

Generalization-Preserved Learning: Closing the Backdoor to Catastrophic Forgetting in Continual Deepfake Detection

Xueyi Zhang¹, Peiyin Zhu², Chengwei Zhang³, Zhiyuan Yan⁴, Jikang Cheng⁵,
Mingrui Lao^{1*}, Siqi Cai^{6*}, Yanming Guo¹

¹National University of Defense Technology ²National University of Singapore

³University of Chinese Academy of Sciences ⁴Peking University

⁵Wuhan University ⁶Harbin Institute of Technology, Shenzhen

zhangxy1998@nudt.edu.cn, laomingrui@vip.sina.cn, caisiqi@hit.edu.cn

Abstract

Existing continual deepfake detection methods typically treat stability (retaining previously learned forgery knowledge) and plasticity (adapting to novel forgeries) as conflicting properties, emphasizing an inherent trade-off between them, while regarding generalization to unseen forgeries as secondary. In contrast, we reframe the problem: stability and plasticity can coexist and be jointly improved through the model's inherent generalization. Specifically, we propose Generalization-Preserved Learning (GPL), a novel framework consisting of two key components: (1) Hyperbolic Visual Alignment, which introduces learnable watermarks to align incremental data with the base set in hyperbolic space, alleviating inter-task distribution shifts; (2) Generalized Gradient Projection, which prevents parameter updates that conflict with generalization constraints, ensuring new knowledge learning does not interfere with previously acquired knowledge. Notably, GPL requires neither backbone retraining nor historical data storage. Experiments conducted on four mainstream datasets (FF++, Celeb-DF v2, DFD, and DFDCP) demonstrate that GPL achieves an accuracy of 92.14%, outperforming replay-based state-of-the-art methods by 2.15%, while reducing forgetting by 2.66%. Moreover, GPL achieves an 18.38% improvement on unseen forgeries using only 1% of baseline parameters, thus presenting an efficient adaptation to continuously evolving forgery techniques.

1. Introduction

The rapid evolution of generative technologies has led to a greater variety and sophistication in facial forgery methods [4, 16, 25, 38, 52], posing significant challenges to conven-

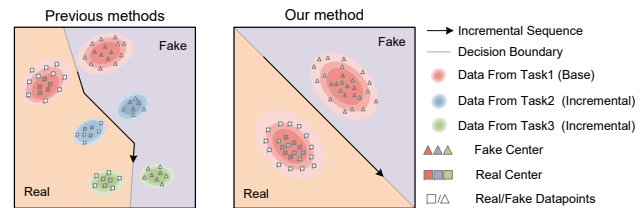


Figure 1. Comparison of inter-task data distributions between previous methods and our method in continual deepfake detection.

tional static detection methods. Continual forgery detection has emerged as a promising strategy, incrementally adjusting models to new forgery methods while preserving acquired knowledge. However, existing approaches are prone to catastrophic forgetting due to incremental distribution shifts, resulting in task recency bias—models tend to perform well on recent tasks but progressively lose their ability to detect earlier forgery types.

Traditional continual deepfake detection methods [21, 36, 41] typically rely on replay-based strategies, storing limited historical data to mitigate forgetting, raising additional storage costs and privacy risks [7, 8, 19, 40]. As illustrated in Fig. 1, combining limited replay data with current-task samples tends to form a long-tailed distribution, leading to incremental shifts in decision boundaries. These shifts progressively cause accumulative forgetting and overfitting to task-specific forgery patterns, significantly weakening the model's capacity to generalize to new forgery domains.

To this end, we propose Generalization-Preserved Learning (GPL), a framework designed from a novel perspective to simultaneously enhance model plasticity and stability. GPL comprises two key components: (1) Hyperbolic Visual Alignment (HVA) and (2) Generalized Gradient Projection (GGP). Notably, GPL requires neither retraining the backbone nor storage of historical data.

*Corresponding author.

Specifically, data distribution drift in incremental learning [29, 45, 49, 50] tends to bias models toward new tasks, causing the forgetting of prior task knowledge. Inspired by Visual Reprogramming [2, 3, 28], the first component HVA in our method introduces a lightweight watermark generator, optimizing only watermark patterns while keeping the backbone model frozen, to align incremental data visually with the base-class data. Moreover, leveraging the hierarchical embedding capabilities of hyperbolic space [15, 34], we map watermark-enhanced incremental data into the Poincaré ball, designing a contrastive loss based on hyperbolic distance to preserve geometric stability between data from both previous and current tasks. This approach mitigates data drift and exploits the exponential growth property of hyperbolic space to achieve tighter hierarchical alignment between incremental and base-class data, thereby preserving inter-task alignment and improving model generalization.

However, relying solely on data alignment is insufficient to prevent interference with previously learned parameters during incremental updates. Therefore, Generalized Gradient Projection selectively restricts parameter updates that conflict with predefined generalization boundaries in gradient optimization. Specifically, we compute the angle between gradients of the new task and base-class tasks. Gradients forming acute or right angles with base-class gradients are updated directly, whereas gradients forming obtuse angles are orthogonally projected along the base-class gradient direction. This approach prevents new-task learning from disturbing established generalization boundaries.

Our study employs backbone and incremental protocols consistent with DFIL [36] and DMP [44], conducting experiments on four mainstream datasets: FaceForensics++ [37], Celeb-DF v2 [26], DeepFake Detection [37], and DFDC-Preview [12]. Compared to state-of-the-art replay-based methods, the GPL framework achieves an average accuracy of 92.14%, surpassing current methods by 2.15%, and reduces average forgetting rate to 1.42%, a reduction of 2.66%. Furthermore, GPL utilizes only 1% of the learnable parameters compared to baseline methods, yet improves average accuracy on unseen forgery domains by 18.38%. This groundbreaking performance validates the dual advantages of the GPL framework in terms of model continual learning and generalization capability.

- **Contribution.** We propose a new perspective on continual deepfake detection, reframing stability and plasticity as mutually reinforcing rather than competing objectives.

1. We propose **Hyperbolic Visual Alignment**, a lightweight watermark-based method for aligning incremental and base-class data distributions in hyperbolic space, effectively mitigating distribution shifts.
2. We develop **Generalized Gradient Projection**, selectively blocking gradient updates that interfere with es-

tablished generalization boundaries, significantly enhancing generalization to unseen forgery domains.

3. Extensive experiments verify the state-of-the-art performance of GPL, significantly outperforming replay-based methods with fewer parameters, while markedly boosting generalization to unseen forgery domains.

2. Related Works

2.1. Generalizable Deepfake Detection

The rapid advancement of deepfake generation techniques has made detecting forgeries increasingly challenging. Early detection methods [53] relied on biological artifacts such as eye blinking and facial asymmetry, but these approaches quickly became ineffective as generative models improved. Recent works have shifted towards data-driven learning to enhance generalization. Some approaches capture domain-invariant forgery cues by disentangling forgery-specific and common features [30, 47], while others use contrastive learning to enforce robust feature representations [5, 18]. Frequency-based methods leverage spatial-frequency correlations to detect artifacts beyond dataset biases [43, 46], and latent space augmentation has been explored to expand the forgery space and mitigate overfitting [48]. *However, these methods still rely on limited seen data, making them susceptible to novel deepfake techniques [9, 42]. To address this, recent research explores continual learning paradigms as a more practical solution for adapting to evolving deepfake manipulations [13, 19].*

2.2. Continual Deepfake Detection

Research on continual deepfake detection remains limited. CoReD [21] introduces knowledge distillation [22, 23, 51] for continual representation learning, while DFIL [36] employs hard and center sample replay to retain forgery patterns. HDP [41] leverages universal adversarial perturbations (UAP) [33] to approximate historical forgery distributions, reducing storage requirements. DMP [44] proposes a dynamic prototype-based replay strategy, where multiple prototypes represent real and fake categories for incremental adaptation. SUR-LID [6] focuses on aligned feature isolation by incrementally constructing feature distributions to reduce task interference. *While these methods mitigate catastrophic forgetting to some extent, they still rely on explicit replay and construct independent decision boundaries for sequential tasks, leading to accumulative forgetting as new distributions override previous ones. Moreover, the lack of effective feature unification across tasks weakens generalization to unseen forgeries.*

3. Preliminaries

Deepfake detection is traditionally a binary classification Task, distinguishing real from fake samples. The rapid evo-

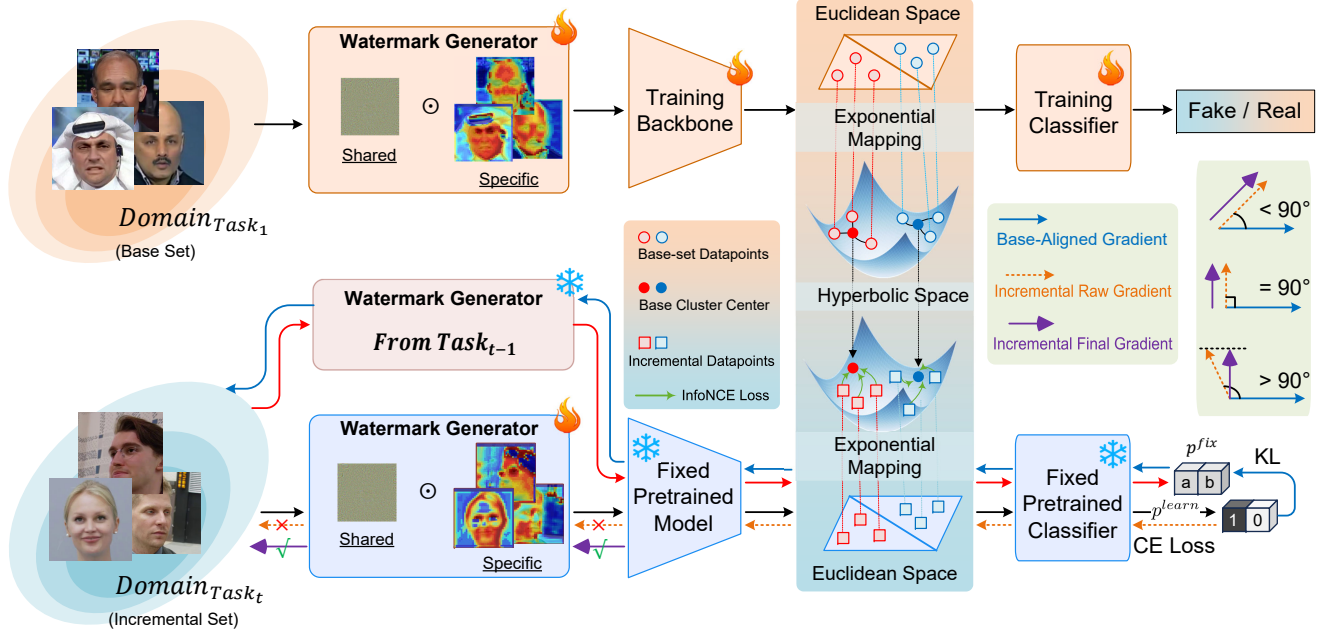


Figure 2. Overview of the Generalization-Preserved Learning framework. GPL integrates Hyperbolic Visual Alignment to align incremental data distributions with base-class data in hyperbolic space, and Generalized Gradient Projection to selectively mitigate interference with previously learned parameters during incremental updates

lution of deepfake techniques introduces novel forgery patterns, which poses a challenge to the generalization ability of pre-trained detection models on unseen data. Continual deepfake learning addresses this by incrementally training on new forgery methods, allowing adaptation to evolving distributions while retaining past knowledge.

We consider a general continual deepfake detection setup, where an incremental dataset sequence is given as $S = \{D_1, D_2, \dots, D_K\}$, where K denotes the total number of datasets. For each dataset D_t ,

$$D_t = \{X_t^{\text{train}}, Y_t^{\text{train}}, X_t^{\text{test}}, Y_t^{\text{test}}\}, \quad (1)$$

represents its corresponding training and test set partitions. Let $\{x, y\} \in D_t$ denote a sample in the dataset, where x represents the input image data, typically expressed as a three-dimensional tensor $x \in \mathbb{R}^{H \times W \times 3}$, and $y \in \{0, 1\}$ indicates whether the sample is real or fake. The model is trained sequentially across different stages.

We define the first training stage ($t = 1$) as the base-set stage, where D_1 serves as the base dataset, and subsequent stages ($t > 1$) as incremental-set stages. In real-world scenarios, the model is initially trained on a large-scale dataset in the base-set stage, representing the pre-deployment pre-training Task. However, in the following incremental-set stages, the number of available labeled deepfake samples is limited, meaning that the amount of training data in the base-set stage is significantly larger than that in the

incremental-set stages:

$$|X_t^{\text{train}}| \ll |X_1^{\text{train}}|, \quad (t > 1). \quad (2)$$

After training at each stage, the model is evaluated on a test set that includes samples from both the current and previous stages.

4. Generalization-Preserved Learning

In this paper, we propose the Generalization-Preserved Learning framework, integrating a lightweight watermark generator with Hyperbolic Visual Alignment and Generalized Gradient Projection, as illustrated in Fig. 2.

4.1. Base-Set Training

In continual deepfake detection tasks, base-set training serves as the foundation for the entire model learning process. The primary objective is to construct a robust initial classification model, initialize watermark parameters, and enable the model to learn fundamental deepfake patterns. This task is trained using the base dataset, denoted as X_1^{train} and Y_1^{train} . The forward process involves enhancing images using a watermark generator, followed by a feature extractor. The extracted features are then processed by the classi-

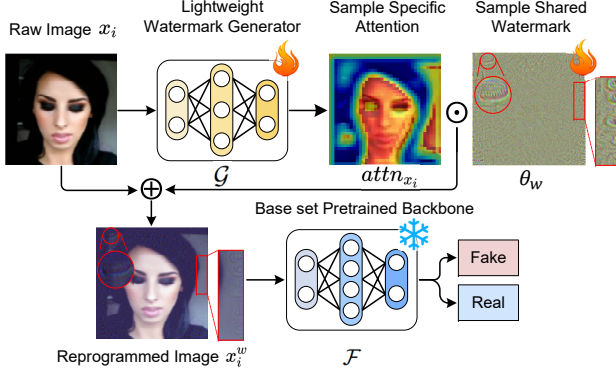


Figure 3. Visualization of the training process of the watermark generator. The raw input image is first passed through a lightweight generator \mathcal{G} , guided by sample-specific attention maps $attn_{x_i}$, and modulated with a sample-shared watermark θ_w to form a perturbation. The reprogrammed image x_i^w is obtained by overlaying the watermark and passed into the frozen backbone \mathcal{F} for deepfake classification.

fier to obtain probabilities, as formulated below:

$$x^w = \mathcal{G}(x, \theta_g^1), \quad (3)$$

$$z = \mathcal{F}(x^w, \theta_f), \quad (4)$$

$$p = \text{Softmax}(\mathcal{C}(z, \theta_c)). \quad (5)$$

Specifically, θ_f represents the pre-trained parameters of the Xception network [10] on ImageNet [11], which serve as the initialization for the feature extractor. The parameters of the watermark generator θ_g^1 and the classification head θ_c are randomly initialized. The raw image and the reprogrammed image are denoted as $x, x^w \in \mathbb{R}^{H \times W \times 3}$, while the feature extractor outputs $z \in \mathbb{R}^d$, where d is the feature dimension. The class probability $p \in \mathbb{R}^2$ is a one-dimensional vector of length equal to the number of classes. Subsequently, cross-entropy loss is used to optimize the classification task. The loss function \mathcal{L}_{cls} for N samples is defined as:

$$\mathcal{L}_{cls} = - \sum_{i=1}^N [y_i \log p_i + (1 - y_i) \log(1 - p_i)]. \quad (6)$$

The lightweight watermark generator overlays learnable perturbations on images to encourage the feature distribution of incremental data to align with that of the base-class data. This mechanism effectively mitigates the issue of inter-task data distribution drift. The watermark generator \mathcal{G} consists of a convolutional network and a globally shared watermark, with the computation process as follows:

$$f_x = \text{Conv}(x, \theta_{Conv}), \quad (7)$$

$$attn_x = U(f_x, size), \quad (8)$$

$$x^w = x + \lambda_w attn_x \odot \theta_w. \quad (9)$$

In the above formulation, \odot denotes element-wise multiplication, $size$ and λ_w is a tunable hyperparameter. $\text{Conv}(x, \theta_{Conv})$ represents a lightweight convolutional neural network, with its detailed architecture provided in Appendix A. The final watermark is obtained through element-wise multiplication of the sample-specific attention map $attn_x \in \mathbb{R}^{H \times W \times 3}$ and a trainable inter-task shared watermark $\theta_w \in \mathbb{R}^{H \times W \times 3}$. The sample-specific attention $attn_x$ adapts to different inputs to optimize watermarking for individual forged instances, while the inter-task shared watermark θ_w is optimized across incremental tasks to ensure distributional consistency between incremental and base set forgeries. The raw image is first processed by a convolutional network to extract feature maps $f_x \in \mathbb{R}^{\frac{H}{s} \times \frac{W}{s} \times 3}$. Then, a spatial nearest-neighbor upsampling operation $U(f_x, s)$ enlarges the feature map by a factor of $size = s$, generating the attention map $attn_x$ so that the watermark shares the same attention weights at the patch size s level.

After training the feature extractor on the base set dataset, we extract feature vectors from all training images, denoted as $Z_1^{\text{train}} = \mathcal{F}(\mathcal{G}(X_1^{\text{train}}, \theta_g^1), \theta_f)$. The feature sets for real and forged samples are defined as $Z_{1, \text{real}}^{\text{train}} = \{z_i \in Z_1^{\text{train}} \mid y_i = 0\}$ and $Z_{1, \text{fake}}^{\text{train}} = \{z_i \in Z_1^{\text{train}} \mid y_i = 1\}$, respectively. Using the K-means [1], we compute N_c cluster centers for both real and forged samples to preserve representative distributions for contrastive learning in incremental tasks: $C_{\text{real}} = \text{Kmeans}(Z_{1, \text{real}}^{\text{train}}, N_c)$ and $C_{\text{fake}} = \text{Kmeans}(Z_{1, \text{fake}}^{\text{train}}, N_c)$. Upon completing base set training, we obtain the pretrained parameters of the feature extractor, classifier, and watermark generator ($\theta_f, \theta_c, \theta_g^1$), along with the cluster centers $C_{\text{real}}, C_{\text{fake}} \in \mathbb{R}^{N_c \times d}$. To balance with the number of incremental task samples, we set $N_c = 25$.

4.2. Incremental-Set Learning

To mitigate forgetting accumulation caused by inter-task domain bias, we propose hyperbolic visual alignment and generalized gradient projection. During incremental training, the feature extractor and classifier parameters remain frozen, while the watermark generator follows a two-step process: first, the previously learned generator from task $t-1$ is frozen ($\theta_g^{t, \text{fix}} = \theta_g^{t-1, \text{learn}}$), and then a copy is made ($\theta_g^{t, \text{learn}} = \theta_g^{t-1, \text{learn}}$) for further learning. The learning process is illustrated in Fig. 3. Specifically, for the first incremental task, $\theta_g^{2, \text{learn}} = \theta_g^{2, \text{fix}} = \theta_g^1$. Each incremental task sample x is processed by both $\theta_g^{t, \text{fix}}$ and $\theta_g^{t, \text{learn}}$, yielding feature representations z^{fix} and z^{learn} , along with class probabilities p^{fix} and p^{learn} . Forward propagation computes the loss based on hyperbolic visual alignment and cross-entropy, while backward propagation applies gradient projection to remove gradients conflicting with the established decision boundary. Further details are provided in

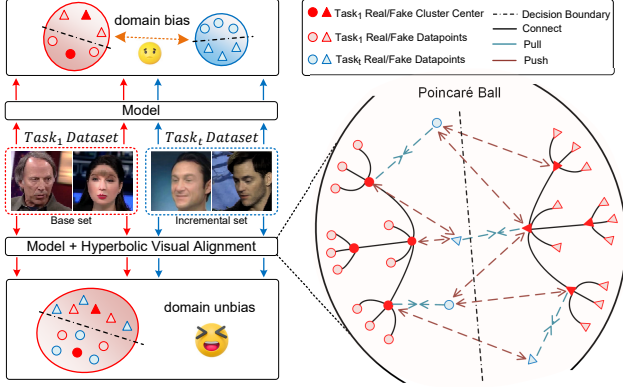


Figure 4. Hyperbolic Visual Alignment for Mitigating Domain Bias in Incremental Learning. The figure illustrates how domain bias arises between base and incremental tasks, causing misalignment in decision boundaries. By leveraging a hyperbolic space (Poincaré Ball), our method aligns features across tasks, preserving structured representations while reducing domain shifts.

the following sections.

4.2.1. Hyperbolic Visual Alignment

Forgery patterns exhibit a natural hierarchical structure based on manipulation location and extent. Traditional Euclidean space, relying on linear metrics, struggles to capture these hierarchical relationships effectively. To address this, we propose hyperbolic visual alignment as shown in Fig. 4, leveraging hyperbolic geometry on Riemannian manifolds to implicitly model hierarchical features [34]. We replace conventional Euclidean distance with hyperbolic distance $d_{\mathbb{H}}(z_1, z_2)$ [14, 31], ensuring tighter feature alignment in a hierarchical space.

$$d_{\mathbb{H}}(z_1, z_2) = \operatorname{acosh} \left(1 + 2 \frac{\|z_1 - z_2\|^2}{(1 - \|z_1\|^2)(1 - \|z_2\|^2)} \right). \quad (10)$$

In this formulation, the inverse hyperbolic cosine function is defined as $\operatorname{acosh}(x) = \ln(x + \sqrt{x^2 - 1})$, and the L2 norm is given by $\|z\|_2 = \sqrt{\sum_i z_i^2}$. The hyperbolic space implicitly models the hierarchical relationships among different forgery methods.

To guide the watermark generator in producing perturbations that facilitate inter-task alignment while preserving distinctions among different forgery types, we incorporate contrastive learning [20, 35] using precomputed cluster centers from the base set. Specifically, the real-class cluster centers C_{real} and fake-class cluster centers C_{fake} serve as positive and negative samples. For a sample feature z^{learn} from the learnable watermark generator in incremental task t , we encourage proximity to the nearest of the 25 same-class cluster centers while pushing it away from all the 25 opposite-class centers. Based on this, we define the con-

trastive loss in hyperbolic space as follows:

$$\mathcal{L}_{\text{ncc}} = -\log \frac{\exp(-d_{\mathbb{H}}(z^{\text{learn}}, C_y)/\tau)}{\sum_{C_j \in \{C_y, C_{-y}\}} \exp(-d_{\mathbb{H}}(z^{\text{learn}}, C_j)/\tau)}, \quad (11)$$

where C_y denotes the nearest same-class cluster center to z^{learn} , while C_{-y} represents the set of opposite-class cluster centers. The parameter τ serves as a temperature scaling factor.

The final training objective combines classification loss and contrastive loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} + \lambda_{\text{ncc}} \mathcal{L}_{\text{ncc}},$$

where \mathcal{L}_{cls} is the cross-entropy-based classification loss, and λ_{ncc} balances the contribution of contrastive loss in the overall optimization.

4.2.2. Generalized Gradient Projection

Relying solely on hyperbolic visual alignment is insufficient to prevent interference with previously learned parameters during incremental updates. To further address this, we introduce the Gradient Projection strategy, which constrains the direction and magnitude of gradient updates, ensuring the model adapts to new data without disrupting the established generalization boundary. Base-Aligned Gradient G_{base} is computed using p^{fix} and p^{learn} via Kullback-Leibler (KL) divergence [17], as illustrated in Fig. 2

$$\mathcal{L}_{\text{KL}} = \sum_j p^{\text{fix}}(y_j|x) \log \frac{p^{\text{learn}}(y_j|x)}{p^{\text{fix}}(y_j|x)}, \quad (12)$$

$$G_{\text{base}} = \nabla_{\theta_g^{k, \text{learn}}} \mathcal{L}_{\text{KL}}. \quad (13)$$

Incremental Raw Gradient is given as follows:

$$G_{\text{task}} = \nabla_{\theta_g^{k, \text{learn}}} \mathcal{L}_{\text{total}}. \quad (14)$$

Directly updating the model with G_{task} may result in catastrophic forgetting. To address this, we adjust the gradient direction by projecting it to better align with G_{base} , thereby preserving learned representations from previous tasks. Incremental Final Gradient is given as follows:

$$G_{\text{align}} = G_{\text{task}} - \lambda_g \frac{G_{\text{task}} \cdot G_{\text{base}}}{\|G_{\text{base}}\|^2} G_{\text{base}}, \quad (15)$$

where λ_g represents the Gradient Projection intensity coefficient. When the angle between G_{task} and G_{base} is $\leq 90^\circ$, the update is applied directly. If the angle exceeds 90° , gradient components conflicting with G_{base} are suppressed. The final parameter update for the watermark generator is formulated as:

$$\theta_g^{k, \text{learn}} = \theta_g^{k, \text{learn}} - \eta G_{\text{align}},$$

where η is the learning rate. This approach effectively regulates the magnitude of model updates, ensuring that incremental tasks do not cause severe disruptions to previously learned representations, thereby mitigating catastrophic forgetting under a replay-free mechanism.

5. Experiment

5.1. Experimental Settings

Datasets. We conduct experiments on a diverse set of widely used deepfake detection datasets, covering multiple forgery techniques to comprehensively evaluate continual learning performance. Specifically, we employ FaceForensics++ (FF++) [37], Celeb-DF v2 (CDF) [26], DeepFake Detection (DFD) [37], and DFDC-Preview (DFDCP) [12]. These datasets include various manipulation methods, spanning both face-swapping and reenactment techniques. We design three incremental task sequences:

$$\begin{aligned} \mathcal{S}_1 &= \{\text{FF++}, \text{DFDCP}, \text{DFD}, \text{CDF}\}, \\ \mathcal{S}_2 &= \{\text{FF++}, \text{DFDCP}, \text{CDF}, \text{DFD}\}, \\ \mathcal{S}_3 &= \{\text{FF++}, \text{DFD}, \text{DFDCP}, \text{CDF}\}. \end{aligned}$$

Implementation Details. For preprocessing, we follow DFIL, using RetinaFace for face detection, alignment, and cropping. All face images are resized to 299×299 . The model undergoes base-set training on FF++ for 20 epochs using Adam with a learning rate of 5×10^{-4} , decayed by 0.5 every 5 epochs. During incremental learning, our method trains for 10 epochs per dataset, decaying the learning rate every 3 epochs. We freeze the backbone and classifier, optimizing only the watermark generator (\mathcal{G}). Training is conducted with 25 real and 25 fake videos per dataset, without replay. For frame sampling, we extract 20 frames per training video and 10 per testing video.

Evaluation Protocol. To evaluate continual deepfake detection, we conduct a four-phase training process, where the model is incrementally trained on new datasets. After each phase, it is tested on the current and all previous test sets. We use three metrics: accuracy (ACC), average accuracy (AA), and average forgetting (AF). AA evaluates the overall classification performance across tasks: $AA = \frac{1}{N} \sum_{i=1}^N ACC_i$, where ACC_i is the accuracy of task i and N is the total number of tasks. AF quantifies forgetting during incremental learning: $AF = \frac{1}{N} \sum_{i=1}^N (ACC_i^{\text{first}} - ACC_i^{\text{last}})$, where ACC_i^{first} is the accuracy of task i after initial training, and ACC_i^{last} is its accuracy in the final phase.

5.2. Comparisons with Existing Methods for Continual Deepfake Detection

To comprehensively evaluate the effectiveness of our method, we compare it with two general incremental learning approaches (LwF and DGR) and three continual deepfake detection methods (DFIL, CoReD, and DMP). **All**

Table 1. Performance comparisons (ACC, AA, and AF) among existing continual deepfake detection methods. All methods use Xception as the backbone to ensure fair comparisons. The best results are highlighted in **bold**, while the second-best results are underlined.

Method	Dataset	Test Set ACC (%)				AA (%) \uparrow	AF (%) \downarrow
		FF++	DFDCP	DFD	CDF		
DGR [39]	FF++	88.86	-	-	-	88.86	-
	DFDCP	78.81	83.89	-	-	81.35	10.05
	DFD	64.31	73.31	89.69	-	75.77	17.56
	CDF	67.33	79.65	78.35	76.50	75.45	12.37
LwF [27]	FF++	95.52	-	-	-	95.52	-
	DFDCP	87.83	81.57	-	-	84.70	7.69
	DFD	76.16	41.78	96.36	-	71.43	19.89
	CDF	67.34	67.43	84.05	87.90	76.68	14.44
CoReD [21]	FF++	95.50	-	-	-	95.50	-
	DFDCP	92.94	87.61	-	-	90.28	2.56
	DFD	86.84	81.07	95.22	-	87.71	7.60
	CDF	74.08	76.59	93.41	80.78	81.21	11.42
DFIL [36]	FF++	95.67	-	-	-	<u>95.67</u>	-
	DFDCP	93.15	88.87	-	-	91.01	<u>2.52</u>
	DFD	90.83	85.42	94.67	-	90.31	4.41
	CDF	86.28	79.53	92.36	83.81	85.49	7.01
DMP [44]	FF++	95.96	-	-	-	95.96	-
	DFDCP	92.71	89.72	-	-	<u>91.22</u>	3.25
	DFD	92.64	86.09	94.84	-	<u>91.19</u>	<u>3.48</u>
	CDF	91.61	84.86	91.81	91.67	<u>89.99</u>	<u>4.08</u>
Ours	FF++	95.67	-	-	-	<u>95.67</u>	-
	DFDCP	94.06	89.85	-	-	91.95	1.61
	DFD	94.11	87.69	95.16	-	92.32	1.86
	CDF	94.19	87.32	94.90	92.16	92.14	1.42

methods adopt Xception as the backbone to ensure fair comparisons. As shown in Tab. 1, we report the test accuracy (ACC), average accuracy (AA), and average forgetting (AF) across different datasets. Our approach achieves the highest average accuracy of 92.14% and the lowest forgetting rate of 1.42% across all three datasets, highlighting its superior adaptation to incremental deepfake detection tasks.

5.3. Ablation Study

5.3.1. Overall Ablation.

To systematically analyze the effectiveness of our proposed method in continual deepfake learning, we conduct an ablation study focusing on the impact of gradient-guided pruning (GGP), hyperbolic visual alignment (HVA), and the replay set. The results shown in Tab. 2. Baseline represents training the watermark generator \mathcal{G} using only the classification loss, without HVA and GGP.

The experimental results show that the **Baseline** model suffers from severe catastrophic forgetting, especially in later incremental stages, leading to significant loss of early-

Table 2. Ablation study on the effects of GGP, HVA, and replay in continual deepfake detection. We report classification accuracy (%) on each test set and two summary metrics: AA (Average Accuracy, \uparrow) and AF (Accuracy Forgetting, \downarrow). Removing GGP or HVA degrades performance.

Method	Dataset	Test Set ACC (%)				AA (%) \uparrow	AF (%) \downarrow
		FF++	DFDCP	DFD	CDF		
Baseline	FF++	95.67	-	-	-	95.67	-
	DFDCP	92.68	77.11	-	-	84.895	2.99
	DFD	91.66	69.51	85.60	-	82.25	5.80
	CDF	92.30	67.07	80.96	75.22	78.88	6.01
w/o GGP	FF++	95.67	-	-	-	95.67	-
	DFDCP	86.80	88.73	-	-	87.76	8.87
	DFD	86.89	73.45	94.47	-	84.93	12.03
	CDF	85.47	77.64	80.39	90.91	83.60	11.79
w/o HVA	FF++	95.67	-	-	-	95.67	-
	DFDCP	92.55	80.48	-	-	86.51	3.12
	DFD	91.06	79.49	88.64	-	86.39	2.80
	CDF	90.30	77.92	83.72	79.29	82.80	4.28
Ours	FF++	95.67	-	-	-	95.67	-
	DFDCP	94.06	89.85	-	-	91.95	1.61
	DFD	94.06	87.69	95.16	-	92.32	1.86
	CDF	94.19	87.32	94.90	92.16	92.14	1.42
w/ replay	FF++	95.67	-	-	-	95.67	-
	DFDCP	94.96	89.94	-	-	92.45	0.71
	DFD	94.23	89.80	95.39	-	92.80	1.29
	CDF	94.20	87.24	95.18	92.04	92.16	1.46

learned forgery features. This indicates that classification supervision alone is insufficient for continual deepfake detection, as the model overfits new tasks while forgetting prior knowledge.

w/o GGP adapts more rapidly to new data in early incremental stages. However, as more incremental tasks are introduced, the accumulated parameter updates exacerbate forgetting, leading to performance degradation. In contrast, **w/o HVA** maintains relatively stable performance over long-term learning, exhibiting a lower forgetting rate on previous tasks. Nevertheless, the absence of contrastive learning prevents effective feature alignment across incremental tasks, resulting in reduced recognition performance for newly introduced forgeries.

The full method (**Ours**) that integrates HVA and GGP achieves better performance. The additional benefits of the **replay** in this setting are limited, further confirming that our method can completely abandon replay while maintaining excellent plasticity for new tasks and stability for previous tasks.

5.3.2. Hyperbolic Space vs. Euclidean Space

By replacing hyperbolic space with Euclidean space, the results in Tab. 3 show a decrease in AA and an increase in

Table 3. Comparison of hyperbolic and Euclidean representation learning.

Method	Dataset	Test Set ACC (%)				AA (%) \uparrow	AF (%) \downarrow
		FF++	DFDCP	DFD	CDF		
Hyperbolic	FF++	95.67	-	-	-	95.67	-
	DFDCP	94.06	89.85	-	-	91.95	1.61
	DFD	94.11	87.69	95.16	-	92.32	1.86
	CDF	94.19	87.32	94.90	92.16	92.14	1.42
Euclidean	FF++	95.67	-	-	-	95.67	-
	DFDCP	92.16	88.57	-	-	90.36	3.51
	DFD	91.98	87.11	91.36	-	90.15	2.575
	CDF	91.59	85.92	89.47	88.25	88.80	2.87

Table 4. Comparison of model performance under different training orders \mathcal{S}_1 , \mathcal{S}_2 , and \mathcal{S}_3 .

Method	Training Order	Test Set ACC (%)				AA (%) \uparrow	AF (%) \downarrow
		D ₁	D ₂	D ₃	D ₄		
DFIL	\mathcal{S}_1	86.28	79.53	92.36	83.81	85.49	7.01
	\mathcal{S}_2	88.95	83.62	86.26	92.51	87.10	4.24
	\mathcal{S}_3	89.61	95.08	85.81	84.14	88.66	1.70
Ours	\mathcal{S}_1	94.19	87.32	94.90	92.16	92.14	1.42
	\mathcal{S}_2	89.91	88.02	88.24	93.59	89.94	3.41
	\mathcal{S}_3	89.76	95.70	90.11	88.92	91.12	1.76

AF. This is because that hyperbolic space provides a natural geometric structure for capturing hierarchical relationships, making it more suitable for modeling data distributions.

5.3.3. Impact of Training Order

We alter the sequence of incremental datasets and analyze its effect. Tab. 4 shows that while performance varies across different training orders, our proposed method consistently achieves the highest accuracy across all sequences.

5.3.4. Evaluation on Unseen Deepfake Datasets

To further evaluate the generalization capability of our method, we train only on base set and the first incremental dataset (DFDCP) and directly test on unseen datasets (DFD and CDF). As shown in Tab. 5, our method significantly outperforms DFIL in unseen datasets testing, improving accuracy by 18.96% on DFD and 17.80% on CDF. Notably, during the incremental phase, our approach requires only 1% of the trainable parameters compared to DFIL.

To further analyze the impact of each component, we conduct an ablation study by removing GGP and HVA. The results indicate that removing GGP results in a 6.02% drop in average accuracy, while removing HVA leads to a 8.36% reduction. This demonstrates that GGP plays a crucial role in preventing catastrophic forgetting, while HVA effectively aligns feature distributions, enhancing generalization to unseen deepfake datasets.

Table 5. Comparison of generalization performance between DFIL and our method on unseen deepfake datasets.

Method	Unseen Datasets Test ACC (%)		Avg. ACC (%)
	DFD	CDF	
DFIL	61.71	67.21	64.46
Ours	80.67	85.01	82.84
w/o GGP	73.76	79.89	76.82
w/o HVA	72.11	76.85	74.48

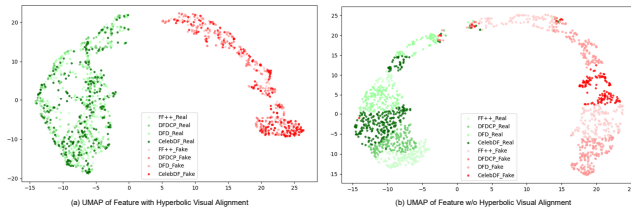


Figure 5. UMAP latent-space visualization illustrating the impact of Hyperbolic Visual Alignment \mathcal{G} on incremental data distributions

5.4. Visualization and Qualitative Analysis

UMAP latent-space visualization. We randomly sample 200 real and 200 fake face images from the FF++, DFDCP, DFD, and CDF datasets and use UMAP [32] to visualize their feature representations in a lower-dimensional space after incremental training. As shown in Fig. 5, the results demonstrate the impact of hyperbolic visual alignment (HVA) on the watermark generator \mathcal{G} in shaping data distributions. Without HVA, the distributions of deepfake samples from incremental datasets exhibit noticeable deviations from the base-set dataset, leading to domain misalignment. In contrast, applying HVA to \mathcal{G} enables incremental samples to align more effectively with the base-set distribution. This alignment suggests that the watermarking strategy guided by \mathcal{G} preserves decision boundaries across tasks, reducing domain shifts and minimizing catastrophic forgetting while maintaining detection efficacy for novel forgeries.

Visualization of Loss Landscape. Previous studies [24] have demonstrated that the flatness of the loss landscape reflects a model’s sensitivity to noise. Flat minima often correspond to improved generalization performance, as they indicate a larger region in parameter space where the loss remains consistently low. This characteristic allows the model to better tolerate distribution shifts or minor perturbations in unseen domains. Conversely, sharp minima suggest high sensitivity to parameter perturbations, indicating that while the model may achieve low loss on training data, it is less robust to distributional changes, such as domain shifts or noise. In the Appendix C, we provide a detailed explanation of the loss landscape.

As shown in Fig. 6, we compare the loss landscapes of

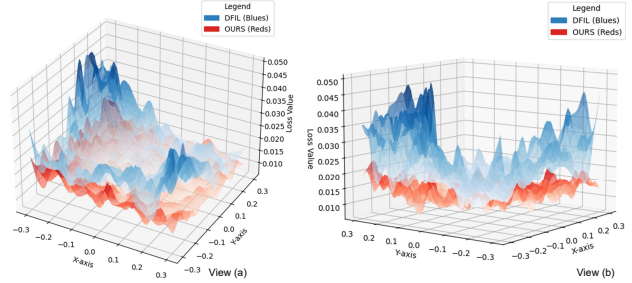


Figure 6. Loss Landscapes visualization comparing DFIL and our method. The X and Y axes represent small perturbations in the model parameter space, while the Z axis denotes the loss value, measuring the model’s error at each point. To visualize the loss landscape, we select the optimized parameter θ^* as the reference and sample two orthogonal direction vectors d_1 and d_2 in the high-dimensional parameter space. The new parameter points are then defined as $\theta(x, y) = \theta^* + xd_1 + yd_2$, and their corresponding loss values $L(x, y) = L(\theta^* + xd_1 + yd_2)$ are computed, forming the loss surface representation.

DFIL (blue) and our method (red). The fluctuating loss surface of DFIL indicates lower robustness to noise and distribution shifts. In contrast, our method exhibits a flatter loss surface, suggesting greater resilience to parameter perturbations and variations in data distribution.

6. Conclusion

In this paper, we propose Generalization-Preserved Learning, a novel continual deepfake detection framework that simultaneously enhances stability and plasticity through model generalization. GPL eliminates the need for backbone retraining or historical data storage. By integrating Hyperbolic Visual Alignment to mitigate incremental distribution shifts and Generalized Gradient Projection to prevent interference with existing generalization boundaries, GPL not only surpasses current replay-based methods in terms of accuracy and forgetting rate but also achieves outstanding detection performance on unseen forgery domains while utilizing significantly fewer parameters. Our findings highlight the crucial role of generalization preservation in continual deepfake detection.

Acknowledgements

This work was supported by National Natural Science Foundation of China (Grant No. 62271432), Shenzhen Science and Technology Program (Shenzhen Key Laboratory Grant No. ZDSYS20230626091302006), and Shenzhen Science and Technology Research Fund (Fundamental Research Key Project Grant No. JCYJ20220818103001002). This work is also supported by the Foundation of NUDT (Grant No. 25-ZZCX-JDZ-39 and ZK24-27).

References

- [1] David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. Technical report, Stanford, 2006. 4
- [2] Chengyi Cai, Zesheng Ye, Lei Feng, Jianzhong Qi, and Feng Liu. Bayesian-guided label mapping for visual reprogramming. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 2
- [3] Chengyi Cai, Zesheng Ye, Lei Feng, Jianzhong Qi, and Feng Liu. Sample-specific masks for visual reprogramming-based prompting. *arXiv preprint arXiv:2406.03150*, 2024. 2
- [4] Junyi Cao, Chao Ma, Taiping Yao, Shen Chen, Shouhong Ding, and Xiaokang Yang. End-to-end reconstruction-classification learning for face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4113–4122, 2022. 1
- [5] Liang Chen, Yong Zhang, Yibing Song, Lingqiao Liu, and Jue Wang. Self-supervised learning of adversarial example: Towards good generalizations for deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18710–18719, 2022. 2
- [6] Jikang Cheng, Zhiyuan Yan, Ying Zhang, Li Hao, Jiaxin Ai, Qin Zou, Chen Li, and Zhongyuan Wang. Stacking brick by brick: Aligned feature isolation for incremental face forgery detection. *arXiv preprint arXiv:2411.11396*, 2024. 2
- [7] Jikang Cheng, Zhiyuan Yan, Ying Zhang, Yuhao Luo, Zhongyuan Wang, and Chen Li. Can we leave deepfake data behind in training deepfake detector? *arXiv preprint arXiv:2408.17052*, 2024. 1
- [8] Jikang Cheng, Ying Zhang, Qin Zou, Zhiyuan Yan, Chao Liang, Zhongyuan Wang, and Chen Li. Ed 4: Explicit data-level debiasing for deepfake detection. *arXiv preprint arXiv:2408.06779*, 2024. 1
- [9] Jongwook Choi, Taehoon Kim, Yonghyun Jeong, Seungryul Baek, and Jongwon Choi. Exploiting style latent flows for generalizing deepfake video detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1133–1143, 2024. 2
- [10] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017. 4
- [11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 4
- [12] Brian Dolhansky, Russ Howes, Ben Pflaum, Nicole Baram, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) preview dataset. *arXiv preprint arXiv:1910.08854*, 2019. 2, 6
- [13] Shichao Dong, Jin Wang, Renhe Ji, Jiajun Liang, Haoqiang Fan, and Zheng Ge. Implicit identity leakage: The stumbling block to improving deepfake detection generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3994–4004, 2023. 2
- [14] Octavian Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic neural networks. *Advances in neural information processing systems*, 31, 2018. 5
- [15] Songwei Ge, Shlok Mishra, Simon Kornblith, Chun-Liang Li, and David Jacobs. Hyperbolic contrastive learning for visual representations beyond objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6840–6849, 2023. 2
- [16] Alexandros Haliassos, Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Lips don't lie: A generalisable and robust approach to face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5039–5049, 2021. 1
- [17] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 5
- [18] Cheng-Yao Hong, Yen-Chi Hsu, and Tyng-Luh Liu. Contrastive learning for deepfake classification and localization via multi-label ranking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17627–17637, 2024. 2
- [19] Baojin Huang, Zhongyuan Wang, Jifan Yang, Jiaxin Ai, Qin Zou, Qian Wang, and Dengpan Ye. Implicit identity driven deepfake face swapping detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4490–4499, 2023. 1, 2
- [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. 5
- [21] Minha Kim, Shahroz Tariq, and Simon S Woo. Cored: Generalizing fake media detection with continual representation using distillation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 337–346, 2021. 1, 2, 6
- [22] Mingrui Lao, Yanming Guo, Yu Liu, Wei Chen, Nan Pu, and Michael S Lew. From superficial to deep: Language bias driven curriculum learning for visual

- question answering. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3370–3379, 2021. 2
- [23] Mingrui Lao, Zheng Li, Yanming Guo, Xueyi Zhang, Siqi Cai, Zhaoyun Ding, and Haizhou Li. Boosting discriminability for robust multimodal entity linking with visual modality missing. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 989–999, 2025. 2
- [24] Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. *Advances in neural information processing systems*, 31, 2018. 8
- [25] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face x-ray for more general face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5001–5010, 2020. 1
- [26] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3207–3216, 2020. 2, 6
- [27] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017. 6
- [28] Zheng Li, Yibing Song, Penghai Zhao, Ming-Ming Cheng, Xiang Li, and Jian Yang. Atprompt: Textual prompt learning with embedded attributes. *arXiv preprint arXiv:2412.09442*, 2024. 2
- [29] Zheng Li, Xueyi Zhang, Yanming Guo, Siqi Cai, and Mingrui Lao. Pencil: Prototype-enhanced compositional learning for class-incremental hand gesture recognition. *IEEE Transactions on Consumer Electronics*, 2025. 2
- [30] Li Lin, Xinan He, Yan Ju, Xin Wang, Feng Ding, and Shu Hu. Preserving fairness generalization in deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16815–16825, 2024. 2
- [31] Shaoteng Liu, Jingjing Chen, Liangming Pan, Chong-Wah Ngo, Tat-Seng Chua, and Yu-Gang Jiang. Hyperbolic visual embedding learning for zero-shot recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9273–9281, 2020. 5
- [32] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018. 8
- [33] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1765–1773, 2017. 2
- [34] Maximillian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. *Advances in neural information processing systems*, 30, 2017. 2, 5
- [35] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 5
- [36] Kun Pan, Yifang Yin, Yao Wei, Feng Lin, Zhongjie Ba, Zhenguang Liu, Zhibo Wang, Lorenzo Cavallaro, and Kui Ren. Dfil: Deepfake incremental learning by exploiting domain-invariant forgery clues. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 8035–8046, 2023. 1, 2, 6
- [37] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1–11, 2019. 2, 6
- [38] Rui Shao, Tianxing Wu, Liqiang Nie, and Ziwei Liu. Deepfake-adapter: Dual-level adapter for deepfake detection. *International Journal of Computer Vision*, 133(6):3613–3628, 2025. 1
- [39] Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. *Advances in neural information processing systems*, 30, 2017. 6
- [40] Kaede Shiohara and Toshihiko Yamasaki. Detecting deepfakes with self-blended images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18720–18729, 2022. 1
- [41] Ke Sun, Shen Chen, Taiping Yao, Xiaoshuai Sun, Shouhong Ding, and Rongrong Ji. Continual face forgery detection via historical distribution preserving. *International Journal of Computer Vision*, pages 1–18, 2024. 1, 2
- [42] Zhimin Sun, Shen Chen, Taiping Yao, Bangjie Yin, Ran Yi, Shouhong Ding, and Lizhuang Ma. Contrastive pseudo learning for open-world deepfake attribution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20882–20892, 2023. 2
- [43] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28130–28139, 2024. 2

- [44] Jiahe Tian, Cai Yu, Xi Wang, Peng Chen, Zihao Xiao, Jizhong Han, and Yesheng Chai. Dynamic mixed-prototype model for incremental deepfake detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 8129–8138, 2024. [2](#), [6](#)
- [45] Songsong Tian, Lusi Li, Weijun Li, Hang Ran, Xin Ning, and Prayag Tiwari. A survey on few-shot class-incremental learning. *Neural Networks*, 169:307–324, 2024. [2](#)
- [46] Yuan Wang, Kun Yu, Chen Chen, Xiyuan Hu, and Silong Peng. Dynamic graph learning with content-guided spatial-frequency relation reasoning for deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7278–7287, 2023. [2](#)
- [47] Zhiyuan Yan, Yong Zhang, Yanbo Fan, and Baoyuan Wu. Ucf: Uncovering common features for generalizable deepfake detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22412–22423, 2023. [2](#)
- [48] Zhiyuan Yan, Yuhao Luo, Siwei Lyu, Qingshan Liu, and Baoyuan Wu. Transcending forgery specificity with latent space augmentation for generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8984–8994, 2024. [2](#)
- [49] Lu Yu, Bartłomiej Twardowski, Xialei Liu, Luis Heranz, Kai Wang, Yongmei Cheng, Shangling Jui, and Joost van de Weijer. Semantic drift compensation for class-incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6982–6991, 2020. [2](#)
- [50] Xianghu Yue, Yiming Chen, Xueyi Zhang, Xiaoxue Gao, Mengling Feng, Mingrui Lao, Huiping Zhuang, and Haizhou Li. Pal: Prompting analytic learning with missing modality for multi-modal class-incremental learning. *arXiv preprint arXiv:2501.09352*, 2025. [2](#)
- [51] Xueyi Zhang, Chengwei Zhang, Tao Wang, Jun Tang, Songyang Lao, and Haizhou Li. Slow-fast time parameter aggregation network for class-incremental lip reading. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 747–756, 2023. [2](#)
- [52] Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu. Multi-attentional deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2185–2194, 2021. [1](#)
- [53] Yipin Zhou and Ser-Nam Lim. Joint audio-visual deepfake detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 14800–14809, 2021. [2](#)