Deepfake Detection that Generalizes Across Benchmarks

Andrii Yermakov[1] Jan Cech[1] Jiri Matas[1] Mario Fritz[2]
[1]Czech Technical University in Prague [2]CISPA Helmholtz Center for Information Security
{yermaand,cechj,matas}@fel.cvut.cz fritz@cispa.de

Abstract

The generalization of deepfake detectors to unseen manipulation techniques remains a challenge for practical deployment. Although many approaches adapt foundation models by introducing significant architectural complexity, this work demonstrates that robust generalization is achievable through a parameter-efficient adaptation of one of the foundational pre-trained vision encoders. The proposed method, GenD, fine-tunes only the Layer Normalization parameters (0.03% of the total) and enhances generalization by enforcing a hyperspherical feature manifold using L2 normalization and metric learning on it.

We conducted an extensive evaluation on 14 benchmark datasets spanning from 2019 to 2025. The proposed method achieves state-of-the-art performance, outperforming more complex, recent approaches in average cross-dataset AUROC. Our analysis yields two primary findings for the field: 1) training on paired real-fake data from the same source video is essential for mitigating shortcut learning and improving generalization, and 2) detection difficulty on academic datasets has not strictly increased over time, with models trained on older, diverse datasets showing strong generalization capabilities.

This work delivers a computationally efficient and reproducible method, proving that state-of-the-art generalization is attainable by making targeted, minimal changes to a pre-trained foundational image encoder model. The code is at: `https://github.com/yermandy/GenD`.

1. Introduction

The proliferation of realistic facial deepfakes raises significant concerns regarding misinformation and malicious use, with AI-manipulated videos – those altered by techniques like face swapping or face reenactment – making detection challenging. Unlike fully synthetic content, such forgeries preserve the original context and leave subtle artifacts that are difficult for humans and machines to detect [10, 24].

A primary issue affecting current detection methods is their limited ability to generalize. A model that has been trained to identify images altered by a particular deepfake generation algorithm often struggles when faced with examples produced by a new generation algorithm.

The generalization gap is the primary issue that we address in this work. Assuming the hypothesis that *adapted* large-scale, pre-trained foundational vision encoder can serve as a general foundation for deepfake detection [40], we build the proposed method in three variants, using Contrastive Language-Image Pre-training (CLIP) [43], Perception Encoder (PE) [8] and DINO [46] models as feature extractors, which are known for their generalizable visual representations.

The proposed method consists of a vision encoder, whose outputs are L2-normalized. We then fine-tune only the parameters of the Layer Normalization blocks [42] while keeping the rest frozen. Additionally, we propose using metric learning in this L2 space to enhance generalization.

We benchmarked the generalization capabilities of the proposed model on 14 deepfake video datasets released between 2019 and 2025, listed in Tab. 1. To our knowledge, this represents the broadest evaluation in the deepfake literature. We show that the proposed model outperforms the most recent state-of-the-art methods on the majority of all available benchmarks.

In summary, our **key contributions** are as follows:

- A novel deepfake detection method called GenD. The method achieves the best average cross-dataset AUROC compared to recently released models.
- The most comprehensive evaluation in the deepfake literature covering datasets released throughout six years of research.
- A demonstration that to achieve the best generalization and prevent shortcut learning, it is essential to construct the training set consisting of real-fake pairs, where the fake video is generated from the real counterpart of the pair.

2. Related Work

Deepfake techniques have rapidly evolved. Key methods include: Face-swap [4], Face-reenactment [9], LipSync [41], and full face video synthesis [14, 48]. The first three approaches manipulate only small, localized areas of the video, typically the facial region, while leaving the rest of the footage untouched, making detection challenging. The techniques are continually improving, leveraging advanced models to produce increasingly seamless and realistic results without obvious, visually perceptible artifacts.

As deepfake production techniques advance, detection methods are also evolving to keep pace. The literature encompasses a taxonomy of detection strategies, which can be divided into two main branches: content-based and signal-based approaches. Content-based methods focus on visible or interpretable inconsistencies, utilizing inductive biases such as apparent blending boundaries [45], unsynchronized audio-visual movements [20], or motion artifacts [1]. Signal-based methods, on the other hand, detect subtle and often invisible traces left in the visual signal, fingerprints that reveal synthetic origins even when no overt artifacts are present. Signal-based techniques have become increasingly accurate, especially as newer deepfakes leave fewer perceptible cues. However, a persistent challenge remains: generalizing to previously unseen manipulation methods.

Recent efforts in detecting general AI-generated content [29, 36, 39, 40] and facial deepfakes specifically [13, 18, 22, 37, 58] have begun to adopt CLIP as a backbone for generalizable deepfake detection. These approaches can be broadly categorized by the type of model adaptation. Some methods, like **Forensics Adapter** [13] (ForAda), introduce *architectural* changes by adding a separate parallel network explicitly trained to identify artifacts left by blending. Other methods operate on the *parameter space*; for instance, **Effort** [58] uses Singular Value Decomposition (SVD) to decompose the model's weights into orthogonal subspaces, freezing the principal components to preserve pre-trained knowledge while fine-tuning the residual components on forgery patterns.

In contrast to these approaches, our work proposes to modify a subset of parameters. LN-Tuning [42] adjusts the affine parameters of Layer Normalization blocks. The efficiency of LN-tuning in the general context was analyzed in [19, 51]. We demonstrate that it achieves competitive or state-of-the-art performance in the deepfake detection task.

As a complimentary approach to the passive detection methods of AI manipulations, there exist active defense approaches that protect real images by incorporating invisible watermarks [2, 3, 6, 61, 62, 65].

3. Method

The proposed method comes in three variants depending on the pre-trained image encoder to which we add two components: L2 normalization of the classification token and a linear classifier. We optimize only 0.03% of network parameters using a weighted combination of cross-entropy, uniformity, and alignment losses in an end-to-end fashion. We call this approach GenD, and its complete overview is shown in Fig. 1.

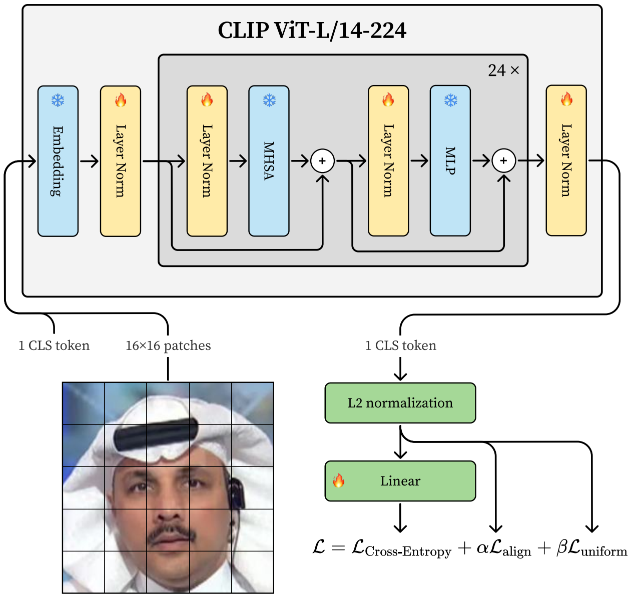

Figure 1. The architecture of GenD (CLIP). The gray rectangle represents the original CLIP ViT image encoder, and green represents the added components. Only rectangles with a fire icon represent layers with trainable parameters.

3.1. Model

Feature extractor. We use three different pre-trained image encoders: 1. CLIP ViT-L/14, trained by OpenAI [43]; 2. PE_{core}L, trained by Meta AI [8]; 3. DINOv3 ViT-L/16 [46], trained by Meta AI. After the encoder processes the image, we take only the 1024-dimensional classification token and discard the rest.

Optimized parameters. The L2-normalized classification token is used as the input to the linear binary classifier, which gives logits for fake and real classes. In addition to the weights $W \in \mathbb{R}^{1024\times 2}$ and biases $b \in \mathbb{R}^2$ of the classification layer, we optimize the parameters of all Layer Normalization blocks [19, 42, 51]. This roughly constitutes 0.03% of the total model parameters; the remaining parameters of the encoder are frozen.

Loss function. The network is trained using a weighted combination of cross-entropy, uniformity, and alignment losses [52]

$$\mathcal{L} = \mathcal{L}_{\text{Cross-Entropy}} + \alpha\mathcal{L}_{\text{align}} + \beta\mathcal{L}_{\text{uniform}}. \tag{1}$$

The cross-entropy loss is applied to fake and real logits, while uniformity and alignment losses are directly applied to the L2-normalized classification token, see Fig. 1.

Let $\mathbf{z}_i$ be the L2-normalized feature of the i-th sample. The alignment loss tightens feature clusters coming from the same class

$$\mathcal{L}_{\text{align}} = \mathop{\mathbb{E}}_{x,y\sim\mathcal{P}^+} \left[\|\mathbf{z}_x - \mathbf{z}_y\|_2^2\right], \tag{2}$$

where $\mathcal{P}^+$ is the distribution of positive pairs. Uniformity encourages the features to be uniformly distributed on the unit hypersphere

$$\mathcal{L}_{\text{uniform}} = \log \mathop{\mathbb{E}}_{x,y\sim\mathcal{P}} \left[e^{-2\|\mathbf{z}_x - \mathbf{z}_y\|_2^2}\right], \tag{3}$$

where x and y are sampled independently from the data distribution $\mathcal{P}$.

Training algorithm. To optimize the model parameters, we use Adam optimizer [30] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ without weight decay. The weight precision is set to bfloat16. The learning rate is scheduled using a cosine cyclic rule [47]. Each cycle starts with a linear warm-up for one epoch, increasing from 10^{-5} to 3×10^{-4}, and then decays using cosine scheduling over nine epochs to 10^{-5}. For most runs, we observed that the training did not improve after two cycles (or 20 epochs). The batch size is set to 128 samples. For CLIP ViT-L/14, the loss coefficients of Eq. (1) are set to $\alpha = 0.1$ and $\beta = 0.5$, found on the validation set.

Video classification. GenD has a single frame as an input. To get video-level probabilities, we uniformly sample 32 frames from a video, calculate softmax probabilities for every frame independently, and average them.

Speed and memory. Every encoder of GenD is of the size of a ViT-L model, consisting of roughly 304M parameters. Our non-optimized for speed and memory implementation in PyTorch can process 120 frames per second on A100 with a batch size of 1, requiring approximately 2GB of VRAM.

3.2. Data

Data preprocessing. We follow a standard practice of dataset preprocessing proposed in DeepfakeBench [56], which consists of the following steps:

1. Sample 32 frames evenly from each video.
2. Extract the largest face using RetinaFace [15].
3. Align the face using predicted landmarks.
4. Calculate the bounding box for the aligned face.
5. Enlarge the bounding box by a 1.3× margin.
6. Crop and resize the face to 224×224.
7. Save the image in lossless format.

Training data. Following the standard protocol, we train the model on the FaceForensics++ dataset [44] (FF++), which consists of 3600 videos, of which 720 are real videos, and the rest, 4×720, come from four different deepfake forgery methods. After preprocessing, we obtain approximately 115k frames. The dataset was released in three different compression levels. Following all of the preceding work, we use the c23 compression version of the dataset. We augment the training set using image augmentations such as random horizontal flipping, random affine transformations, Gaussian blurring, color jitter, and JPEG compression, all implemented in the `torchvision` package. In ablations, we also experimented with changing the training set to FFIW [67] and DSv2 [5].

Training samples pairing. We train the model on a dataset consisting of real-fake pairs, where each fake video is generated from its real counterpart. We hypothesize and empirically confirm in Sec. 4.4 that the dataset constructed in this way forces the network to learn the subtle, low-level artifacts of the manipulation process, rather than exploiting superficial, high-level differences between unrelated videos, thus preventing shortcut learning.

Validation data. We noticed that the FF++ validation set is very similar to the training set. That makes the comparison of different training experiments on such a validation ineffective. Each experiment achieves the best validation AUROC quickly, shows almost no signs of overfitting, and does not provide any estimate of cross-dataset generalization. The FF++ validation set does not provide crucial information on how the methods will perform when trained for an extended period. Therefore, we create a custom validation set, which consists of data from validation or training splits of other datasets, namely CDFv3 [35], DSv1 and DSv2 [5], and FFIW [67]. We ensure that test samples from these datasets are not present in the custom validation set.

4. Experiments

We demonstrate the generalizability of GenD by presenting a series of experiments. First, we introduce the diverse range of benchmark datasets and the evaluation metric. Next, we conduct a comprehensive cross-dataset evaluation, where we train the model on the FF++ dataset and compare its performance against numerous state-of-the-art (SOTA) methods on other benchmarks. Following this, a detailed ablation study is presented to systematically analyze the contribution of each component of the proposed method. We then investigated two hypotheses: one demonstrating the critical importance of training on paired real-fake data for robust generalization and another exploring the evolution of deepfake detection difficulty over the years, which shows that training on older, diverse datasets can be more beneficial than using solely recent datasets. We finish the experiments with robustness evaluation against common image degradation techniques.

We ensure the statistical significance of the results by repeating each experiment 5 times with different training

Table 1. Test datasets. The number of real and fake videos in the datasets. Negative values: the number of videos missing in our experiments, compared to the original datasets, due to face detector failures. Values in rows with datasets marked $*$ represent the number of videos randomly subsampled from the original dataset. Gen. means the number of generators used in the dataset.

Year	Dataset	Gen.	Real		Fake	
2019	FF++ [44]	4	140		560	
2019	DF [44]	1	140		140	
2019	F2F [44]	1	140		140	
2019	FS [44]	1	140		140	
2019	NT [44]	1	140		140	
2019	DFD [17]	5	363		3068	-2
2019	UADFV [60]	1	49		49	
2019	DFDC [16]	8	2500	-185	2500	-111
2020	FSh [32]	1	140		140	
2020	CDFv2 [34]	1	178		340	
2021	FFIW [67]	3	1738	-3	1738	-3
2021	KoDF [31] $*$	6	403		1106	
2021	FAVC [28]	4	500		20566	-22
2022	DFDM [27]	5	590	-2	1720	-2
2024	PGF [25]	10	762		13605	
2024	IDF [54] $*$	9	18834		2323	
2024	DSv1.1 [5]	5	1416		1497	
2025	DSv2 [5]	6	1863		1416	
2025	CDFv3 [35]	22	178		2540	-1

seeds. Every reported AUROC for GenD except for the Tab. 2 is averaged over 5 seeds; standard deviations can be found in the supplementary text.

4.1. Test benchmarks

For reporting the performance of the model in cross-dataset settings, we are using Celeb-DF-v2 (CDFv2) [34], Celeb-DF++ (CDFv3) [35], DeepFake Detection Challenge (DFDC) [16], Google's DFD dataset [17], Face Forensics in the Wild (FFIW) [67], DeepSpeak v1.1 (DSv1.1) and DeepSpeak v2.0 (DSv2) [5], Korean DeepFake Detection Dataset (KoDF) [31], FakeAVCeleb (FAVC) [28], DeepFakes from Different Models (DFDM) [27], PolyGlotFake (PGF) [25], and IDForge (IDF) [54]. The statistics for the test part of the datasets are shown in Tab. 1.

We are using the video-level area under the ROC curve (AUROC) as the main comparison and optimization metric. See the supplementary material for more evaluation metrics such as average precision (AP) or equal error rate (EER).

4.2. Cross-dataset evaluation

We compared the proposed model in a cross-dataset fashion following the standard protocol, that is, by training on FF++ and testing on the rest of the datasets as in previous work. Tab. 2 compares the video-level AUROC of the proposed model with SOTA approaches. AUROC values for the competing methods were taken directly from the original papers. The model achieves SOTA results on two datasets: CDFv2 and FFIW, and gets high results on the DFDC dataset.

The proposed method achieves these results simply by averaging individual frame softmax probabilities, without complex post-processing or architectural modifications as described in Sec. 3.1. This contrasts with many contemporary methods that rely on complex multimodal analyses, such as modeling temporal inconsistencies between video frames or detecting audio-video discrepancies [20–22, 37, 59, 66]. The performance of the proposed method suggests that the features of the visual encoder, when properly tuned, are highly effective for this detection task.

We also evaluated the generalization of the proposed method against current SOTA models across a wide range of benchmark datasets. For this experiment, we selected the most recent open-source SOTA models that have made their code and pre-trained weights publicly available, namely ForAda [13] and Effort [58]. To ensure that each model was used as intended by its original authors, we utilized their officially released weights and integrated their data preprocessing pipelines which were used during model training. After reproducing the results on the test datasets that were used in these papers, we obtained the same or very similar results, presented in Tab. 5. The in-distribution results on FF++ are shown in Tab. 3.

Subsequently, we compared the pre-trained models with the proposed model on the comprehensive collection of test datasets from Tab. 1. The results of this large-scale evaluation are presented in Tab. 4. This methodology ensures that the methods were compared on the same data. From the results of this experiment, it is seen that the proposed approach generalized better, which is proved by higher average performance across all datasets.

4.3. Ablation studies

We conducted ablation experiments to understand the contribution of each component; see Tab. 6. There are three setups tested for every backbone. Setup 3 is the proposed GenD:

1. **Baseline**: the setup is similar to [40]. We freeze the backbone and train a linear classifier with cross-entropy loss on top of the feature space of the classification token.
2. **Baseline + LN**: adds tuning of Layer Normalization parameters to setup 1.
3. **Baseline + LN + UA**: L2-normalizes classification token and adds uniformity and alignment losses to setup 2.

Layer Normalization tuning. As shown in Tab. 6, LN-tuning consistently achieves substantial gains in AUROC across all benchmarks.

Table 2. Video-level AUROC (%) in cross-dataset testing of models trained on the FF++ dataset. Results of other methods are taken from their original papers. Values with superscript citations are taken from the papers referenced in superscripts. Video input means that the model takes a sequence of frames and models temporal correlations between them.

Model	Year	Publication	Input	Backbone	CDFv2	DFD	DFDC	FFIW
LipForensics [20]	2021	CVPR	Video	ResNet-18	82.4	–	73.5	–
FTCN [66]	2021	ICCV	Video	3D ResNet-50	86.9	–	74.0	74.5[45]
RealForensics [21]	2022	CVPR	Video	Modified CSN	86.9	–	75.9	–
SBI [45]	2022	CVPR	Frame	EFNB4	93.2	82.7	72.4	84.8
AUNet [64]	2023	CVPR	Video	Xception+ART	92.8	**99.2**	73.8	81.5
StyleDFD [12]	2024	CVPR	Video	3D ResNet-50	89.0	96.1	–	–
LSDA [57]	2024	CVPR	Frame	EFNB4	91.1	–	77.0	72.4[58]
LAA-Net [38]	2024	CVPR	Frame	EFNB4	95.4	98.4	86.9	–
AltFreezing [53]	2024	CVPR	Video	3D ResNet-50	89.5	98.5	–	–
NACO [63]	2024	ECCV	Video	ViT-B/16	89.5	–	76.7	–
RAE [50]	2024	ECCV	Frame	ViT-B/16	95.5	99.0	80.2	–
TALL++ [55]	2024	IJCV	Video	Swin-B	92.0	–	78.5	–
ProDet [11]	2024	NeurIPS	Frame	EFNB4	92.5	–	77.0	–
UDD [18]	2025	AAAI	Frame	CLIP ViT-B/16	93.1	95.5	81.2	–
P&P [59]	2025	CVPR	Video	CLIP ViT-L/14	94.7	96.5	84.3	92.1
DFD-FCG [22]	2025	CVPR	Video	CLIP ViT-L/14	95.0	–	81.8	–
ForAda [13]	2025	CVPR	Frame	CLIP ViT-L/14	95.7	97.2	**87.2**	–
Effort [58]	2025	ICML	Frame	CLIP ViT-L/14	95.6	96.5	84.3	92.1
GenD (CLIP)	2025	–	Frame	CLIP ViT-L/14	**96.0**	97.0	87.1	92.8
GenD (PE)	2025	–	Frame	PE_{core}L	95.8	96.5	81.6	93.3
GenD (DINO)	2025	–	Frame	DINOv3 ViT-L/16	92.2	96.6	84.7	**94.5**

Table 3. Video-level test AUROC (%) on in-domain FF++ dataset calculated by us. GenD results are the means over five seeds.

Method	DF	F2F	FS	NT	Mean
ForAda [13]	**99.7**	97.0	98.6	91.9	96.8
Effort [58]	99.4	93.2	98.4	84.6	93.9
GenD (CLIP)	99.5	98.1	98.7	95.5	98.0
GenD (PE)	**99.7**	99.3	99.1	**97.5**	**98.9**
GenD (DINO)	99.6	**99.4**	**99.4**	96.7	98.8

We explore further parameter-efficient fine-tuning (PEFT) strategies, the goal of which is to modify only a relatively small number of parameters while keeping most parameters untouched, thus efficiently introducing several degrees of freedom. PEFT is more effective for generalization than full fine-tuning (FFT) in low-data regimes [23], which is particularly applicable to deepfake detection, where collected datasets and benchmarks have a limited number of samples. This work explores three PEFT methods: LoRA [26], LN-tuning [33, 42], and bias tuning (BitFit) [7, 33].

We found that FFT leads to rapid overfitting and poor generalization across datasets. This observation is consistent with recent findings from Effort [58] that FFT degrades the generalization performance.

Similarly to [58], we fine-tuned the parameters of every transformer weight matrix using LoRA to preserve the generalization. Setting LoRA to rank one, the lowest possible, allowed the model to reach a near-perfect training AUROC (99.99%) in two epochs, resulting in rapid overfitting.

We observed that BitFit [7] and LN-tuning [42] have stable training dynamics by first improving the validation AUROC, reaching a peak, and overfitting after some epochs. In comparison, FFT and LoRA show no improvement in validation AUROC even after a single training epoch.

Experimental results showed that in combination with other components, LN-tuning achieves the best performance and prevents overfitting compared to other PEFT strategies, such as BitFit, LoRA, or FFT. These results suggest that LN tuning works as a good tuning strategy for this task. A more detailed ablation comparing these techniques can be found in the supplementary material.

Uniformity and Alignment. We hypothesized that representations that effectively exploit the latent space could help us improve generalization. We used uniformity and alignment loss [52], leading to slower overfitting and better validation AUROC. We suppose that adding these losses helps

Table 4. Cross-dataset video-level AUROC (%) for reproduced methods. The highest score in each column is in bold. Results for GenD are the averages over five training seeds; standard deviations can be found in the supplementary.

Method	2019 UADFV	2019 DFD	2019 DFDC	2020 FSh	2020 CDFv2	2021 FFIW	2021 KoDF	2021 FAVC	2022 DFDM	2024 PGF	2024 IDF	2024 DSv1.1	2025 DSv2	2025 CDFv3	Mean
ForAda [13]	**99.4**	**97.2**	**87.3**	82.0	**95.7**	90.6	88.2	93.1	97.1	86.6	90.8	81.8	72.8	75.6	88.4
Effort [58]	97.4	95.2	85.4	**91.2**	93.2	92.5	88.1	92.4	98.2	84.9	96.0	82.1	64.4	78.7	88.5
GenD (CLIP)	99.2	96.4	86.4	86.6	94.6	91.5	84.9	96.0	99.6	89.6	97.8	**90.1**	77.7	85.9	91.2
GenD (PE)	97.7	96.8	82.2	87.6	95.0	**93.7**	85.1	97.3	98.3	92.3	97.9	87.8	78.6	**89.5**	91.4
GenD (DINO)	98.6	96.2	85.6	88.8	92.5	92.9	**89.7**	**98.4**	**99.8**	**92.4**	**98.2**	86.9	**79.4**	83.5	**91.6**

Table 5. Comparison of Reported / Reproduced video-level AUROC (%) for reimplemented deepfake detectors.

Method	CDFv2	DFD	DFDC	FFIW
ForAda [13]	95.7 / 95.7	97.2 / 97.2	87.2 / 87.3	– / 90.6
Effort [58]	95.6 / 93.2	96.5 / 95.2	84.3 / 85.4	92.1 / 92.5

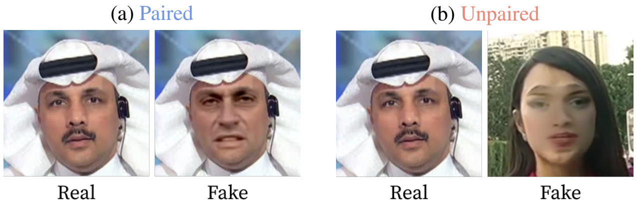

Figure 2. Samples from (a) Paired and (b) Unpaired datasets.

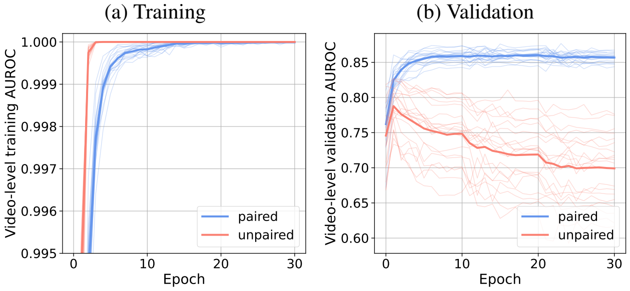

Figure 3. Video-level AUROC for (a) Training and (b) Validation, averaged over 20 randomly sampled paired and 20 unpaired datasets from FF++. The image encoder is PE_{core}L.

to utilize the hyperspherical space more effectively, resulting in improved generalization.

4.4. Importance of training on paired dataset

A fundamental question in training deepfake detectors is how to construct training data more effectively to promote generalization. In this section, we empirically validate the hypothesis that achieving better generalization requires *a training dataset in which each fake video has a real counterpart from which it was produced*.

To empirically validate this hypothesis, we designed an experiment with two differently constructed training datasets. The first dataset, denoted as paired, consists of data where for each fake video there exists its pair – a real counterpart from which this fake was created. In the second dataset, unpaired, fake videos never have its pair – a real counterpart. Examples of samples from these two datasets can be found in Fig. 2.

In particular, we generate 20 different paired and 20 unpaired dataset splits from the FF++ training set. Every split shares the real part but has a different fake part. In the paired dataset, every fake video shares the background with a video from the real part. In an unpaired dataset, every fake video does not have a corresponding real video with the same background, see Fig. 2. We ensure that the number of fake and real videos between the paired and unpaired datasets is the same.

We plot AUROC for training and validation progress of PE_{core}L image encoder in Fig. 3. Models trained on the unpaired datasets consistently overfit more quickly and achieve a lower validation AUROC, as indicated by the sharper increase in training and the decline in the validation AUROC. In contrast, models trained using paired datasets achieve a higher validation AUROC and have a lower tendency to overfit throughout the training process. We observe the same behavior on other pre-trained backbones and training sets; see supplementary material for more experiments.

We attribute the discrepancy in validation AUROC between training on paired and unpaired datasets to shortcut learning. When the model is presented with real and fake videos from the same source, it is constrained to focus on the low-level inconsistencies and artifacts introduced by the deepfake algorithm, learning a more robust and generalizable representation left by the forgery algorithm. This experiment highlights the critical importance of data pairing as a fundamental principle for developing deepfake detectors that can generalize to unseen data.

Table 6. Ablation study. Impact of model components on test video-level AUROC in cross-dataset settings. Each setup was trained on the FF++ training set. Reported values are averages over 5 training seeds. Setups #3, #6, #9 are the proposed GenD.

#	Setup	UADFV	DFD	DFDC	FSh	CDFv2	FFIW	KoDF	FAVC	DFDM	PGF	IDF	DSv1.1	DSv2	CDFv3	Mean
1	CLIP + LP	94.6	89.2	75.3	77.6	74.6	80.7	81.8	83.0	77.8	62.7	68.7	64.5	57.8	75.6	76.0
2	#1 + LN	99.0	96.1	85.4	**86.9**	93.8	**91.5**	84.4	95.5	99.3	87.5	95.0	89.1	74.0	**86.8**	90.3
3	#2 + UA	**99.2**	**96.4**	**86.4**	86.6	**94.6**	91.5	**84.9**	**96.0**	**99.6**	**89.6**	**97.8**	**90.1**	**77.7**	85.9	**91.2**
4	PE$_{\text{core}}$L + LP	76.1	76.4	67.2	80.3	72.2	75.7	65.6	71.3	73.3	52.8	74.6	65.6	59.5	79.5	70.7
5	#4 + LN	**98.4**	96.4	81.9	**88.1**	94.2	93.0	84.5	96.5	97.9	91.2	97.7	**88.0**	77.0	**89.9**	91.0
6	#5 + UA	97.7	**96.8**	**82.2**	87.6	**95.0**	**93.7**	**85.1**	**97.3**	**98.3**	**92.3**	**97.9**	87.8	**78.6**	89.5	**91.4**
7	DINOv3L + LP	84.2	79.8	69.0	71.9	68.2	74.7	75.4	81.1	85.9	65.0	82.0	72.2	70.2	68.3	74.9
8	#7 + LN	97.3	94.8	84.4	88.7	91.7	90.4	**90.6**	98.4	98.8	**93.6**	97.5	**89.1**	**82.4**	80.1	91.3
9	#8 + UA	**98.6**	**96.2**	**85.6**	**88.8**	**92.5**	**92.9**	89.7	**98.4**	**99.8**	92.4	**98.2**	86.9	79.4	**83.5**	**91.6**

4.5. Evolution of detection difficulty over the years

We hypothesize that successfully training a deepfake detection system requires exposure not only to the most recent forgery techniques but also to older ones. Relying solely on recent datasets may result in poor generalization, as detectors might overfit to the artifacts specific to modern generation methods while ignoring the broader spectrum of manipulations seen over time.

To test this, we trained three models on datasets introduced in different years: FF++ (2019), FFIW (2021), and DSv2 (2025). These models were then evaluated on a diverse set of benchmarks spanning from 2019 to 2025, with test AUROC plotted against dataset release year in Fig. 4.

Although the model trained on the DSv2 dataset performs well on its own test set, it generalizes poorly to older benchmarks such as FF++, DFDM, and CDFv2. In contrast, the model trained on the older but diverse FF++ dataset achieves competitive or even superior performance across a wider range of test datasets, including many recent ones.

These results indicate that training on only recent deepfakes does not guarantee a strong generalization. Instead, models trained on older datasets with diverse manipulation techniques often perform more robustly across time. This suggests that newer deepfakes do not entirely subsume or replace the challenges posed by earlier generation techniques.

To build a generalizable deepfake detector, it is crucial to curate training datasets that span both old and new forgery techniques. Focusing exclusively on recent data risks overfitting to current trends and missing broader forgery patterns.

4.6. Robustness to image degradations

We evaluated the robustness of the GenD to commonly used image degradations, which can be used to remove forgery patterns and decrease the performance of deepfake detection. For statistical significance, we average the results of GenD trained using 5 different training seeds. The results are shown in Fig. 5. Visual examples of images processed over various levels of degradations can be found in the supplementary.

We selected three methods, such as Gaussian blur, JPEG compression, and image resizing. For Gaussian blur, we varied the sigma σ and selected the proportional kernel size $k = 2 \cdot \lceil 3\sigma \rceil + 1$. For resizing, we took the facial image and resized it to 224^2, 114^2, or 64^2 pixels by using 6 different interpolation techniques: nearest, lanczos, bilinear, bicubic, box, hamming. After that, each downsized image was upsampled to the model's input resolution by every method's preprocessing technique. Results are averaged over the interpolation methods.

5. Limitations and Future Work

Although the proposed method, GenD, demonstrates a strong generalization, it is important to acknowledge its limitations, which also present avenues for future research.

Robustness to perturbations and adversarial attacks. One limitation is that the model's performance against targeted adversarial attacks [49] specifically designed to evade detection has not been evaluated.

Limited temporal modeling. We tried to model the temporal signal using a simple single-layer self-attention block between classification tokens from different frames in a video. Nevertheless, our experiments did not show any noticeable improvements compared to the proposed method, which operates at the frame level and aggregates predictions via simple averaging. Although this allows for architectural simplicity, it ignores temporal dynamics that could offer additional forgery cues, such as inconsistencies in lip synchronization or frame-level motion artifacts.

Dependence on facial region. Our preprocessing pipeline assumes access to a detectable and alignable face. Videos with occlusions, extreme poses, or low resolution may lead to failures in face detection or misaligned crops, which can degrade performance.

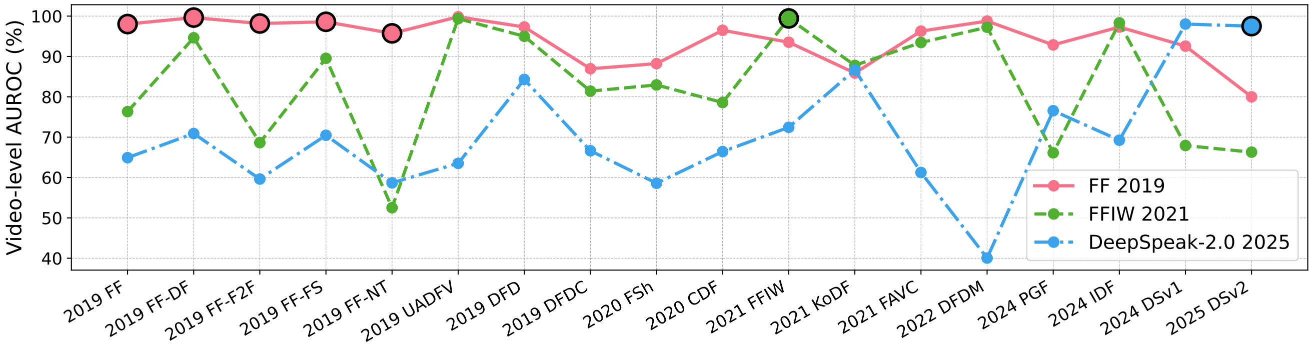

Figure 4. Evolution of detection difficulty over time. Each video-level AUROC is computed on the test set of the corresponding benchmark. Each curve represents a model trained on a single dataset. Highlighted circles indicate the model's in-dataset performance.

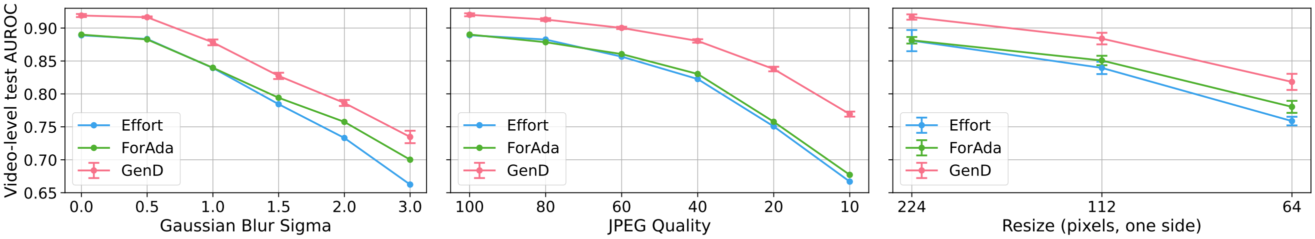

Figure 5. Robustness to image degradations for GenD (PE$_{\text{core}}$L), ForAda [13], and Effort [58]. Video-level AUROC (%) is calculated across all 14 test datasets. Error bars for GenD are computed from models trained with 5 different training seeds. In the resize, we also average every method across 6 interpolation strategies: nearest, lanczos, bilinear, bicubic, box, hamming.

Limited demographic diversity in training data. Training data lacks sufficient diversity in terms of identity, ethnicity, age, and gender. This can result in biased detection performance, where the model may generalize poorly to underrepresented demographic groups. Ensuring demographic balance and evaluating fairness across subgroups are important next steps toward responsible deployment. See the supplementary for typical failure cases.

Incremental learning. The current framework does not address the challenge of incremental learning. As new forgery methods emerge, the model would require complete retraining. Developing a framework that can adapt to new manipulation techniques without catastrophically forgetting previously learned ones is a critical direction for real-world applicability.

6. Conclusions

We addressed the persistent challenge of generalization in deepfake detection. Although recent approaches explored increasingly complex architectural and parameter-space modifications to adapt large foundation models, we show that generalization is achieved by tuning existing parameters of the image encoder (CLIP, PE, DINO). We demonstrated that such encoder is transformed into a state-of-the-art deepfake detection model using three additional components: LN-tuning, L2 normalization, and a linear classifier. State-of-the-art results are achieved within a few hours of training on a single A100 GPU, making the proposed approach computationally efficient and reproducible.

Our extensive evaluation, conducted on one of the most comprehensive collections of deepfake benchmarks to date, confirms the efficacy of this approach, consistently matching or outperforming more complex state-of-the-art methods. Beyond performance, our work provides several insights for the community. We experimentally demonstrated that: 1) the difficulty of academic datasets did not substantially increase over time, and training on old data still gives strong generalization capabilities, compared to, for example, training on only recent datasets, and 2) training with paired real-fake data from the same source video is critical for mitigating shortcut learning and achieving better generalization. These findings indicate that data pairing and training set diversity are key factors in the development of generalizable deepfake detectors.

Acknowledgments. The research was supported by the National Recovery Plan project CEDMO 2.0 NPO (MPO 60273/24/21300/21000), the EC Digital Europe Programme project CEDMO 2.0 no. 101158609, the CTU Student Grant SGS23/173/OHK3/3T/13, and the OP VVV project CZ.02.1.01/0.0/0.0/16_019/0000765 RCI.

References

[1] Shruti Agarwal, Hany Farid, Yuming Gu, Mingming He, Koki Nagano, and Hao Li. Protecting world leaders against deep fakes. In *CVPR Workshops*, 2019. 2

[2] Vishal Asnani, Xi Yin, Tal Hassner, Sijia Liu, and Xiaoming Liu. Proactive image manipulation detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15386–15395, 2022. 2

[3] Vishal Asnani, Xi Yin, Tal Hassner, and Xiaoming Liu. MALP: manipulation localization using a proactive scheme. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12343–12352, 2023. 2

[4] Sanoojan Baliah, Qinliang Lin, Shengcai Liao, Xiaodan Liang, and Muhammad Haris Khan. Realistic and efficient face swapping: A unified approach with diffusion models. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1062–1071. IEEE, 2025. 1

[5] Sarah Barrington, Matyas Bohacek, and Hany Farid. Deepspeak dataset v1.0. *arXiv preprint arXiv:2408.05366*, 2024. 3, 4

[6] Filippo Bartolucci, Iacopo Masi, and Giuseppe Lisanti. Perturb, attend, detect and localize (PADL): Robust proactive image defense. *IEEE Access*, 2025. 2

[7] Elad Ben Zaken, Yoav Goldberg, and Shauli Ravfogel. BitFit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1–9, Dublin, Ireland, 2022. Association for Computational Linguistics. 5

[8] Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Rasheed, et al. Perception encoder: The best visual embeddings are not at the output of the network. *arXiv preprint arXiv:2504.13181*, 2025. 1, 2

[9] Stella Bounareli, Christos Tzelepis, Vasileios Argyriou, Ioannis Patras, and Georgios Tzimiropoulos. One-shot neural face reenactment via finding directions in GAN's latent space. *International Journal of Computer Vision*, 132(8): 3324–3354, 2024. 1

[10] Sergi D Bray, Shane D Johnson, and Bennett Kleinberg. Testing human ability to detect 'deepfake' images of human faces. *Journal of Cybersecurity*, 9(1):tyad011, 2023. 1

[11] Jikang Cheng, Zhiyuan Yan, Ying Zhang, Yuhao Luo, Zhongyuan Wang, and Chen Li. Can we leave deepfake data behind in training deepfake detector? In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 5

[12] 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. 5

[13] Xinjie Cui, Yuezun Li, Ao Luo, Jiaran Zhou, and Junyu Dong. Forensics adapter: Adapting CLIP for generalizable face forgery detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19207–19217, 2025. 2, 4, 5, 6, 8

[14] Google DeepMind. Veo 3, 2025. https://deepmind.google/models/veo/. 2

[15] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5203–5212, 2020. 3

[16] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge (DFDC) dataset. *arXiv preprint arXiv:2006.07397*, 2020. 4

[17] Nicholas Dufour, Andrew Gully, Per Karlsson, Alexey Victor Vorbyov, Thomas Leung, Jeremiah Childs, and Christoph Bregler. Deepfakes Detection Dataset by Google & Jigsaw. https://research.google/blog/contributing-data-to-deepfake-detection-research/, 2019. 4

[18] Xinghe Fu, Zhiyuan Yan, Taiping Yao, Shen Chen, and Xi Li. Exploring unbiased deepfake detection via token-level shuffling and mixing. *arXiv preprint arXiv:2501.04376*, 2025. 2, 5

[19] Angeliki Giannou, Shashank Rajput, and Dimitris Papailiopoulos. The expressive power of tuning only the normalization layers. *arXiv preprint arXiv:2302.07937*, 2023. 2

[20] 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. 2, 4, 5

[21] Alexandros Haliassos, Rodrigo Mira, Stavros Petridis, and Maja Pantic. Leveraging real talking faces via self-supervision for robust forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14950–14962, 2022. 5

[22] Yue-Hua Han, Tai-Ming Huang, Kai-Lung Hua, and Jun-Cheng Chen. Towards more general video-based deepfake detection through facial component guided adaptation for foundation model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22995–23005, 2025. 2, 4, 5

[23] Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. Parameter-efficient fine-tuning for large models: A comprehensive survey. *Transactions on Machine Learning Research*, 2024. 5

[24] Ammarah Hashmi, Sahibzada Adil Shahzad, Chia-Wen Lin, Yu Tsao, and Hsin-Min Wang. Unmasking illusions: Understanding human perception of audiovisual deepfakes. *arXiv preprint arXiv:2405.04097*, 2024. 1

[25] Yang Hou, Haitao Fu, Chunkai Chen, Zida Li, Haoyu Zhang, and Jianjun Zhao. Polyglotfake: A novel multilingual and multimodal deepfake dataset. In *International Conference on Pattern Recognition*, pages 180–193. Springer, 2024. 4

[26] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen.

LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. 5

[27] Shan Jia, Xin Li, and Siwei Lyu. Model attribution of face-swap deepfake videos. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 2356–2360. IEEE, 2022. 4

[28] Hasam Khalid, Shahroz Tariq, Minha Kim, and Simon S Woo. FakeAVCeleb: A novel audio-video multimodal deepfake dataset. *arXiv preprint arXiv:2108.05080*, 2021. 4

[29] Sohail Ahmed Khan and Duc-Tien Dang-Nguyen. Clipping the deception: Adapting vision-language models for universal deepfake detection. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, pages 1006–1015, 2024. 2

[30] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 3

[31] Patrick Kwon, Jaeseong You, Gyuhyeon Nam, Sungwoo Park, and Gyeongsu Chae. Kodf: A large-scale korean deepfake detection dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10744–10753, 2021. 4

[32] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. Advancing high fidelity identity swapping for forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5074–5083, 2020. 4

[33] Ming Li, Jike Zhong, Chenxin Li, Liuzhuozheng Li, Nie Lin, and Masashi Sugiyama. Vision-language model fine-tuning via simple parameter-efficient modification. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14394–14410, Miami, Florida, USA, 2024. Association for Computational Linguistics. 5

[34] 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. 4

[35] Yuezun Li, Delong Zhu, Xinjie Cui, and Siwei Lyu. Celeb-df++: A large-scale challenging video deepfake benchmark for generalizable forensics. *arXiv preprint arXiv:2507.18015*, 2025. 3, 4

[36] Huan Liu, Zichang Tan, Chuangchuang Tan, Yunchao Wei, Jingdong Wang, and Yao Zhao. Forgery-aware adaptive transformer for generalizable synthetic image detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10770–10780, 2024. 2

[37] Weifeng Liu, Tianyi She, Jiawei Liu, Boheng Li, Dongyu Yao, and Run Wang. Lips are lying: Spotting the temporal inconsistency between audio and visual in lip-syncing deepfakes. *Advances in Neural Information Processing Systems*, 37:91131–91155, 2024. 2, 4

[38] Dat Nguyen, Nesryne Mejri, Inder Pal Singh, Polina Kuleshova, Marcella Astrid, Anis Kacem, Enjie Ghorbel, and Djamila Aouada. Laa-net: Localized artifact attention network for quality-agnostic and generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17395–17405, 2024. 5

[39] Huy H Nguyen, Junichi Yamagishi, and Isao Echizen. Exploring self-supervised vision transformers for deepfake detection: A comparative analysis. In *Proceedings of the IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10, 2024. 2

[40] Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24480–24489, 2023. 1, 2, 4

[41] Ziqiao Peng, Jiwen Liu, Haoxian Zhang, Xiaoqiang Liu, Songlin Tang, Pengfei Wan, Di Zhang, Hongyan Liu, and Jun He. Omnisync: Towards universal lip synchronization via diffusion transformers. *arXiv preprint arXiv:2505.21448*, 2025. 1

[42] Wang Qi, Yu-Ping Ruan, Yuan Zuo, and Taihao Li. Parameter-efficient tuning on layer normalization for pre-trained language models. *arXiv preprint arXiv:2211.08682*, 2022. 1, 2, 5

[43] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PmLR, 2021. 1, 2

[44] 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. 3, 4

[45] 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. 2, 5

[46] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025. 1, 2

[47] Leslie N Smith. Cyclical learning rates for training neural networks. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 464–472. IEEE, 2017. 3

[48] Synthesia, 2024. `https://www.synthesia.io`. 2

[49] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014. 7

[50] Jiahe Tian, Cai Yu, Xi Wang, Peng Chen, Zihao Xiao, Jiao Dai, Jizhong Han, and Yesheng Chai. Real appearance modeling for more general deepfake detection. In *European Conference on Computer Vision*, pages 402–419. Springer, 2024. 5

[51] Taha ValizadehAslani and Hualou Liang. Layernorm: A key component in parameter-efficient fine-tuning. *arXiv preprint arXiv:2403.20284*, 2024. 2

[52] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International conference on machine learning*, pages 9929–9939. PMLR, 2020. 2, 5

[53] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, and Houqiang Li. Altfreezing for more general video face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4129–4138, 2023. 5

[54] Junhao Xu, Jingjing Chen, Xue Song, Feng Han, Haijun Shan, and Yu-Gang Jiang. Identity-driven multimedia forgery detection via reference assistance. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 3887–3896, 2024. 4

[55] Yuting Xu, Jian Liang, Lijun Sheng, and Xiao-Yu Zhang. Learning spatiotemporal inconsistency via thumbnail layout for face deepfake detection. *International Journal of Computer Vision*, 132(12):5663–5680, 2024. 5

[56] Zhiyuan Yan, Yong Zhang, Xinhang Yuan, Siwei Lyu, and Baoyuan Wu. Deepfakebench: A comprehensive benchmark of deepfake detection. In *Advances in Neural Information Processing Systems*, pages 4534–4565. Curran Associates, Inc., 2023. 3

[57] 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. 5

[58] Zhiyuan Yan, Jiangming Wang, Peng Jin, Ke-Yue Zhang, Chengchun Liu, Shen Chen, Taiping Yao, Shouhong Ding, Baoyuan Wu, and Li Yuan. Orthogonal subspace decomposition for generalizable AI-generated image detection. In *Proceedings of the International Conference on Machine Learning*, 2025. 2, 4, 5, 6, 8

[59] Zhiyuan Yan, Yandan Zhao, Shen Chen, Mingyi Guo, Xinghe Fu, Taiping Yao, Shouhong Ding, Yunsheng Wu, and Li Yuan. 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*, pages 12615–12625, 2025. 4, 5

[60] Xin Yang, Yuezun Li, and Siwei Lyu. Exposing deep fakes using inconsistent head poses. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8261–8265. IEEE, 2019. 4

[61] Yuankun Yang, Chenyue Liang, Hongyu He, Xiaoyu Cao, and Neil Zhenqiang Gong. Faceguard: Proactive deepfake detection. *arXiv preprint arXiv:2109.05673*, 2021. 2

[62] Rui Zhai, Rongrong Ni, Yu Chen, Yang Yu, and Yao Zhao. Defending fake via warning: Universal proactive defense against face manipulation. *IEEE Signal Processing Letters*, 30:1072–1076, 2023. 2

[63] Daichi Zhang, Zihao Xiao, Shikun Li, Fanzhao Lin, Jianmin Li, and Shiming Ge. Learning natural consistency representation for face forgery video detection. In *European Conference on Computer Vision*, pages 407–424. Springer, 2024. 5

[64] Tianchen Zhao, Xiang Xu, Mingze Xu, Hui Ding, Yuanjun Xiong, and Wei Xia. Learning self-consistency for deepfake detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15023–15033, 2021. 5

[65] Yuan Zhao, Bo Liu, Tianqing Zhu, Ming Ding, Xin Yu, and Wanlei Zhou. Proactive image manipulation detection via deep semi-fragile watermark. *Neurocomputing*, 585:127593, 2024. 2

[66] Yinglin Zheng, Jianmin Bao, Dong Chen, Ming Zeng, and Fang Wen. Exploring temporal coherence for more general video face forgery detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15044–15054, 2021. 4, 5

[67] Tianfei Zhou, Wenguan Wang, Zhiyuan Liang, and Jianbing Shen. Face forensics in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5778–5788, 2021. 3, 4