViT-DCDCA are more stable, meaning our ViT-CDCA and ViT-DCDCA forget less than ResNet18. Therefore, we observe that a more generalized model could forget less in continual learning. The insight behind this observation is intuitive. Although different FAS datasets have domain gaps between each other, there could be some shared knowledge and information. A generalized model can extract generalized knowledge across different datasets. Thus, when learning on a new dataset, the model can learn related cues about previous datasets, and thus it suffers from less catastrophic forgetting. Therefore, developing a model generalized to the unseen domain is essential for DCL-FAS.

4.3. Quantitative Analysis

Impact of generalized architectures. In this section, we analyze quantitatively the performance of finetuning different architectures in Protocol-1 and Protocol-2. As shown in Table 2, we can see that ResNet18 has the worst performance of mAA, $mABT$, and $mAGA$, indicating ResNet-18 is not as generalized as the other ViT architectures. Also, as ViT-Adapter merely uses linear layers as adapters, it lacks FAS-specific inductive bias to extract generalized FAS features. Thus, ViT-Adapter also achieves poorer generalization performance of $mAGA$ than ViT-ConvA/CDCA/DCDCA and suffers more severe catastrophic forgetting with higher negative values of $mABT$. Meanwhile, we can see the effectiveness of our ViT-DCDCA as it achieves better $mABT$ and $mAGA$ than ViT-ConvA and ViT-CDCA. Thus, our proposed method of dynamically adapting central difference cues can achieve better generalization performance on unseen domains and less forgetting in the domain continual learning setting.

Table 4. Results of our ViT-DCDCA with Proxy Prototype (PPCR) and without Proxy Prototype (CR).

Algorithm	Protoco-1			Prococol-2		
	mAA (%)	$mABT$ (%)	$mAGA$ (%)	mAA (%)	$mABT$ (%)	$mAGA$ (%)
CR	**93.71**	-3.57	81.87	**94.11**	1.77	**80.31**
PPCR	93.54	**-3.48**	**82.06**	94.03	**-1.72**	80.08

Table 5. Results of our ViT-DCDCA with PPCR, EWC, and LWF. '*' means the results are obtained by our re-implementation based on the officially released code.

Method	Protocol-1			Protocol-2		
	mAA (%)	$mABT$ (%)	$mAGA$ (%)	mAA (%)	$mABT$(%)	$mAGA$ (%)
EWC*[21]	92.35	-0.163	78.41	92.32	-0.52	78.43
LWF*[28]	91.61	**-0.97**	78.68	91.22	**0.14**	77.85
PPCR (Ours)	**93.54**	-3.48	**82.06**	**94.03**	-1.72	**80.08**

Table 6. Result (%) comparisons to state-of-the-art. '*' means the results are obtained by our re-implementation based on the officially released code.

Backbone	Method	Protocol-1			Protocol-2		
		mAA	$mABT$	$mAGA$	mAA	$mABT$	$mAGA$
ViT-base	S-Prompt*[57]	81.71	**-1.73**	51.12	82.14	**-0.65**	52.06
ViT-base	DCDCA-PASS*[69]	76.72	-5.19	63.48	72.31	-10.18	70.87
ViT-base	DCDCA-PPCR ours	**93.54**	-3.29	**82.06**	**94.03**	-1.72	**80.08**

Efficacy of PPCR. In Table 3, we compare models using finetuning (FT) or the proposed PPCR in continual learning. With different architectures, our PPCR can achieve significantly better $mABT$ and $mAGA$ than FT. Thus, our PPCR can help to generalize more and forget less than FT in DCL-FAS. In Table 4, 'CR' means using $\mathcal{L}_{Con}$ in the optimization but does not include the proxy prototypes. The performance of 'CR' and 'PPCR' are comparable in terms of $mAGA$, and PPCR forgets less with better $mABT$ than CR in both Protocol-1 and Protocol-2. Thus, the proposed proxy protocol can help reduce forgetting.

Comparison with existing continual learning methods. We implement two benchmark continual learning methods EWC [21] and LWF [28] to train our ViT-DCDCA and make comparisons to ViT-DCDCA with our PPCR. EWC and LWF achieve better $mABT$ than our PPCR, as they are designed specifically to forget less in continual learning. However, both EWC and LWF do not consider the general-

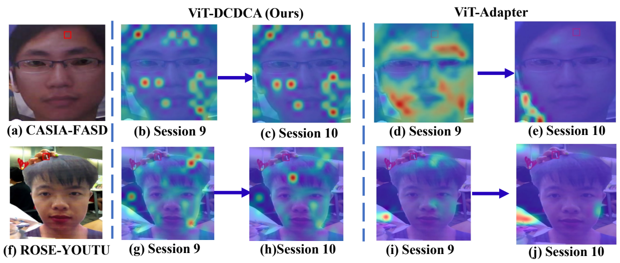

Figure 6. Visual attention maps in Protocol-2. Red means high activation. (a) shows an attack example from $\mathcal{D}_9$ (CASIA-FASD) of Protocol-2. (b) and (c) are attention maps of (a) from our ViT-DCDCA in Sessions 9 and 10, (d) and (e) are from ViT-Adapter. (f) is an attack example from the unseen ROSE-YOUTU. (g) and (h) are (f)'s attention maps from our ViT-DCDCA in Sessions 9 and 10, and (i) and (j) are from ViT-Adapter.

ization capability, and they achieve significantly less mAA and $mAGA$ performance than our PPCR.

We also compare two most similar works in Table 6. In Table 6, the backbone is all ViT-Base to make a fair comparison. We implement S-Prompt without the DCDCA, which [57] utilizes a pool of prompts to store previous knowledge. With the prompt pool, the previous knowledge can be well preserved and S-Prompt achieves the best performance of $mABT$. However, the prompt designed in [57] cannot deal with unseen domain knowledge, and thus the generalization performance ($mAGA$) is poor. On the other hand, we optimize our ViT-DCDCA model by using the PASS method [69], which conducts prototype augmentation with self-supervised learning. Since the proxy task, rotation prediction, used in the self-supervised learning is not spoofing-aware and thus the benefits from the prototype augmentation are limited for the FAS task. In summary, our proposed DCDCA and the proposed prototype-based PPCR can provide a competitive option when making a trade-off between forgetting ($mABT$) and generalization performance ($mAGA$).

4.4. Visualization and Analysis

In Figure 6, we visualize the attention maps to analyze the generalization and forgetting behavior of ViT-DCDCA and ViT-Adapter in Protocol-2. Face areas in Figure 6(b) and Figure 6(c) are highly activated. Moreover, the two attention maps are highly overlapped, meaning ViT-DCDCA's knowledge about CASIA-FASD is rarely forgotten from Session 9 to Session 10. By contrast, From Figure 6(d) to Figure 6(e), the activation maps change dramatically, meaning the ViT-Adapter's knowledge about CASIA-FASD is largely forgotten, which corresponds to Figure 5(i) that ViT-Adapter forgets much more knowledge than our ViT-DCDCA about the CASIA-FASD dataset in Session 10. On the other hand, in Figure 6(g) and Figure 6(h), the paper edges areas are activated for spoofing classification by our ViT-DCDCA. Meanwhile, in Figure 6(i) and Figure 6(j), the background areas but not the face areas are activated, which indicates the corresponds to Figure 5(l) that ViT-Adapter achieves much worse generalization performance than ViT-DCDCA.

5. Conclusion

In this paper, we raise the concern of privacy in continual FAS and formulate the FAS in rehearsal-free Domain Continual Learning settings. We devise more practical protocols for evaluation and experimentally find that models with better unseen domain generalization capability can also have less forgetting during continual learning. For better generalization performance, we develop a Dynamic Central Difference Convolutional Adapter to adapt ViT continually. Besides, we propose Proxy Prototype Contrastive Regularization to provide previous knowledge from previous model weights to reduce forgetting. Extensive experiments on two protocols demonstrate that the proposed method generalizes better in the unseen domain and forgets less previous knowledge compared to baseline methods. Beyond the face anti-spoofing task, our proposed method and protocols could be referenced by other security-related topics, such as deepfake detection[40], anti-spoofing for audio [41], and so on.

6. Acknowledement

This work was carried out at the Rapid-Rich Object Search (ROSE) Lab, Nanyang Technological University (NTU), Singapore. The research is supported in part by the Defence Science and Technology Agency, Singapore, under project agreement No. DST000ECI22000195, A*STAR under its A*STAR MBRC Strategic Positioning Fund (SPF) – ASTAR-P&G Collaboration (Award APG2013/138), and the Science and Technology Foundation of Guangzhou Huangpu Development District under Grant 2022GH15. This research is also partially supported by Sichuan Science and Technology Program (No. 2022NSFSC0551). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of the funding agencies.

References

[1] Sushil Bhattacharjee, Amir Mohammadi, and Sébastien Marcel. Spoofing deep face recognition with custom silicone masks. In *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–7. IEEE, 2018. 5

[2] Zinelabinde Boulkenafet, Jukka Komulainen, Lei Li, Xiaoyi Feng, and Abdenour Hadid. Oulu-npu: A mobile face presentation attack database with real-world variations. In *2017 12th IEEE International Conference on Automatic Face & Gesture recognition (FG 2017)*, pages 612–618. IEEE, 2017. 5

[3] Rizhao Cai, Zhi Li, Renjie Wan, Haoliang Li, Yongjian Hu, and Alex C. Kot. Learning Meta Pattern for Face Anti-

Spoofing. *IEEE Transactions on Information Forensics and Security*, 17:1201–1213, 2022. 3

[4] Zhihong Chen, Taiping Yao, Kekai Sheng, Shouhong Ding, Ying Tai, Jilin Li, Feiyue Huang, and Xinyu Jin. Generalizable representation learning for mixture domain face anti-spoofing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1132–1139, 2021. 3

[5] Ivana Chingovska, André Anjos, and Sébastien Marcel. On the effectiveness of local binary patterns in face anti-spoofing. In *2012 BIOSIG-proceedings of the international conference of biometrics special interest group (BIOSIG)*, pages 1–7. IEEE, 2012. 5

[6] Yawen Cui, Wanxia Deng, Haoyu Chen, and Li Liu. Uncertainty-aware distillation for semi-supervised few-shot class-incremental learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2023. 3

[7] Yawen Cui, Wanxia Deng, Xin Xu, Zhen Liu, Zhong Liu, Matti Pietikäinen, and Li Liu. Uncertainty-guided semi-supervised few-shot class-incremental learning with knowledge distillation. *IEEE Transactions on Multimedia*, 2022. 3

[8] Yawen Cui, Zitong Yu, Rizhao Cai, Xuna Wang, Alex Chichung Kot, and Li Liu. Generalized few-shot continual learning with contrastive mixture of adapters. *ArXiv*, abs/2302.05936, 2023. 3

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2, 4, 7

[10] Anjith George, Zohreh Mostaani, David Geissenbuhler, Olegs Nikisins, André Anjos, and Sébastien Marcel. Biometric face presentation attack detection with multi-channel convolutional neural network. *IEEE Transactions on Information Forensics and Security*, 15:42–55, 2019. 5, 8

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 4

[12] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR, 2019. 2

[13] Hsin-Ping Huang, Deqing Sun, Yaojie Liu, Wen-Sheng Chu, Taihong Xiao, Jinwei Yuan, Hartwig Adam, and Ming-Hsuan Yang. Adaptive Transformers for Robust Few-shot Cross-domain Face Anti-spoofing. In *European Conference on Computer Vision*, 2022. 1, 2, 3, 4

[14] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. *arXiv preprint arXiv:2203.12119*, 2022. 4

[15] Shan Jia, Chuanbo Hu, Guodong Guo, and Zhengquan Xu. A database for face presentation attack using wax figure faces. In *International Conference on Image Analysis and Processing*, pages 39–47. Springer, 2019. 5, 8

[16] Yunpei Jia, Jie Zhang, Shiguang Shan, and Xilin Chen. Single-Side Domain Generalization for Face Anti-Spoofing. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8481–8490, 2020. 1, 4

[17] Shibo Jie and Zhi-Hong Deng. Convolutional bypasses are better vision transformer adapters. *arXiv preprint arXiv:2207.07039*, 2022. 4

[18] Ronald Kemker and Christopher Kanan. Fearnet: Brain-inspired model for incremental learning. *arXiv preprint arXiv:1711.10563*, 2017. 3

[19] Ronald Kemker, Marc McClure, Angelina Abitino, Tyler Hayes, and Christopher Kanan. Measuring catastrophic forgetting in neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018. 3

[20] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673, 2020. 4

[21] 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. 3, 8

[22] Chenqi Kong, Shiqi Wang, and Haoliang Li. Digital and physical face attacks: Reviewing and one step further. *arXiv preprint arXiv:2209.14692*, 2022. 1, 2

[23] Chenqi Kong, Kexin Zheng, Yibing Liu, Shiqi Wang, Anderson Rocha, and Haoliang Li. M3fas: An accurate and robust multimodal mobile face anti-spoofing system. *arXiv preprint arXiv:2301.12831*, 2023. 1

[24] Chenqi Kong, Kexin Zheng, Shiqi Wang, Anderson Rocha, and Haoliang Li. Beyond the pixel world: A novel acoustic-based face anti-spoofing system for smartphones. *IEEE Transactions on Information Forensics and Security*, 17:3238–3253, 2022. 1

[25] Haoliang Li, Peisong He, Shiqi Wang, Anderson Rocha, Xinghao Jiang, and Alex C Kot. Learning generalized deep feature representation for face anti-spoofing. *IEEE Transactions on Information Forensics and Security*, 13(10):2639–2652, 2018. 2

[26] Haoliang Li, Wen Li, Hong Cao, Shiqi Wang, Feiyue Huang, and Alex C Kot. Unsupervised domain adaptation for face anti-spoofing. *IEEE Transactions on Information Forensics and Security*, 13(7):1794–1809, 2018. 1, 3, 5

[27] Zhi Li, Rizhao Cai, Haoliang Li, Kwok-Yan Lam, Yongjian Hu, and Alex C Kot. One-class knowledge distillation for face presentation attack detection. *IEEE Transactions on Information Forensics and Security*, 2022. 3

[28] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947, 2017. 3, 8

[29] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International Conference on Machine Learning*, pages 6028–6039. PMLR, 2020. 5

[30] Ajian Liu, Zichang Tan, Jun Wan, Sergio Escalera, Guodong Guo, and Stan Z Li. CASIA-SUR CeFA: A Benchmark for Multi-Modal Cross-Ethnicity Face Anti-Spoofing. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1179–1187, 2021. 5

[31] Ajian Liu, Chenxu Zhao, Zitong Yu, Jun Wan, Anyang Su, Xing Liu, Zichang Tan, Sergio Escalera, Junliang Xing, Yanyan Liang, et al. Contrastive context-aware learning for 3d high-fidelity mask face presentation attack detection. *IEEE Transactions on Information Forensics and Security*, 17:2497–2507, 2022. 5

[32] Siqi Liu, Baoyao Yang, Pong C Yuen, and Guoying Zhao. A 3d mask face anti-spoofing database with real world variations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 100–106, 2016. 5

[33] Shubao Liu, Ke-Yue Zhang, Taiping Yao, Mingwei Bi, Shouhong Ding, Jilin Li, Feiyue Huang, and Lizhuang Ma. Adaptive normalized representation learning for generalizable face anti-spoofing. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1469–1477, 2021. 3

[34] Shubao Liu, Ke-Yue Zhang, Taiping Yao, Kekai Sheng, Shouhong Ding, Ying Tai, Jilin Li, Yuan Xie, and Lizhuang Ma. Dual reweighting domain generalization for face presentation attack detection. *arXiv preprint arXiv:2106.16128*, 2021. 2, 3

[35] Yuchen Liu, Yabo Chen, Wenrui Dai, Mengran Gou, Chun-Ting Huang, and Hongkai Xiong. Source-free domain adaptation with contrastive domain alignment and self-supervised exploration for face anti-spoofing. In *European Conference on Computer Vision*, pages 511–528. Springer, 2022. 5

[36] Yuchen Liu, Yabo Chen, Mengran Gou, Chun-Ting Huang, Yaoming Wang, Wenrui Dai, and Hongkai Xiong. Towards unsupervised domain generalization for face anti-spoofing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 3

[37] Yaojie Liu, Amin Jourabloo, and Xiaoming Liu. Learning Deep Models for Face Anti-Spoofing: Binary or Auxiliary Supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 389–398, Salt Lake City, UT, 2018. 5

[38] Yuchen Liu, Yaoming Wang, Yabo Chen, Wenrui Dai, Chenglin Li, Junni Zou, and Hongkai Xiong. Promoting semantic connectivity: Dual nearest neighbors contrastive learning for unsupervised domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3510–3519, 2023. 3

[39] David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. *Advances in Neural Information Processing Systems*, 30, 2017. 6, 7

[40] Anwei Luo, Chenqi Kong, Jiwu Huang, Yongjian Hu, Xiangui Kang, and Alex C Kot. Beyond the prior forgery knowledge: Mining critical clues for general face forgery detection. *arXiv preprint arXiv:2304.12489*, 2023. 9

[41] Anwei Luo, Enlei Li, Yongliang Liu, Xiangui Kang, and Z Jane Wang. A capsule network based approach for detection of audio spoofing attacks. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6359–6363. IEEE, 2021. 9

[42] Davide Maltoni and Vincenzo Lomonaco. Continuous learning in single-incremental-task scenarios. *Neural Networks*, 116:56–73, 2019. 3

[43] German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, 2019. 3

[44] Daniel Pérez-Cabo, David Jiménez-Cabello, Artur Costa-Pazo, and Roberto J López-Sastre. Learning to learn face-pad: a lifelong learning approach. In *2020 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–9. IEEE, 2020. 2, 3

[45] Yunxiao Qin, Zitong Yu, Longbin Yan, Zezheng Wang, Chenxu Zhao, and Zhen Lei. Meta-teacher for Face Anti-Spoofing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Early Access:1–1, 2021. 3

[46] Yunxiao Qin, Weiguo Zhang, Jingping Shi, Zezheng Wang, and Longbin Yan. One-class adaptation face anti-spoofing with loss function search. *Neurocomputing*, 417:384–395, 2020. 3

[47] Yunxiao Qin, Chenxu Zhao, Xiangyu Zhu, Zezheng Wang, Zitong Yu, Tianyu Fu, Feng Zhou, Jingping Shi, and Zhen Lei. Learning meta model for zero-and few-shot face anti-spoofing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11916–11923, 2020. 3

[48] Raghavendra Ramachandra and Christoph Busch. Presentation attack detection methods for face recognition systems: A comprehensive survey. *ACM Computing Surveys (CSUR)*, 50(1):1–37, 2017. 1, 2

[49] Mohammad Rostami, Leonidas Spinoulas, Mohamed Hussein, Joe Mathai, and Wael Abd-Almageed. Detection and continual learning of novel face presentation attacks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14851–14860, 2021. 2, 3

[50] Rui Shao, Xiangyuan Lan, Jiawei Li, and Pong C. Yuen. Multi-Adversarial Discriminative Deep Domain Generalization for Face Presentation Attack Detection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10015–10023, 2019. 1

[51] Rui Shao, Xiangyuan Lan, and Pong C Yuen. Regularized fine-grained meta face anti-spoofing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11974–11981, 2020. 3

[52] Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. *Advances in Neural Information Processing Systems*, 30, 2017. 3

[53] Hung-Yu Tseng, Hsin-Ying Lee, Jia-Bin Huang, and Ming-Hsuan Yang. Cross-domain few-shot classification via learned feature-wise transformation. In *International Conference on Learning Representations*, 2019. 4

[54] Guoqing Wang, Hu Han, Shiguang Shan, and Xilin Chen. Cross-domain face presentation attack detection via multi-domain disentangled representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6678–6687, 2020. 2, 3

[55] Guoqing Wang, Hu Han, Shiguang Shan, and Xilin Chen. Unsupervised adversarial domain adaptation for cross-domain face presentation attack detection. *IEEE Transactions on Information Forensics and Security*, 16:56–69, 2020. 3

[56] Jiong Wang, Zhou Zhao, Weike Jin, Xinyu Duan, Zhen Lei, Baoxing Huai, Yiling Wu, and Xiaofei He. Vlad-vsa: Cross-domain face presentation attack detection with vocabulary separation and adaptation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1497–1506, 2021. 2

[57] Yabin Wang, Zhiwu Huang, and Xiaopeng Hong. S-prompts learning with pre-trained transformers: An occam's razor for domain incremental learning. *Advances in Neural Information Processing Systems*, 35:5682–5695, 2022. 3, 8, 9

[58] Di Wen, Hu Han, and Anil K Jain. Face spoof detection with image distortion analysis. *IEEE Transactions on Information Forensics and Security*, 10(4):746–761, 2015. 5

[59] Zitong Yu, Rizhao Cai, Yawen Cui, Xin Liu, Yongjian Hu, and Alex Kot. Rethinking vision transformer and masked autoencoder in multimodal face anti-spoofing. *arXiv preprint arXiv:2302.05744*, 2023. 5

[60] Zitong Yu, Rizhao Cai, Zhi Li, Wenhan Yang, Jingang Shi, and Alex Chichung Kot. Benchmarking joint face spoofing and forgery detection with visual and physiological cues. *arXiv preprint arXiv:2208.05401*, 2022. 2

[61] Zitong Yu, Xiaobai Li, Xuesong Niu, Jingang Shi, and Guoying Zhao. Face anti-spoofing with human material perception. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*, pages 557–575. Springer, 2020. 1

[62] Zitong Yu, Ajian Liu, Chenxu Zhao, Kevin HM Cheng, Xu Cheng, and Guoying Zhao. Flexible-modal face anti-spoofing: A benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6345–6350, 2023. 5

[63] Zitong Yu, Yunxiao Qin, Xiaobai Li, Chenxu Zhao, Zhen Lei, and Guoying Zhao. Deep learning for face anti-spoofing: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 1, 2

[64] Zitong Yu, Jun Wan, Yunxiao Qin, Xiaobai Li, Stan Z. Li, and Guoying Zhao. NAS-FAS: Static-Dynamic Central Difference Network Search for Face Anti-Spoofing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(9):3005–3023, 2021. 4, 5

[65] Zitong Yu, Chenxu Zhao, Zezheng Wang, Yunxiao Qin, Zhuo Su, Xiaobai Li, Feng Zhou, and Guoying Zhao. Searching Central Difference Convolutional Networks for Face Anti-Spoofing. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5294–5304, 2020. 2, 4

[66] Shifeng Zhang, Ajian Liu, Jun Wan, Yanyan Liang, Guodong Guo, Sergio Escalera, Hugo Jair Escalante, and Stan Z Li. CASIA-SURF: A Large-scale Multi-modal benchmark for face anti-spoofing. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2(2):182–193, 2020. 5, 8

[67] Yuanhan Zhang, Zhenfei Yin, Yidong Li, Guojun Yin, Junjie Yan, Jing Shao, and Ziwei Liu. Celeba-spoof: Large-scale face anti-spoofing dataset with rich annotations. In *European Conference on Computer Vision (ECCV)*, 2020. 5

[68] Zhiwei Zhang, Junjie Yan, Sifei Liu, Zhen Lei, Dong Yi, and S. Z. Li. A face anti-spoofing database with diverse attacks. In *IAPR International Conference on Biometrics*, pages 26–31, 2012. 5

[69] Fei Zhu, Xu-Yao Zhang, Chuang Wang, Fei Yin, and Cheng-Lin Liu. Prototype augmentation and self-supervision for incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5871–5880, 2021. 3, 8, 9