Preserving Knowledge across Space and Time for Continual Video Deepfake Detection

Taehoon Kim[1⋆], Jongwook Choi[1⋆], Heejae Jo[2], Byungmin Park[1], and Jongwon Choi[1,2,3†]

[1] GS. of AI, Chung-Ang University, Republic of Korea
[2] Dept. of Advanced Imaging, GSAIM, Chung-Ang University, Republic of Korea
[3] Dept. of Metaverse Convergence, Chung-Ang University, Republic of Korea
{kimth,cjw,jzoe,byungmin_park}@vilab.cau.ac.kr, choijw@cau.ac.kr

Abstract. The continuous emergence of high-quality video deepfakes requires detectors that continually adapt to new forgery patterns, yet existing approaches, which are designed for deepfake images, fail to capture video-specific cues. Unlike deepfake images that contain only spatial artifacts, deepfake videos leave distinct evidence along both spatial and temporal axes, necessitating the separate preservation of each modality during sequential model updates. To overcome this limitation, we introduce a continual deepfake video detection framework, Modality-Specific Frequency Distillation (MSFD), that explicitly decomposes video features into spatial, temporal, and spatiotemporal modalities in the frequency domain. This decomposition enables independent preservation of each modality, as different deepfake video types exhibit varying reliance on spatial and temporal cues across tasks. Furthermore, MSFD adopts a cross-modality decorrelation loss that encourages spatiotemporal representations to remain orthogonal to single-modality cues. Extensive experiments show that our framework achieves stronger adaptation and preserves performance more effectively than state-of-the-art methods across diverse continual deepfake video scenarios.

Keywords: Continual Learning · Deepfake Detection · Video Processing

1 Introduction

To address the continuously evolving video deepfakes, continual learning approaches have been proposed to enable detection models to adapt to newly emerging forgery techniques while preserving their ability to detect previously learned ones [4, 19, 36]. The rapid advancement of deepfake generation techniques has led to the continuous emergence of new video forgery methods, with the quality and realism of generated videos improving significantly [55]. Recent video generation approaches [1,50,61] produce high-quality video deepfakes that pose substantial challenges to deepfake detection systems trained on limited forgery types. While recent video deepfake detectors improve robustness

⋆ Equal contribution. † Corresponding author.

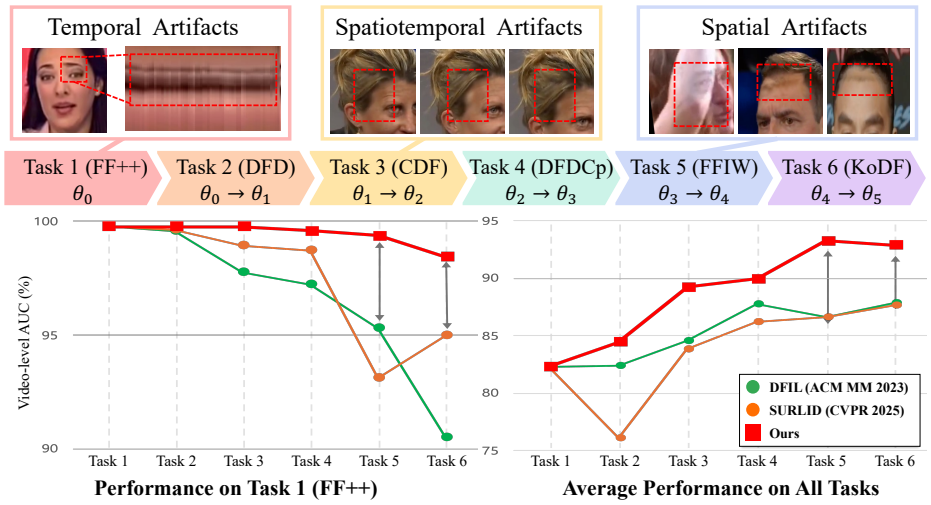
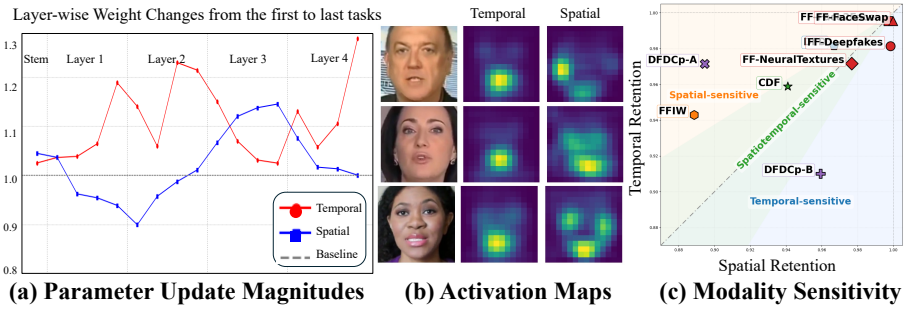


Fig. 1: Performance in the continual video deepfake detection. The top row illustrates representative temporal, spatiotemporal, and spatial artifacts across sequential tasks. The bottom row shows the performance degradation on the initial task (FF++) and average performance across all tasks as the model is continually trained on six sequential deepfake video datasets (FF++, DFD, CDF, DFDCp, FFIW, and KoDF). Our method (red) maintains the highest performance with the least degradation compared to existing continual deepfake detection approaches.

by modeling generalizable spatiotemporal representations [12, 51, 62], they are typically trained once under a fixed distribution and evaluated in a static setting. In practice, however, forgery techniques evolve after deployment and are encountered sequentially, requiring detectors to be updated over time.

However, most existing continual deepfake detection frameworks [4, 19, 32, 36, 46] were primarily designed under image-level assumptions, which limit their ability to capture video-level spatiotemporal artifacts. Deepfake videos have complementary but distinct evidence in spatial artifacts (*e.g.*, blending boundaries) and temporal inconsistencies (*e.g.*, flickering, unnatural motion) [34, 51, 62]. Such image-centric designs are therefore inadequate for videos, where deepfakes exhibit characteristics that differ from those observed in manipulated images [51, 62]. Consequently, these image-based approaches struggle when applied to video-based detection, often leading to unbalanced adaptation and severe forgetting, as shown in Fig. 1.

Our preliminary experiments further support this observation (Fig. 2). When we trained models using only spatial or only temporal cues in a continual-training setting, their layer-wise weight updates and activation responses diverged substantially (Fig. 2 (a) and (b)). Moreover, performance retention under spatial and temporal perturbations varies markedly across generation paradigms—some rely primarily on spatial cues, others on temporal cues, and some on both—even



(a) Parameter Update Magnitudes **(b) Activation Maps** **(c) Modality Sensitivity**

Fig. 2: Analysis of Spatial and Temporal Modalities. We visualize (a) parameter update magnitudes across layers relative to the iCaRL [39] baseline and (b) activation patterns when models are continually trained with either spatial-only or temporal-only cues, alongside (c) a dataset-wise sensitivity map showing the AUC retention ratio of dataset-specific finetuned models under spatial and temporal perturbations. Detailed settings are provided in the supplementary material.

within the same dataset (*e.g.*, DFDCp-A vs. DFDCp-B) (Fig. 2 (c)). These results explain why existing methods that treat spatial and temporal cues as a single entangled representation fail to retain both during continual updates, and motivate the core design of our framework: preserving spatial, temporal, and spatiotemporal representations separately.

Motivated by these observations, we develop Modality-Specific Frequency Distillation (MSFD), which decomposes features into spatial, temporal, and spatiotemporal modalities and preserves each by aligning the current model's frequency spectra with those of the previous model at each continual learning step. This design enables independent preservation of each modality, as different deepfake video types carry distinct spatial and temporal characteristics that need to be retained separately (Fig. 2 (c)). We perform this decomposition in the spectral domain because it offers a simple yet highly effective approach specifically for deepfakes. Unlike complex disentanglement in the raw feature space, frequency transforms explicitly isolate distinct forgery signatures: spatial forgeries manifest high-frequency boundary artifacts captured by 2D spatial spectra [15, 30], while point-wise 1D temporal frequency spectra capture fine-grained temporal irregularities characteristic of temporal manipulations [20]. We further employ a 3D spatiotemporal spectrum to capture joint space-time dynamics inaccessible to single-modality transforms, and introduce a cross-modality decorrelation loss that encourages the spatiotemporal representation to encode frequency patterns specific to the joint use of spatial and temporal cues while remaining orthogonal to those captured by single-modality branches.

Through extensive experiments across realistic continual deepfake video detection scenarios, including sequential domain shifts, emerging forgery techniques, and restricted data, our method shows strong adaptability while preserving prior knowledge. We further confirm its robustness through evaluations

on diverse perturbation settings, and our ablation studies validate the effectiveness of each component in the modality-specific design. Our contributions are summarized as follows:

- We introduce the first framework that addresses the preservation of spatial and temporal forgery cues for continual deepfake video detection.
- We design the Modality-Specific Frequency Distillation (MSFD) framework, which decomposes video features into three complementary frequency modalities to preserve modality-specific deepfake cues.
- We propose a Cross-Modality Decorrelation Loss (CDL) that encourages spatiotemporal frequency representations orthogonal to single-modality cues.
- We conduct comprehensive experiments across multiple continual deepfake video settings, demonstrating our proposed framework's ability to improve generalization, robustness, and forgetting mitigation.

2 Related Work

Video-based Deepfake Detection. Early deepfake detectors [6, 15, 16, 30, 43, 49, 56] focused on spatial artifacts in static images, failing to account for temporal inconsistencies that often expose video manipulations. Later studies [5, 10, 11, 20, 33, 34, 51, 53, 57, 62] showed that modeling spatiotemporal representations improves robustness to unseen forgery types. For example, FTCN [62] leverages temporal artifacts as transferable cues, while AltFreezing [51] disentangles spatial and temporal branches to stabilize detection, and recent detectors further highlight the importance of balanced spatiotemporal fusion [12, 34, 57]. Despite these advances, most detectors still assume a fixed closed-world distribution. However, new deepfake techniques are continuously evolving after deployment, necessitating detectors that can be incrementally updated over time.

Continual Deepfake Detection. Continual deepfake detection aims to adapt detectors to newly emerging forgeries while retaining performance on previously learned ones [19, 24, 36, 45, 46, 60]. Existing methods such as CoReD [19], DFIL [36], and SURLID [4] preserve prior knowledge via distillation or feature alignment, but are designed under image-level assumptions. Unlike deepfake images that contain only spatial artifacts, video deepfakes present complementary evidence along both spatial and temporal axes, whose divergent characteristics make a single entangled representation insufficient to retain during continual updates. SARB-DF [38], the only prior effort toward continual deepfake video detection, addresses only spatial cues and follows generic regularization (*e.g.*, EWC [21]) rather than deepfake-specific mechanisms. These limitations motivate our design that separately preserves spatial, temporal, and spatiotemporal forensic cues throughout continual model updates.

Continual Learning for Video Representations. Continual learning has been widely studied for video understanding, where models are sequentially adapted to evolving spatial and temporal dynamics [3, 14, 23, 28, 48]. These methods are typically designed to preserve high-level semantic representations (*e.g.*,

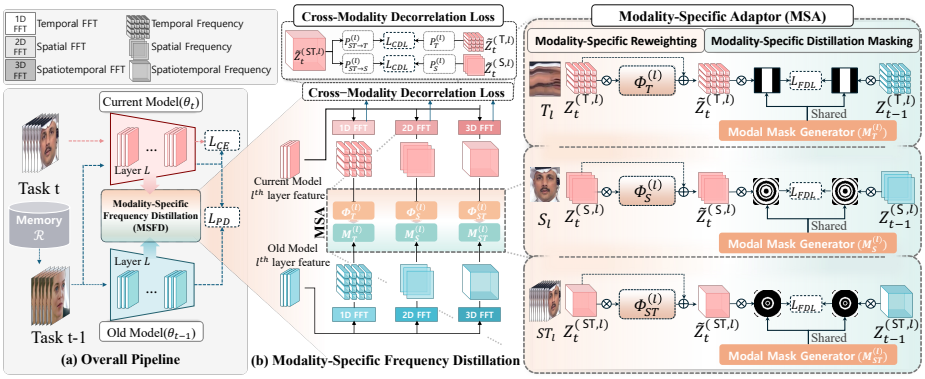


Fig. 3: Our proposed framework, Modality-Specific Frequency Distillation (MSFD) for continual deepfake video detection. (a) Overall pipeline: The current model (θ_t) is trained on task t data and memory ($\mathcal{R}$), distilling knowledge from the frozen previous model (θ_{t-1}). (b) Modality-Specific Frequency Distillation (MSFD): Intermediate features from l^{th} layer are decomposed into temporal (T), spatial (S), and spatiotemporal (ST) frequency spectra via 1D, 2D, and 3D FFTs. The Modality-Specific Adaptor (MSA) applies modality-specific reweighting ($\Phi_d^{(l)}$) and distillation masking ($M_d^{(l)}$) to compute frequency-domain distillation loss ($\mathcal{L}_{\text{FDL}}$), while the cross-modality decorrelation loss ($\mathcal{L}_{\text{CDL}}$) is computed in parallel.

playing tennis) under sequential updates. However, deepfake video detection depends on modality-sensitive, low-level forensic signals (*e.g.*, frequency artifacts and fine-grained temporal inconsistencies) that are fragile and can be easily washed out under semantics-focused objectives. As a result, directly applying conventional video continual learning techniques leads to suboptimal adaptation, motivating forensic-aware preservation targets for continual deepfake detection.

3 Proposed Method

We address continual deepfake video detection, where the model encounters a sequence of T tasks. At each task t, the previous model θ_{t-1} is frozen as the teacher, and the current model θ_t is updated using the task data $\mathcal{D}_t$ and a small subset of prior data stored in the replay memory $\mathcal{R}$. The primary objective at task t is to effectively learn the new deepfake domain introduced in the current dataset, while maintaining high accuracy on all previously learned tasks by preserving knowledge from the previous model θ_{t-1}.

3.1 Framework Overview

Our proposed framework, Modality-Specific Frequency Distillation (MSFD), employs a unified optimization scheme that leverages frequency-domain distillation and cross-modality regularization. The overall pipeline, visualized in Fig. 3,

explicitly decomposes intermediate features [37, 58] into three complementary modalities—spatial, temporal, and spatiotemporal—via Fast Fourier Transforms (FFT), enabling distillation of distinct deepfake artifacts per modality. Our framework involves three key mechanisms: (1) the modality-wise frequency transform explicitly decomposes intermediate features into frequency spectra for each modality; (2) the Modality-Specific Adaptor (MSA) performs adaptive calibration and selective masking on the current model's spectra to compute the frequency-domain distillation loss $\mathcal{L}_{FDL}$, thereby more effectively preserving modality-specific cues; (3) the Cross-Modality Decorrelation Loss (CDL) $\mathcal{L}_{CDL}$ promotes distinct spatiotemporal representations while remaining orthogonal to those captured by single-modality branches. These components are integrated with standard classification loss $\mathcal{L}_{CE}$ and logit distillation loss $\mathcal{L}_{PD}$.

3.2 Modality-wise Frequency Transform

At task t, the current model θ_t is optimized with respect to the frozen previous model θ_{t-1} using replay samples from $\mathcal{R}$. For a set of intermediate layers L in the video backbone, we extract l^{th} layer's feature maps $F_{t-1}^{(l)}, F_t^{(l)} \in \mathbb{R}^{C_l \times T_l \times H_l \times W_l}$ from both models, where $l \in L$, C_l denotes the number of channels, T_l is the temporal dimension, and H_l, W_l are the spatial dimensions at l^{th} layer.

To explicitly decompose these representations into modality-specific components, we apply Fast Fourier Transform (FFT) along different dimensions for feature maps $F_t^{(l)}$ and $F_{t-1}^{(l)}$ to obtain three complementary magnitude spectra for each modality $d \in \{\mathsf{T}, \mathsf{S}, \mathsf{ST}\}$, corresponding to temporal, spatial, and spatiotemporal modalities. We then take the magnitude to construct modality-wise spectra $Z_t^{(d,l)}$ and $Z_{t-1}^{(d,l)}$.

- **Temporal (T):** 1D FFT along the temporal axis T_l to capture temporal frequency patterns.
- **Spatial (S):** 2D FFT over the spatial dimensions (H_l, W_l) to extract spatial frequency characteristics.
- **Spatiotemporal (ST):** 3D FFT across both temporal and spatial axes (T_l, H_l, W_l) to capture joint spatiotemporal frequency, targeting coupled space–time artifacts not recoverable from single-modality spectra alone.

3.3 Modality-Specific Adaptor

To more effectively handle modality-specific cues during distillation, we introduce the Modality-Specific Adaptor (MSA), which integrates modality-specific reweighting and modality-specific distillation masking.

Modality-Specific Reweighting. For each modality $d \in \{\mathsf{T}, \mathsf{S}, \mathsf{ST}\}$ and l^{th} layer, the modality-specific reweighting module applies a learnable, modality-wise reweighting parameter $\Phi_d^{(l)}$ to recalibrate the current model's spectrum $Z_t^{(d,l)}$ via a residual transformation:

$$\widetilde{Z}_t^{(d,l)} = \Phi_d^{(l)} \odot Z_t^{(d,l)} + Z_t^{(d,l)}, \tag{1}$$

where the reweighting parameter $\Phi_d^{(l)}$ is defined with a modality-dependent shape and is automatically broadcast to match the spectrum $Z_t^{(d,l)} \in \mathbb{R}^{B\times C\times T_l\times H_l\times W_l}$, enabling element-wise spectral modulation. Specifically, for the temporal modality (T), the reweighting parameter spans only the temporal axis, resulting in $\Phi_{\mathsf{T}}^{(l)} \in \mathbb{R}^{1\times C\times T_l\times 1\times 1}$. For the spatial modality (S), the parameter covers the spatial dimensions, giving $\Phi_{\mathsf{S}}^{(l)} \in \mathbb{R}^{1\times C\times 1\times H_l\times W_l}$. Finally, for the spatiotemporal modality (ST), all axes are reweighted jointly, yielding $\Phi_{\mathsf{ST}}^{(l)} \in \mathbb{R}^{1\times C\times T_l\times H_l\times W_l}$.

Modality-Specific Distillation Masking. The modality-specific distillation masking module learns a sparse, binary mask $M_d^{(l)}$ to select discriminative frequency bands based on data, rather than hand-crafted priors. The frequency spectrum's radial axis is discretized into K_d bins. A learnable logit vector $\mathbf{a}_{d,k}^{(l)} \in \mathbb{R}^2$ for each bin k is converted into a binary pass/block decision $m_{d,k}^{(l)} \in \{0,1\}$ using the Gumbel-Softmax operator $\mathcal{G}(\cdot;\tau)$:

$$m_{d,k}^{(l)} = \mathcal{G}(\mathbf{a}_{d,k}^{(l)};\tau)\cdot \mathbf{e}_{\text{pass}}, \tag{2}$$

where $\mathbf{e}_{\text{pass}} = [0,1]^\top$ selects the "pass" component, and the temperature τ is linearly annealed from 1.0 to 0.05 to encourage sharp selections.

The radial mask at frequency location $\mathbf{u}$ is then defined as:

$$M_d^{(l)}(\mathbf{u}) = m_{d,\kappa_d^{(l)}(\mathbf{u})}^{(l)} \in \{0,1\}, \tag{3}$$

which directly maps each frequency location to its corresponding bin's pass/block decision. The learned radial mask is broadcast to modality-specific dimensions, such as $(1,1,T_l,1,1)$ for temporal, $(1,1,1,H_l,W_l)$ for spatial, and $(1,1,T_l,H_l,W_l)$ for spatiotemporal, and then applied element-wise to each corresponding spectrum to obtain the final masked representations:

$$\widehat{Z}_t^{(d,l)} = M_d^{(l)} \odot \widetilde{Z}_t^{(d,l)}, \quad \widehat{Z}_{t-1}^{(d,l)} = M_d^{(l)} \odot Z_{t-1}^{(d,l)}, \tag{4}$$

where $\odot$ denotes element-wise multiplication after broadcasting. We empirically set the radial bins K_d for each modality as $K_{\mathsf{T}} = 4$, $K_{\mathsf{S}} = 16$, and $K_{\mathsf{ST}} = 32$.

3.4 Training Objective

Frequency-domain distillation loss. The frequency-domain distillation loss then enforces knowledge transfer in the most informative spectral regions:

$$\mathcal{L}_{\text{FDL}} = \sum_{l\in\mathcal{L}} \sum_{d\in\{\mathsf{T},\mathsf{S},\mathsf{ST}\}} \lambda_d \|\widehat{Z}_t^{(d,l)} - \widehat{Z}_{t-1}^{(d,l)}\|_2^2, \tag{5}$$

where λ_d is a modality-wise weighting coefficient.

Cross-modality decorrelation loss. To ensure the spatiotemporal representation captures unique joint space-time dynamics and remains orthogonal to the single-modality cues, we introduce the Cross-Modality Decorrelation Loss

(CDL). For each l-th layer, CDL measures the correlation between the spatiotemporal spectrum and its spatial or temporal counterparts:

$$\mathcal{L}_{\text{CDL}} = \sum_{l \in L} \begin{bmatrix} \mathcal{L}_{\text{corr}}(P^{(l)}_{\mathsf{ST}\to\mathsf{S}}(\widetilde{Z}^{(\mathsf{ST},l)}_t), P^{(l)}_{\mathsf{S}}(\widetilde{Z}^{(\mathsf{S},l)}_t)) \\ +\mathcal{L}_{\text{corr}}(P^{(l)}_{\mathsf{ST}\to\mathsf{T}}(\widetilde{Z}^{(\mathsf{ST},l)}_t), P^{(l)}_{\mathsf{T}}(\widetilde{Z}^{(\mathsf{T},l)}_t)) \end{bmatrix}, \tag{6}$$

where $P^{(l)}_d$ represents modality-specific projection heads that map the spectra into an alignment space. Before projection, spatial spectra are obtained by averaging over the temporal axis, temporal spectra by averaging across spatial dimensions, and spatiotemporal spectra are used directly without reduction. Each spectrum is then projected through its corresponding head ($P^{(l)}_{\mathsf{S}}$, $P^{(l)}_{\mathsf{T}}$, $P^{(l)}_{\mathsf{ST}\to\mathsf{S}}$, $P^{(l)}_{\mathsf{ST}\to\mathsf{T}}$) into an alignment space with dimensionality equal to the number of frequency bins for that modality (*e.g.*, K_{S}, K_{T}).

For projected tensors $A, B \in \mathbb{R}^{B\times C\times k}$, the mean per-channel Pearson correlation is defined as:

$$\text{corr}(A, B) = \frac{1}{BC} \sum_{b=1}^{B} \sum_{c=1}^{C} \left(\frac{1}{k} \sum_{i=1}^{k} A'_{b,c,i} B'_{b,c,i} \right), \tag{7}$$

where A' and B' are channel-wise z-score normalized along the frequency dimension. In practice, $\mathcal{L}_{\text{corr}}(A, B)$ is computed as the mean squared correlation, $\mathcal{L}_{\text{corr}}(A, B) = \text{corr}(A, B)^2$, equally penalizing positive and negative correlations.

Final objective loss. The final objective function integrates classification, logit-level distillation [19, 39], and our frequency-based regularization terms:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \lambda_{\text{PD}}\mathcal{L}_{\text{PD}} + \lambda_{\text{FDL}}\mathcal{L}_{\text{FDL}} + \lambda_{\text{CDL}}\mathcal{L}_{\text{CDL}}, \tag{8}$$

where $\mathcal{L}_{\text{CE}}$ is the standard binary cross-entropy loss computed on the combined data ($\mathcal{D}_t \cup \mathcal{R}$). The term $\mathcal{L}_{\text{PD}}$ is the memory-based logit-level distillation loss, which is applied exclusively to samples from the replay memory $\mathcal{R}$ to align the student's logits with the frozen previous model's predictions. Finally, λ_{PD}, λ_{FDL}, and λ_{CDL} are empirically set weighting coefficients.

4 Experiments

Datasets. We conduct experiments on ten publicly available video-based deepfake datasets spanning a broad range of forgery techniques: FF++ [42], DFD [8], CDF [27], DFDCp [7], FFIW [63], KoDF [22], FSh [26], DFo [17], DF40 [55], and AIGVDBench [31]. The first nine are face-centric video datasets capturing the temporal dynamics of facial forgeries; AIGVDBench additionally covers non-facial AI-generated content to evaluate generalization beyond face-centric scenarios. For DF40 [55], we adopt six manipulation methods: BlendFace [44], FaceDancer [41], and MobileSwap [54] for Face Swap (FS), and HyperReenact [2], MCNet [13], and SadTalker [59] for Face Reenactment (FR) deepfakes.

Continual Video Deepfake Detection Protocols. We evaluate our framework under three complementary continual learning protocols tailored to video-based deepfake detection.

- *Protocol 1: Dataset Incremental (DI).* This protocol emulates a practical scenario where detectors must continually adapt as new deepfake datasets emerge. The model is sequentially trained on FF++ (Task 1), followed by DFD, CDF, DFDCp, FFIW, and KoDF. For a balanced evaluation, we sample 200 real and 200 fake videos from each newly introduced dataset.
- *Protocol 2: Fake Incremental (FI).* This protocol evaluates robustness against structural shifts across generation paradigms in AIGVDBench [31], inspired by [4]: Text-to-Video (T2V) setup using EasyAnimate [52], Image-to-Video (I2V) via Pyramid-Flow [18], Video-to-Video (V2V) via LTX [9], and closed-source generation (SORA [35]), each relying on distinct generative priors that produce fundamentally different forgery patterns. Specifically, subsequent tasks introduce only new fake videos without any real data.
- *Protocol 3: Few-shot Incremental (FSI).* Following the continual deepfake detection setting [36], this protocol evaluates adaptation under a constrained data regime (only 25/25 real/fake videos per task, excluding the first task), which mirrors realistic scenarios where new forgeries first appear with limited samples. The model is sequentially updated across FF++, DFDCp, DFD, and CDF.

Evaluation Metrics. For each task t, we compute the video-level AUC, denoted AUC_t, and summarize the overall performance as: $AUC_{all} = \frac{1}{T}\sum_{t=1}^{T} AUC_t$. Forgetting is measured as the average performance drop relative to when each task was first learned: $F_{avg} = \frac{1}{T-1}\sum_{t=1}^{T-1}\left(a_{t,t} - a_{T,t}\right)$, where $a_{t,t}$ denotes the AUC on Task t immediately after learning it, and $a_{T,t}$ is the AUC on the same task after completing all T tasks. A higher AUC_{all} and lower F_{avg} indicate better retention and a more favorable stability–plasticity balance. For Protocol 3, following [36], we report clip-level accuracy with Average Accuracy (AA) and Average Forgetting (AF), where AA is the mean accuracy over all seen tasks, and AF measures the average performance drop from when each task was first learned.

Comparison Methods. We compare against a broad spectrum of continual learning baselines: naive fine-tuning, EWC [21], DCE [25], replay-based methods (ER [40], iCaRL-NME/FC [39], IDER [29]), video continual learning methods (Drift [14], HCE [28], vCLIMB [48], STSP [3], ESSENTIAL [23]), and continual deepfake detection methods (SARB-DF [38], DFIL [36], HDP [45], SURLID [4]). Unless otherwise specified, all methods adopt 3D ResNet-18 [47] and are initialized from Task 1 weights for a fair comparison. SARB-DF and ESSENTIAL are evaluated with their native backbones, as their architectures cannot be directly adapted.

Implementation Details. For preprocessing, as described in [62], we detect and align faces using RetinaFace. The model processes a 32-frame clip at a spatial resolution of 224×224. During training, up to 270 frames are extracted from each video to form multiple 32-frame clips via temporal sampling, while evaluation uses up to 110 frames to generate non-overlapping or minimally overlapping clips. Video-level scores are computed by averaging the clip-level logits. All models are trained for 20 epochs per task using the Adam optimizer with

Table 1: Comparison under Protocol 1: Dataset-Incremental (DI). We evaluate continual adaptation as new deepfake video datasets are sequentially introduced, and report video-level AUC (%) after the final task, overall AUC_{all}, and average forgetting F_{avg}. Unless otherwise noted, all methods use 3DResNet-18 as backbone; all replay-based methods set memory size to 200 clips. Best results are in **bold** and second best are underlined. ↑/↓ indicates higher/lower is better.

Method	Backbone	FF++	DFD	CDF	DFDCp	FFIW	KoDF	AUC_{all}↑	F_{avg}↓
▷*General continual learning methods*									
Finetuning	3DResNet-18	89.47	89.29	71.13	73.82	84.98	99.77	84.74	14.07
EWC [21]		81.76	86.16	67.67	69.07	83.23	99.89	81.30	18.69
ER [40]		92.36	94.21	72.14	74.13	91.68	99.75	87.38	10.77
iCaRL-NME [39]		86.67	92.22	74.76	80.59	87.60	96.86	86.45	8.14
iCaRL-FC [39]		96.82	94.65	80.50	79.20	88.77	99.07	89.84	6.81
DCE [25]		**99.63**	96.05	72.97	83.93	90.40	90.91	88.98	**0.07**
IDER [29]		95.32	**96.26**	**93.62**	80.09	83.80	97.48	91.10	2.97
▷*Video continual learning methods*									
vCLIMB [48]	3DResNet-18	80.07	90.46	66.61	77.71	85.51	96.36	82.79	11.27
Drift [14]		85.03	82.80	67.56	76.14	94.36	**99.99**	84.31	14.53
HCE [28]		81.54	91.56	62.18	78.96	92.08	99.28	84.27	13.06
STSP [3]		99.48	92.67	83.76	77.15	82.78	98.35	89.03	3.98
ESSENTIAL [23]	ViT-B/16 (CLIP)	68.10	95.50	74.40	**89.70**	41.30	99.30	78.05	20.04
▷*Continual deepfake detection methods*									
DFIL [36]	3DResNet-18	90.51	92.02	78.86	74.58	91.52	99.76	87.88	10.02
HDP [45]		92.22	86.20	68.58	70.21	83.07	99.95	83.37	16.14
SURLID [4]		94.98	90.25	83.20	81.30	76.81	99.12	87.61	8.19
SARB-DF [38]	Xception + Transformer	81.99	90.60	83.26	83.55	80.46	99.16	86.50	11.63
Ours	3DResNet-18	98.42	91.15	90.17	84.14	**94.88**	99.36	**93.02**	2.29

a learning rate of 1×10^{-4} and a batch size of 8. The replay memory stores 200 clips (each containing 32 frames) for replay-based continual learning methods. We adopt iCaRL [39] with a linear classifier as a baseline and employ its exemplar-based memory update strategy to focus on the effectiveness of our preservation mechanism. Additional implementation details are provided in the Supplementary.

4.1 Comparisons on Continual Deepfake Video Detection

Comparison on Protocol 1: Dataset Incremental (DI). As shown in Table 1, our method consistently outperforms existing continual learning strategies under Protocol 1, achieving the highest AUC_{all}. Conventional video continual learning methods (*e.g.*, vCLIMB [48], HCE [28]) primarily focus on preserving high-level semantic representations, which are insufficient for retaining low-level forensic cues, resulting in pronounced forgetting. Continual deepfake detection methods (*e.g.*, DFIL [36], SURLID [4]) rely on a single joint representation under image-level assumptions, failing to disentangle spatial and temporal forensic signals. MSFD overcomes these limitations by independently preserving modality-specific forensic cues in the frequency domain, achieving the highest overall AUC_{all}. While DCE [25] and IDER [29] show competitive forgetting, this comes at the cost of reduced plasticity. Furthermore, as shown in Fig. 4, our method maintains near-zero forgetting throughout the sequence while achieving a steady increase in AUC, demonstrating both stability and adaptability.

Table 2: Protocol 2: Fake-Incremental. We evaluate continual adaptation to structurally distinct generation paradigms and report AUC (%) after the final task, overall AUC_{all}, and average forgetting F_{avg}.

Method	T2V	I2V	V2V	SORA	$AUC_{all}\uparrow$	$F_{avg}\downarrow$
iCaRL-FC [39]	98.49	87.99	96.51	96.71	94.93	0.30
IDER [29]	**99.69**	84.87	95.92	96.51	94.25	-0.32
STSP [3]	99.57	83.28	93.65	89.53	91.51	**-0.37**
DFIL [36]	93.24	83.88	95.35	97.74	92.55	4.59
SURLID [4]	97.31	83.94	96.97	95.62	93.46	0.25
Ours	98.89	**91.47**	**97.49**	**98.84**	**96.67**	1.37

Table 3: Protocol 3: Few-shot Incremental. We evaluate continual adaptation under a restricted data regime and report clip-level accuracy (%) after the final task, Average Accuracy (AA), and Average Forgetting (AF), following [36].

Method	FF++	DFDCp	DFD	CDF	AA ↑	AF ↓
iCaRL-FC [39]	95.95	68.23	88.08	75.53	81.95	2.39
IDER [29]	95.50	**77.04**	**94.67**	66.67	83.47	**-2.10**
STSP [3]	87.81	76.78	52.86	**82.90**	75.09	1.59
DFIL [36]	94.50	63.14	81.93	79.08	79.66	5.47
SURLID [4]	95.98	71.91	93.19	71.24	83.08	0.58
Ours	**97.10**	72.04	90.32	75.13	**83.65**	-1.25

Comparison on Protocol 2: Fake Incremental (FI). MSFD achieves the highest overall AUC_{all} across generation paradigms (Table 2). We select the five most competitive methods from Protocol 1, including iCaRL-FC [39], IDER [29], STSP [3], DFIL [36], and SURLID [4], for evaluation under Protocols 2 and 3. The gain is particularly notable on I2V, where MSFD outperforms all baselines by a large margin, suggesting that modality-specific frequency distillation effectively preserves generation-paradigm-specific forensic cues that single-representation approaches fail to retain across tasks. While STSP and IDER report lower forgetting, this comes at the cost of substantially lower overall accuracy, indicating that they sacrifice adaptability to new paradigms to retain prior knowledge. MSFD achieves a more favorable balance, attaining the highest AUC_{all} with moderate forgetting.

Comparison on Protocol 3: Few-shot Incremental (FSI). Under the few-shot setting (Table 3), MSFD achieves the highest AA while also exhibiting negative forgetting (AF $= -1.25$), indicating positive backward transfer to previously learned tasks. Although IDER reports a lower AF, its overall AA is lower than that of MSFD, suggesting that MSFD provides a stronger few-shot adaptability. Notably, STSP, which showed competitive performance in Protocol 1, suffers significant degradation under both Fake-Incremental and Few-shot Incremental settings, revealing that conventional video continual learning approaches lack the robustness to handle the extreme distribution shifts and data scarcity inherent in evolving deepfake scenarios.

4.2 Ablation Study and Analysis

We conduct ablation studies to validate the contribution of each component in our framework. Further analyses are provided in the supplementary material.

Analysis on Memory Size Sensitivity. We analyze the sensitivity of our framework to the size of the replay memory buffer. As shown in Fig. 5, our method (red) consistently outperforms all baselines across memory sizes from 50 to 400 in both AUC_{all} and F_{avg} under Protocol 1. Most notably, with only 50 memory samples, our framework already matches or surpasses DFIL and

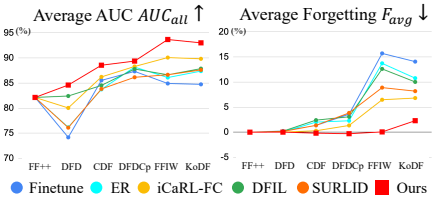


Fig. 4: Performance trends across sequential tasks. We visualize the average performance AUC_{all} and forgetting F_{avg} across sequential tasks under Protocol 1.

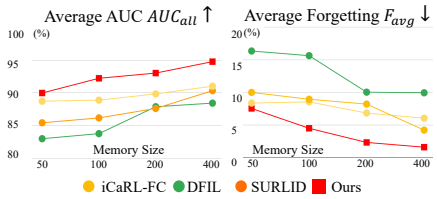


Fig. 5: Sensitivity for memory buffer size. We report average performance AUC_{all} and forgetting F_{avg} across a memory size from 50 to 400 in Protocol 1.

SURLID operating with 400 samples. This efficiency stems from our modality-specific distillation, which focuses on the most discriminative spectral components as replay exemplars, preserving critical knowledge without requiring large memory budgets. These results confirm that MSFD remains effective even in resource-constrained continual learning scenarios.

Forgetting Analysis Across Modalities. As shown in Fig. 6, our method consistently occupies the upper-right across most benchmarks, demonstrating superior preservation of spatial and temporal forgery cues under continual updates. To analyze modality-specific forgetting behavior beyond aggregate metrics, Fig. 6 visualizes the spatial and temporal retention of each method across benchmarks in Protocol 1, defined as the ratio of the final model's AUC under spatial and temporal perturbation (grid shuffling/frame shuffling) to its AUC when first learning each dataset. While competing methods scatter toward lower retention on one or both modalities, our results confirm that explicitly preserving spatial and temporal representations is critical for robust continual deepfake detection. Details will be provided in the supplementary material.

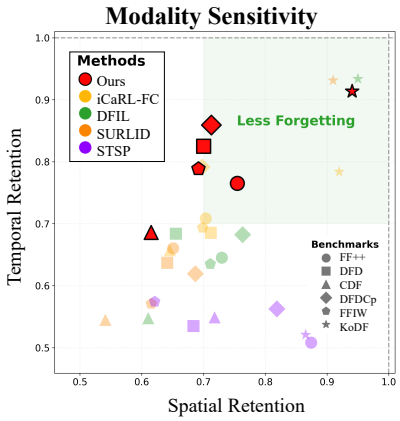


Fig. 6: Forgetting analysis across spatial and temporal modalities. Each point represents a method–dataset pair, plotting spatial and temporal retention, defined as the ratio of the final model's AUC under each perturbation to its AUC when first learned.

Contribution of Components. Each proposed component addresses a distinct bottleneck, and their combination achieves the best balance between adaptation and retention (Table 4-(a)). Multi-layer feature distillation provides only marginal improvement, confirming that

Table 4: Ablation studies of MSFD. (a) Contribution of major components (MSFD, MSA, and CDL) over the baseline. Multi-layer Distillation extends naive feature distillation to multiple intermediate layers of the baseline. (b) Effect of modality decomposition: S (spatial), T (temporal), and ST (spatiotemporal). (c) Effect of CDL on Protocol 1 and robustness on FF++/DF40 (FS: face swap, FR: face reenactment). (d) Breakdown of MSA (MSR and MSDM). (e) Mask types: fixed high/low-pass vs. learnable Gumbel–Softmax in MSDM. (f) Backbone generalization under continual learning settings. Unless otherwise specified, all results are reported under Protocol 1.

(a) Components.

Model	MSFD	MSA	CDL	$AUC_{all}\uparrow$	$F_{avg}\downarrow$
Baseline				89.84	6.81
Multi-layer Distill.				90.64	6.34
MSFD (Naive)	✓			92.65	4.09
MSFD + MSA	✓	✓		92.14	3.20
MSFD + CDL	✓		✓	92.77	2.65
Ours (Full)	✓	✓	✓	**93.02**	**2.29**

(b) Modalities.

Modality	S	T	ST	$AUC_{all}\uparrow$	$F_{avg}\downarrow$
w/o Decompose	–	–	–	90.64	6.34
Spatial Only	✓	–	–	90.69	5.29
Temporal Only	–	✓	–	91.45	4.41
Spatiotemp. Only	–	–	✓	92.06	3.27
Full (Ours)	✓	✓	✓	**93.02**	**2.29**

(c) Effect of CDL.

Method	Protocol 1		FF++		DF40	
	$AUC_{all}\uparrow$	$F_{avg}\downarrow$	FS	FR	FS	FR
w/o CDL	92.14	3.20	99.5	97.5	92.8	75.6
Ours	**93.02**	**2.29**	**99.6**	**97.8**	**93.3**	**85.6**

(d) MSA components.

Configuration	MSR	MSDM	$AUC_{all}\uparrow$	$F_{avg}\downarrow$
w/o MSA	–	–	92.77	2.65
+ MSR Only	✓	–	92.97	2.43
+ MSDM Only	–	✓	91.91	3.08
MSA (Full)	✓	✓	**93.02**	**2.29**

(e) Mask type in MSDM.

Mask Type	Trainable	$AUC_{all}\uparrow$	$F_{avg}\downarrow$
None	–	92.97	2.43
High-pass	Fixed	92.94	2.53
Low-pass	Fixed	92.38	3.14
Ours	**Learnable**	**93.02**	**2.29**

(f) Backbone.

Backbone	Method	$AUC_{all}\uparrow$	$F_{avg}\downarrow$
R3D-18	iCaRL-FC	89.84	6.81
	Ours	**93.02**	**2.29**
R(2+1)D-18	iCaRL-FC	86.10	8.75
	Ours	**89.24**	**2.17**
FTCN	iCaRL-FC	74.05	16.39
	Ours	**75.34**	**8.03**

preserving entangled spatial-temporal features in the raw space is insufficient. MSFD (Naive) yields a substantial gain by decomposing features into modality-specific frequency spectra, validating frequency-domain decomposition as the key enabler. MSA improves stability by suppressing noisy frequency bands, while CDL reduces cross-modality redundancy. The full model achieves the strongest overall performance, confirming their complementary roles.

Analysis of Modality Decomposition. All three frequency modalities are necessary, and their combination yields the most robust performance (Table 4-(b)). Spatial-only (S) provides the smallest gain, as spatial cues are well preserved by standard distillation. Temporal-only (T) shows a larger improvement, consistent with prior findings that temporal cues are more generalizable forensic signals [11, 62]. Spatiotemporal-only (ST) achieves the strongest result by capturing joint space-time dynamics inaccessible to 1D or 2D transforms. The full model further improves, confirming complementarity across the three spectra.

Effect of Cross-modality Decorrelation Loss (CDL). CDL improves both accuracy and stability by preventing the ST branch from redundantly encoding cues already preserved by the spatial and temporal branches (Table 4-(c)). Without CDL, such redundancy wastes capacity and compounds forgetting when overlapping representations shift during later tasks. Furthermore, we find the largest improvement on Face Reenactment (FR) in DF40 [55], an advanced variant of FF++. Since FR modifies expression dynamics, its artifacts appear as temporally inconsistent spatial deformations, making spatiotemporal cues more critical than in Face Swap (FS).

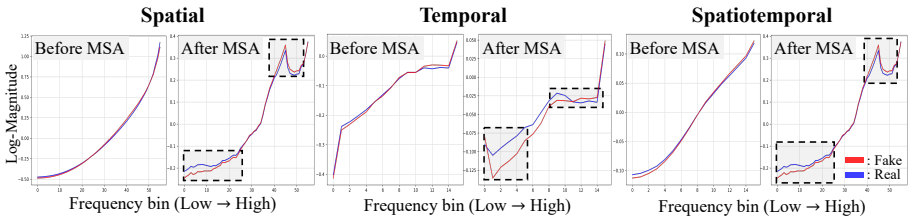


Fig. 7: Effect of the Modality-Specific Adaptor (MSA) on frequency distributions. We visualize frequency spectra of real (blue) and fake (red) samples at Layer 1 across three modalities (spatial, temporal, and spatiotemporal). For each modality, the left plot shows spectra before MSA and the right plot after MSA.

Analysis of the Modality-Specific Adaptor (MSA). The combination of Modality-Specific Reweighting (MSR) and Modality-Specific Distillation Masking (MSDM) achieves the highest performance and lowest forgetting, demonstrating strong synergy between the two components as shown in Table 4-(d). MSR adaptively calibrates channel-wise spectral responses to reduce forgetting through modality-aware spectral reweighting, while MSDM selectively highlights discriminative frequency components via a Gumbel–Softmax mechanism. Their complementary roles, MSR for calibration and MSDM for frequency-wise selection, enable comprehensive spectral adaptation across all modalities.

To further validate the learnable masking mechanism, we compare the modality-specific distillation masking against a fixed frequency filter. As shown in Table 4-(e), replacing the learnable mask with a static high-pass or low-pass filter degrades performance, indicating that fixed frequency selection cannot adapt to task-specific spectral characteristics. In contrast, our learnable mask dynamically allocates attention across frequency regions, achieving the best performance and confirming that adaptive frequency selection is crucial for capturing discriminative spectral cues and maintaining stability in continual deepfake video detection.

To analyze the impact of MSA, we visualize the frequency distributions of real and fake samples at Layer 1, representing low-level features, across spatial, temporal, and spatiotemporal modalities using our model. As shown in Fig. 7, the top row presents the distributions before applying MSA, while the bottom row illustrates those after applying MSA. Without MSA, the spectral responses of real and fake samples largely overlap. In contrast, after MSA is introduced, the separation between real and fake distributions becomes more pronounced across all modalities, demonstrating that MSA enables the model to learn forgery-relevant cues even from the low-level representation stage.

Backbone Generalization Analysis. Our modality-aware frequency distillation generalizes across architecturally diverse backbones, consistently outperforming the baseline on all tested networks (Table 4-(f)). To further validate this generality, we evaluate our framework across different backbone architectures under identical continual learning settings, including the standard 3D ResNet-18 and R(2+1)D-18 video backbones [47], as well as the FTCN [62], a deepfake-

specific backbone. Note that FTCN is originally equipped with a transformer, from which we remove the transformer modules due to memory constraints.

Failure Case. Despite consistent overall gains, our proposed method exhibits two failure modes. *(i) Race-specific distribution shifts*: after learning KoDF (predominantly Asian faces), MSFD's forgetting rises from near-zero to ∼2 (Fig. 4). This suggests a demographic shift in KoDF: differences in facial appearance change the spatial/frequency patterns learned earlier, causing more forgetting. *(ii) Residual asymmetric retention*: the spatial and temporal branches still degrade at uneven rates (Fig. 6).

5 Conclusion and Future Work

In this work, we introduced a novel framework for continual deepfake video detection that mitigates catastrophic forgetting through modality-specific frequency-domain knowledge preservation. Specifically, we proposed a Modality-Specific Frequency Distillation (MSFD) strategy that decomposes intermediate representations into spatial, temporal, and spatiotemporal spectra, enabling modality-specific knowledge transfer. We further introduced the Modality-Specific Adaptor (MSA) for discriminative spectral calibration and masking, along with the Cross-Modality Decorrelation Loss (CDL) to encourage orthogonal frequency representations across modalities. Through comprehensive experiments across multiple continual learning protocols, unseen forgery domains, and perturbation conditions, our method demonstrated superior generalization, robustness, and stability compared to existing continual deepfake detection frameworks.

Future Work. Building upon the strong performance of memory-based replay methods in continual deepfake video detection, future work will explore deepfake video-specific memory update strategies that explicitly consider both spatial and temporal information to guide exemplar selection and maintenance. Additionally, developing adaptive mechanisms that dynamically balance the importance of each modality per task could further improve robustness across diverse generation paradigms and reduce sensitivity to hyperparameter choices.

Acknowledgements

This work was supported in part by the Institute of Information & Communication Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2025-02263841, Development of a Real-time Multimodal Framework for Comprehensive Deepfake Detection Incorporating Common Sense Error Analysis; RS-2021-II211341, Artificial Intelligence Graduate School Program (Chung-Ang University)), and by [Ministry of Science and ICT] (MSIT) through the [National Research Foundation of Korea] (NRF) funded by the Ministry of Science and ICT (RS-2025-16066667).

References

1. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., Jampani, V., Rombach, R.: Stable video diffusion: Scaling latent video diffusion models to large datasets. CoRR **abs/2311.15127** (2023). https://doi.org/10.48550/ARXIV.2311.15127, https://doi.org/10.48550/arXiv.2311.15127
2. Bounareli, S., Tzelepis, C., Argyriou, V., Patras, I., Tzimiropoulos, G.: Hyperreenact: one-shot reenactment via jointly learning to refine and retarget faces. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 7149–7159 (2023)
3. Cheng, H., Yang, S., Wang, C., Zhou, J.T., Kot, A.C., Wen, B.: Stsp: Spatial-temporal subspace projection for video class-incremental learning. In: European Conference on Computer Vision. pp. 374–391. Springer (2024)
4. Cheng, J., Yan, Z., Zhang, Y., Hao, L., Ai, J., Zou, Q., Li, C., Wang, Z.: Stacking brick by brick: Aligned feature isolation for incremental face forgery detection. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 13927–13936 (2025)
5. Choi, J., Kim, T., Jeong, Y., Baek, S., Choi, J.: Exploiting style latent flows for generalizing deepfake video detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1133–1143 (2024)
6. Chollet, F.: Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1251–1258 (2017)
7. Dolhansky, B., Howes, R., Pflaum, B., Baram, N., Canton-Ferrer, C.: The deepfake detection challenge (DFDC) preview dataset. CoRR **abs/1910.08854** (2019), http://arxiv.org/abs/1910.08854
8. Dufour, N., Gully, A.: Contributing Data to Deepfake Detection Research — ai.googleblog.com. https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html (2019), [Accessed 30-07-2023]
9. HaCohen, Y., Chiprut, N., Brazowski, B., Shalem, D., Moshe, D., Richardson, E., Levin, E., Shiran, G., Zabari, N., Gordon, O., Panet, P., Weissbuch, S., Kulikov, V., Bitterman, Y., Melumian, Z., Bibi, O.: Ltx-video: Realtime video latent diffusion. CoRR **abs/2501.00103** (2025). https://doi.org/10.48550/ARXIV.2501.00103, https://doi.org/10.48550/arXiv.2501.00103
10. Haliassos, A., Mira, R., Petridis, S., Pantic, M.: Leveraging real talking faces via self-supervision for robust forgery detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14950–14962 (June 2022)
11. Haliassos, A., Vougioukas, K., Petridis, S., Pantic, M.: 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. pp. 5039–5049 (2021)
12. Han, Y.H., Huang, T.M., Hua, K.L., Chen, J.C.: Towards more general video-based deepfake detection through facial component guided adaptation for foundation model. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22995–23005 (2025)
13. Hong, F.T., Xu, D.: Implicit identity representation conditioned memory compensation network for talking head video generation. In: ICCV (2023)

14. Hu, Y., Hou, J., Liu, X., Sun, X., Guo, W.: Video domain incremental learning for human action recognition in home environments. In: International Conference on Image and Graphics. pp. 316–327. Springer (2025)
15. Jeong, Y., Kim, D., Min, S., Joe, S., Gwon, Y., Choi, J.: Bihpf: Bilateral high-pass filters for robust deepfake detection. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 48–57 (2022)
16. Jeong, Y., Kim, D., Ro, Y., Choi, J.: Frepgan: robust deepfake detection using frequency-level perturbations. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 1060–1068 (2022)
17. Jiang, L., Li, R., Wu, W., Qian, C., Loy, C.C.: DeeperForensics-1.0: A large-scale dataset for real-world face forgery detection. In: CVPR (2020)
18. Jin, Y., Sun, Z., Li, N., Xu, K., Jiang, H., Zhuang, N., Huang, Q., Song, Y., Mu, Y., Lin, Z.: Pyramidal flow matching for efficient video generative modeling. In: International Conference on Learning Representations. vol. 2025, pp. 23378–23402 (2025)
19. Kim, M., Tariq, S., Woo, S.S.: Cored: Generalizing fake media detection with continual representation using distillation. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 337–346 (2021)
20. Kim, T., Choi, J., Jeong, Y., Noh, H., Yoo, J., Baek, S., Choi, J.: Beyond spatial frequency: Pixel-wise temporal frequency-based deepfake video detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 11198–11207 (October 2025)
21. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al.: Overcoming catastrophic forgetting in neural networks. Proceedings of the national academy of sciences **114**(13), 3521–3526 (2017)
22. Kwon, P., You, J., Nam, G., Park, S., Chae, G.: Kodf: A large-scale korean deepfake detection dataset. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 10744–10753 (October 2021)
23. Lee, J., Bae, K., Min, K., Park, G.M., Choi, J.: Essential: Episodic and semantic memory integration for video class-incremental learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17546–17556 (2025)
24. Li, C., Huang, Z., Paudel, D.P., Wang, Y., Shahbazi, M., Hong, X., Van Gool, L.: A continual deepfake detection benchmark: Dataset, methods, and essentials. In: Winter Conference on Applications of Computer Vision (WACV) (2023)
25. Li, L., Zhou, D.W., Ye, H.J., Zhan, D.C.: Addressing imbalanced domain-incremental learning through dual-balance collaborative experts. ICML (2025)
26. Li, L., Bao, J., Yang, H., Chen, D., Wen, F.: Faceshifter: Towards high fidelity and occlusion aware face swapping. arXiv preprint arXiv:1912.13457 (2019)
27. Li, Y., Yang, X., Sun, P., Qi, H., Lyu, S.: Celeb-df: A large-scale challenging dataset for deepfake forensics. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3207–3216 (2020)
28. Liang, S., Zhu, K., Zhai, W., Liu, Z., Cao, Y.: Hypercorrelation evolution for video class-incremental learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 3315–3323 (2024)
29. Liu, Z., Li, Y., Gao, H., Li, Y., Kong, L., Sun, L., Huang, W.: IDER: IDempotent experience replay for reliable continual learning. In: The Fourteenth International Conference on Learning Representations (2026), https://openreview.net/forum?id=Vr5f3kRvLD

30. Luo, Y., Zhang, Y., Yan, J., Liu, W.: Generalizing face forgery detection with high-frequency features. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16317–16326 (June 2021)
31. Ma, L., Xue, Z., Wang, Y., Yan, Z., Xu, J., Jiang, X., Yu, H., Liao, Y., Bi, Z.: Your one-stop solution for ai-generated video detection. arXiv preprint arXiv:2601.11035 (2026)
32. Ma, X., Tian, J., Cai, Y., Chai, Y., Li, Z., Dai, J., Zang, L., Han, J.: Hidd: Human-perception-centric incremental deepfake detection. In: 2024 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6. IEEE (2024)
33. Masi, I., Killekar, A., Mascarenhas, R.M., Gurudatt, S.P., AbdAlmageed, W.: Two-branch recurrent network for isolating deepfakes in videos. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16. pp. 667–684. Springer (2020)
34. Nguyen, D., Astrid, M., Kacem, A., Ghorbel, E., Aouada, D.: Vulnerability-aware spatio-temporal learning for generalizable deepfake video detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10786–10796 (2025)
35. OpenAI: Sora. `https://openai.com/index/sora/` (2024), [Accessed 28-06-2026]
36. Pan, K., Yin, Y., Wei, Y., Lin, F., Ba, Z., Liu, Z., Wang, Z., Cavallaro, L., Ren, K.: Dfil: Deepfake incremental learning by exploiting domain-invariant forgery clues. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 8035–8046 (2023)
37. Pham, C., Nguyen, V.A., Le, T., Phung, D., Carneiro, G., Do, T.T.: Frequency attention for knowledge distillation. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 2277–2286 (2024)
38. Prathibha, P., Tamizharasan, P., Panthakkan, A., Mansoor, W., Al Ahmad, H.: Sarb-df: A continual learning aided framework for deepfake video detection using self-attention residual block. IEEE Access **12**, 189088–189101 (2024)
39. Rebuffi, S.A., Kolesnikov, A., Sperl, G., Lampert, C.H.: icarl: Incremental classifier and representation learning. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 2001–2010 (2017)
40. Rolnick, D., Ahuja, A., Schwarz, J., Lillicrap, T., Wayne, G.: Experience replay for continual learning. Advances in neural information processing systems **32** (2019)
41. Rosberg, F., Aksoy, E.E., Alonso-Fernandez, F., Englund, C.: Facedancer: Pose- and occlusion-aware high fidelity face swapping. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 3454–3463 (January 2023)
42. Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M.: Faceforensics++: Learning to detect manipulated facial images. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1–11 (2019)
43. Shiohara, K., Yamasaki, T.: Detecting deepfakes with self-blended images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18720–18729 (2022)
44. Shiohara, K., Yang, X., Taketomi, T.: Blendface: Re-designing identity encoders for face-swapping. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2023)
45. Sun, K., Chen, S., Yao, T., Sun, X., Ding, S., Ji, R.: Continual face forgery detection via historical distribution preserving. International Journal of Computer Vision **133**(3), 1067–1084 (2025)

46. Tian, J., Yu, C., Wang, X., Chen, P., Xiao, Z., Han, J., Chai, Y.: Dynamic mixed-prototype model for incremental deepfake detection. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 8129–8138 (2024)
47. Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M.: A closer look at spatiotemporal convolutions for action recognition. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 6450–6459 (2018)
48. Villa, A., Alhamoud, K., Escorcia, V., Caba, F., Alcázar, J.L., Ghanem, B.: vclimb: A novel video class incremental learning benchmark. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19035–19044 (2022)
49. Wang, S.Y., Wang, O., Zhang, R., Owens, A., Efros, A.A.: Cnn-generated images are surprisingly easy to spot... for now. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8695–8704 (2020)
50. Wang, Y., Chen, X., Ma, X., Zhou, S., Huang, Z., Wang, Y., Yang, C., He, Y., Yu, J., Yang, P., et al.: Lavie: High-quality video generation with cascaded latent diffusion models. International Journal of Computer Vision **133**(5), 3059–3078 (2025)
51. Wang, Z., Bao, J., Zhou, W., Wang, W., Li, H.: Altfreezing for more general video face forgery detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4129–4138 (2023)
52. Xu, J., Zou, X., Huang, K., Chen, Y., Liu, B., Cheng, M., Shi, X., Huang, J.: Easyanimate: A high-performance long video generation method based on transformer architecture. arXiv preprint arXiv:2405.18991 (2024)
53. Xu, Y., Liang, J., Jia, G., Yang, Z., Zhang, Y., He, R.: Tall: Thumbnail layout for deepfake video detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22658–22668 (2023)
54. Xu, Z., Hong, Z., Ding, C., Zhu, Z., Han, J., Liu, J., Ding, E.: Mobilefaceswap: A lightweight framework for video face swapping. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 2973–2981 (2022)
55. Yan, Z., Yao, T., Chen, S., Zhao, Y., Fu, X., Zhu, J., Luo, D., Wang, C., Ding, S., Wu, Y., Yuan, L.: DF40: toward next-generation deepfake detection. In: Globersons, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J.M., Zhang, C. (eds.) Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024 (2024), http://papers.nips.cc/paper_files/paper/2024/hash/34239f60eca7ce9bee5280aaf81362d8-Abstract-Datasets_and_Benchmarks_Track.html
56. Yan, Z., Zhang, Y., Fan, Y., Wu, B.: Ucf: Uncovering common features for generalizable deepfake detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 22412–22423 (2023)
57. Yan, Z., Zhao, Y., Chen, S., Guo, M., Fu, X., Yao, T., Ding, S., Wu, Y., Yuan, L.: Generalizing deepfake video detection with plug-and-play: Video-level blending and spatiotemporal adapter tuning. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 12615–12625 (2025)
58. Yu, F., Pan, H., Zhang, K., Guan, J., Jiang, H.: Uhkd: A unified framework for heterogeneous knowledge distillation via frequency-domain representations. arXiv preprint arXiv:2510.24116 (2025)
59. Zhang, W., Cun, X., Wang, X., Zhang, Y., Shen, X., Guo, Y., Shan, Y., Wang, F.: Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8652–8661 (2023)

60. Zhang, X., Zhu, P., Zhang, C., Yan, Z., Cheng, J., Lao, M., Cai, S., Guo, Y.: Generalization-preserved learning: Closing the backdoor to catastrophic forgetting in continual deepfake detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3798–3808 (October 2025)
61. Zhang, Y., Liu, Y., Xia, B., Peng, B., Yan, Z., Lo, E., Jia, J.: Magicmirror: Id-preserved video generation in video diffusion transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14464–14474 (2025)
62. Zheng, Y., Bao, J., Chen, D., Zeng, M., Wen, F.: Exploring temporal coherence for more general video face forgery detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 15044–15054 (October 2021)
63. Zhou, T., Wang, W., Liang, Z., Shen, J.: Face forensics in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5778–5788 (June 2021)