Learning Forgery Signals from Synthetic Depth Targets for Generalizable Deepfake Detection

Marko Brodarič[1,4] Walter Scheirer[2] Deepak Kumar Jain[3] Peter Peer[4] Vitomir Štruc[1]
[1]University of Ljubljana, Faculty of Electrical Engineering, Slovenia
[2]University of Notre Dame, USA [3]Dalian University of Technology, China
[4]University of Ljubljana, Faculty of Computer and Information Science, Slovenia
marko.brodaric@fe.uni-lj.si

Abstract

Deepfake detection remains challenging under distribution shifts caused by unseen generation pipelines, post-processing, and unconstrained capture conditions. While most existing detectors primarily reason in RGB appearance space, facial manipulations also disturb geometric structure and consistency, making depth an attractive complementary cue. Existing depth-aware methods, however, typically use depth as an auxiliary input or guidance signal and may therefore underutilize the representational richness of modern monocular depth foundation models. To address this limitation, we propose a mechanism that adapts a pretrained monocular depth foundation model into a forgery-aware representation extractor for deepfake detection. Based on the adapted model, we then propose a novel depth-driven detector, termed FADepth, that uses the resulting depth representation as the primary source of evidence and complements it with RGB appearance cues. Extensive experiments on six benchmark datasets and with 24 state-of-the-art detectors show that FADepth yields highly competitive performance, achieving an overall mAUC of 88.70*. The source code of the model is available at* https://github.com/markobrodaric/FADepth.

1. Introduction

The rapid progress of deep generative models has enabled the creation of highly realistic manipulated facial imagery and videos, commonly referred to as deepfakes, raising concerns related to fraud, misinformation, and the loss of trust in digital media. As a result, deepfake detection has become an increasingly important component of modern biometric and multimedia forensics pipelines [26, 32]. While recent detectors achieve strong performance on the datasets and manipulation types seen during training, they often degrade under distribution shifts caused by unseen generation pipelines, post-processing, and unconstrained capture conditions [4, 36, 37]. To address this challenge, a growing body of work focuses on generalizable detectors that reduce reliance on generator-specific artifacts through pseudo-deepfake generation, manipulation-agnostic learning, anomaly detection methodology and more robust representation design [1, 13, 14, 22, 29, 38, 40].

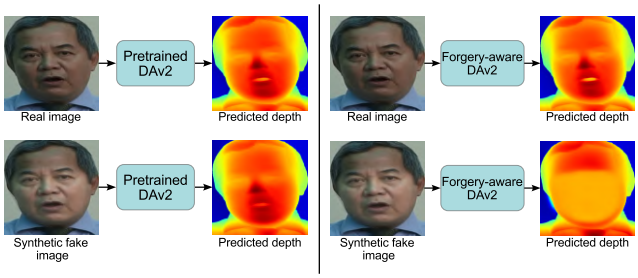

Figure 1. **Learning forgery-aware depth prediction for deepfake detection.** We introduce a novel concept, where a pretrained depth-prediction model (e.g., DepthAnything; DAv2) is adapted, such that altered facial regions produce detectable depth-responses while the depth predictions for authentic content remain unchanged. Based on this idea, we develop a depth-centric deepfake detector, named FADepth, that leverages the forgery-aware depth-prediction model for generalizable deepfake detection.

Despite this progress, most existing detectors still reason primarily in RGB appearance space. This is limiting, since facial manipulations do not only alter texture and color statistics, but also disturb the geometric structure and consistency of the face. Depth information is therefore a natural complementary cue, as it captures coarse 3D facial layout and local geometric variations that are not directly reflected in RGB appearance. Recent works have indeed shown that depth information can support deepfake detection [20, 34]. However, these approaches typically treat depth as an auxiliary signal for standard RGB-based deepfake detectors, e.g., through auxiliary depth maps, depth-guided objectives, or RGB-depth inconsistency modeling. Such formulations demonstrate the value of geometric cues, but still largely underutilize the representational potential of depth information as a key cue for deepfake detection.

At the same time, recent monocular depth foundation models, such as Depth Anything V2 (DAv2) [39], have shown that large-scale pretraining can yield highly trans-

ferable geometric representations. We argue that such representations are highly relevant for deepfake detection, but only if they are encouraged to encode forgery-specific signals rather than optimized solely for depth estimation. Off-the-shelf depth models are trained to predict scene geometry through depth, not to expose manipulated facial regions. Directly using their depth outputs may therefore miss much of their forensic value. This raises the key challenge addressed in this work: How to adapt a pretrained depth representation into a forgery-aware and complementary feature space for deepfake detection without requiring ground-truth depth annotations for manipulated faces?

To address this challenge, we propose to adapt a pretrained depth representation using synthetic depth supervision derived from manipulated training samples, so that it becomes sensitive to forgery-specific signals, as illustrated in Fig. 1. Following recent generalizable deepfake detection literature [13, 29], we implement this idea by generating samples with so-called *pseudo manipulations* that simulate common deepfake artifacts. A frozen pretrained depth model is then used to obtain reference depth predictions, which are converted into synthetic targets such that authentic samples retain their original depth structure while manipulated regions receive attenuated depth responses. Training on these targets encourages the adapted model to encode forgery-induced geometric inconsistencies in its representation. We incorporate this idea into a powerful deepfake detector termed **FADepth** (**F**orgery-**A**ware **Depth** detector), which combines forgery-aware depth features with complementary RGB evidence. Here, the adapted depth representation serves as the primary source of geometric evidence, while an RGB branch contributes appearance-based cues. This design preserves both generic geometric priors and forgery-aware depth signals, and achieves strong cross-dataset generalization, reaching an overall mAUC of 88.70 across six benchmarks without being trained on any of them directly. Through the research presented in this work, we make the following key contributions in this paper:

- We introduce the idea of learning forgery signals in pretrained depth representations through synthetic depth supervision, enabling depth models to become useful for deepfake detection without ground-truth depth annotations for manipulated faces.
- We develop FADepth, a depth-centric deepfake detector that incorporates the adapted depth representation as the primary source of evidence and combines it with complementary RGB cues in a unified framework.
- We show that learning forgery signals in the depth feature space yields representations that preserve geometry while producing localized responses to manipulated regions, enabling strong generalization and state-of-the-art performance in deepfake detection.

2. Related work

This section briefly reviews some of the most relevant and closely related works in generalizable deepfake detection and depth-aware manipulation analysis. For a more comprehensive coverage of the field, the reader is referred to some of the excellent surveys out there [26, 31, 32, 35]

Data–Centric Approaches. A central challenge in deepfake detection is generalization to unseen manipulations and real-world conditions. While early detectors often performed well under in-distribution evaluations, they typically relied on forgery-specific traces that transfer poorly across datasets and post-processing pipelines. A prominent strategy to address this limitation is to train with *synthetic or pseudo manipulations* that simulate inconsistencies shared across forgeries, rather than generator-specific fingerprints. Face X-Ray [17], for example, introduced blending-boundary supervision to emphasize compositing artifacts, while SBI [29] extended this idea through self-blending. SeeABLE [13] explored softened discrepancy learning in a one-class setting, and more recent work [16] further advances this direction with multivariate soft blending and CLIP-based alignment. Related approaches, including CORE [25], PCL+I2G [41], AltFreezing [36], LSDA [37], ProDet [4], and TSRL [15], highlight that both training strategy and data construction are critical for ensuring competitive cross-dataset generalization.

Manipulation–Agnostic Representations. Complementary to the above data-simulation approaches, several recent methods aim to learn manipulation-agnostic representations that capture shared evidence across different forgery processes [2, 8]. UCF [38], for instance, explicitly targets features common across manipulation types, CFM [22] seeks to discover critical clues beyond prior forgery knowledge, and FDML [40] disentangles content, identity, and artifact information. Related efforts such as IID [11], LESB [30], and RFFR [28] further emphasize reducing bias and improving robustness through better representation learning. Collectively, these works suggest that robust detection depends on learning stable cues that persist across generators, identities, and post-processing conditions.

Cue–Driven Detection. Another line of work focuses on capturing cues that are difficult for forgeries to preserve. Frequency-based detectors, such as SRM-based models [23] and SPSL [21], exploit spectral inconsistencies introduced by synthesis and resampling. Temporal approaches, including LipForensics [9] and BSF [12], model lip dynamics and temporal frequency, while Mover [10] reasons about facial-part consistency. While highly effective in many settings, cue-driven approaches depend on the stability of specific forgery signals under compression, resizing, and changes in the manipulation pipeline, assumptions that may not always hold in different deployment scenarios.

Depth–Based Detectors. Most closely related to our work

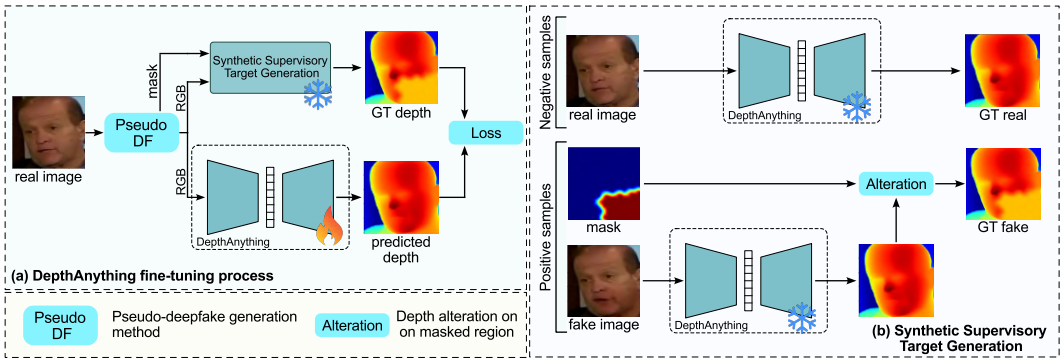

Figure 2. **Forgery-aware depth adaptation.** Starting from real/authentic and pseudo-manipulated face images, we construct depth supervision signals that preserve the original structure for real samples while altering it inside manipulated regions. Training on these targets encourages the adapted model (i.e., DepthAnything V2 in our case) to retain its response on authentic content but produce distinct, localized depth deviations for manipulated areas. This learned behavior can then be exploited for highly competitive deepfake detection.

are methods that incorporate facial depth or 3D structure into the detection process. DepthFake [24] showed that monocular depth maps can improve RGB-based detectors by serving as an additional input. Liang *et al.* [20] proposed a depth-guided triplet formulation, while Wang *et al.* [33, 34] integrated depth through attention mechanisms and RGB–depth inconsistency modeling. These works demonstrate that facial geometry provides useful forensic evidence, but they largely treat depth as an auxiliary signal supporting an RGB-based detector.

Relation to Our Work. The prior work surveyed above highlights two key themes: (i) generalizable detection requires representations that go beyond generator-specific artifacts, and (ii) complementary cues beyond RGB can improve robustness. Our approach builds on both insights, but shifts the focus from using depth as an auxiliary signal to making the depth representation itself forgery-aware. Rather than appending depth maps or using depth to support an RGB-based detector, we adapt a pretrained monocular depth foundation model [39] to encode manipulation-induced geometric inconsistencies and incorporate the adapted model into a deepfake detector. In this way, our work connects depth-based detection with representation learning for generalizable deepfake detection.

3. Methodology

In this section, we present our approach for deepfake detection, which consists of two main contributions. The first is a **Forgery-Aware Depth Adaptation Procedure** that reshapes a pretrained monocular depth foundation model into a representation extractor sensitive to manipulation-induced geometric inconsistencies. This adaptation is not intended to improve metric depth estimation itself; instead, synthetic supervisory depth targets are used to steer the model toward a depth-derived feature space that responds systematically to forged facial regions. The second contribution is **FADepth**, our novel depth-centric deepfake detector built on top of the adapted representation. In designing FADepth, we aim to preserve the generic geometric priors of the original depth model, retain sensitivity to forgery-specific structure introduced during adaptation, and benefit from complementary RGB evidence without shifting the focus away from depth. The resulting design therefore combines adapted depth features with the original depth representation and augments them with RGB cues in a unified detection framework. Details on both contributions are provided in the following sections.

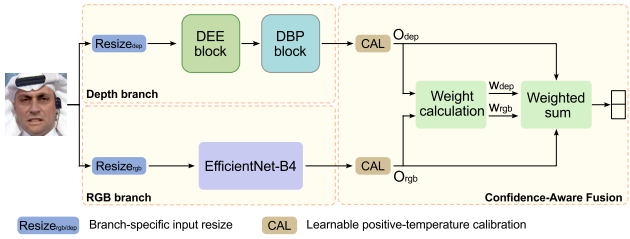

Figure 3. **Overview of FADepth.** The proposed detector follows a two-branch design that combines geometric and appearance cues. A depth branch extracts and fuses representations from the original and adapted DAv2 encoders to capture both generic and forgery-aware signals, while a parallel RGB branch provides complementary appearance cues. The two predictions are combined through confidence-aware fusion to produce the final deepfake decision.

3.1. Forgery-Aware Depth Adaptation

We begin by adapting a pretrained Depth Anything V2 (DAv2) [39] model into a forgery-aware version using synthetic supervisory depth targets derived from real and facial images with proxy manipulations. The key idea is to preserve the reference depth structure for authentic samples and selectively suppress it inside manipulated regions for proxy-manipulated samples, as illustrated in Figs. 1 and 2.

Synthetic Supervisory Targets. Given a real RGB face crop $\mathbf{x}^r \in \mathbb{R}^{3\times H\times W}$, we generate a pseudo-manipulated sample $\mathbf{x}^f \in \mathbb{R}^{3\times H\times W}$ using the manipulation approach

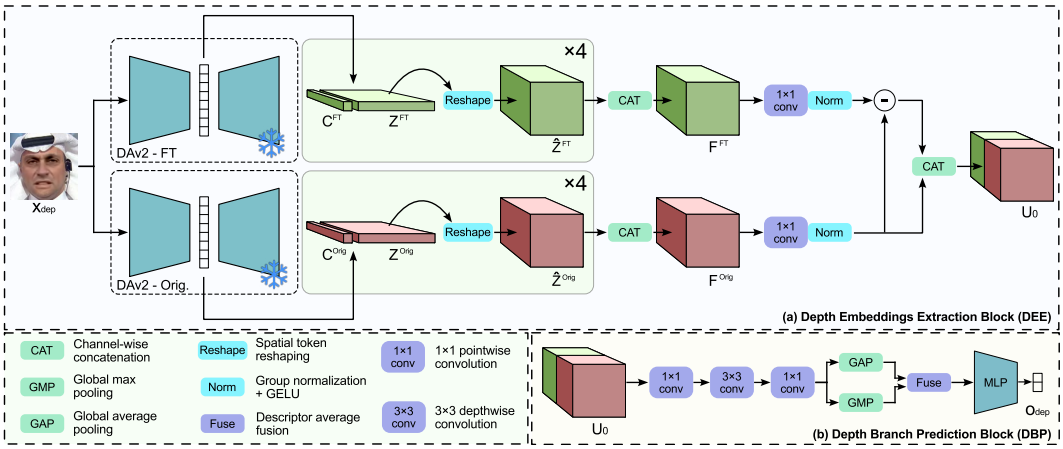

Figure 4. **Depth branch of FADepth.** The branch combines features from the original and adapted DAv2 encoders to capture complementary geometric information. In the DEE block, the representations are fused by emphasizing the adapted features and their deviation from the original encoder. The DBP block then refines and classifies the result to produce the depth-based prediction.

from [29], which in addition to the synthetic manipulation also generates a corresponding manipulation mask $\mathbf{m} \in \mathbb{R}^{H\times W}$. The normalized mask is defined as

$$\mathbf{P} = \frac{\mathbf{m}}{\max(\mathbf{m})}, \tag{1}$$

so that $\mathbf{P} \in [0,1]^{H\times W}$. A frozen pretrained DAv2 model $T(\cdot)$ is then used to obtain reference depth predictions

$$\mathbf{D}^r = T(\mathbf{x}^r), \qquad \mathbf{D}^f = T(\mathbf{x}^f), \tag{2}$$

which form the basis for the synthetic supervisory targets. For authentic samples, the target is simply the reference prediction,

$$\widetilde{\mathbf{D}}^r = \mathbf{D}^r. \tag{3}$$

For proxy-manipulated samples, the depth map is altered only inside the manipulated region. Let

$$\mu(\mathbf{D}^f) = \frac{1}{HW}\sum_{u=1}^{H}\sum_{v=1}^{W}\mathbf{D}^f_{u,v} \tag{4}$$

denote the global mean value of the reference depth map. The target for the manipulated sample is then defined as

$$\begin{aligned}\widetilde{\mathbf{D}}^f &= \mathbf{D}^f + \mathbf{P}\odot\left(\mu(\mathbf{D}^f) - \mathbf{D}^f\right) \\ &= (1-\mathbf{P})\odot\mathbf{D}^f + \mathbf{P}\odot\mu(\mathbf{D}^f),\end{aligned} \tag{5}$$

where $\odot$ denotes the Hadamard product and the scalar $\mu(\mathbf{D}^f)$ is broadcast over the spatial domain. This construction suppresses local geometric variation inside the manipulated region by contracting the masked depth values toward the image-level mean depth, while leaving the unmasked regions unchanged. As a result, the supervisory signal remains faithful to the original depth map for authentic content, but deviates from it in a controlled and spatially localized manner for synthetically manipulated regions. The process is illustrated in Figure 2 (b).

Training/Adaptation Objective. Let $S(\cdot)$ denote the trainable DAv2 model initialized from the same pretrained weights as the frozen reference model. The adapted model is trained to regress the synthetic supervisory targets, using $\widetilde{\mathbf{D}}^r$ for authentic samples and $\widetilde{\mathbf{D}}^f$ for pseudo-manipulated samples. The regression objective is a pointwise Huber loss with threshold $\delta = 0.5$,

$$\mathcal{L}_{\text{depth}} = \ell_{\text{Huber}}\Big(S(\mathbf{x}), \widetilde{\mathbf{D}}; \delta = 0.5\Big), \tag{6}$$

where $\widetilde{\mathbf{D}}$ denotes the corresponding synthetic target. Because the targets for manipulated inputs explicitly attenuate local depth structure inside the forged region, the optimization does not seek better depth estimates in a metric sense. Instead, it encourages the model to encode deviations from expected geometric patterns directly in its internal representation. The resulting adapted model therefore preserves the original depth behavior on authentic content while developing systematic, spatially localized responses to manipulation-induced geometric inconsistencies.

3.2. FADepth: Depth-Centric Deepfake Detector

The overall design of **FADepth** is illustrated in Fig. 3. Our goal is to combine two complementary sources of evidence for deepfake detection: (1) *geometry*, captured through depth, and (2) *appearance*, captured in RGB. To exploit these complementary cues, FADepth adopts a two-branch architecture. The *depth branch* combines features from the original and adapted (forgery-aware) depth encoder, while a parallel *RGB branch* provides appearance-based evidence. The outputs of the two branches are then *fused in a confidence-aware manner* to produce the final prediction. We note that our design does not use DAv2 for explicit depth prediction, instead the learned encoder features are exploited as a representation for detection.

FADepth Inputs. Given a face crop $\mathbf{x} \in \mathbb{R}^{3\times H\times W}$, we feed it to both branches with branch-specific resolutions:

$$\mathbf{x}_{\text{dep}} = \mathcal{R}_{518}(\mathbf{x}), \qquad \mathbf{x}_{\text{rgb}} = \mathcal{R}_{380}(\mathbf{x}). \tag{7}$$

where $\mathcal{R}_s(\cdot)$ denotes image resizing to $s \times s$ pixels. The depth branch uses 518×518 inputs to match the design of the DAv2 encoder, whereas the RGB branch uses 380×380 inputs to ensure an appropriate resolution for the spatial feature extractor, i.e., EfficientNet-B4 in our design.

The Depth Branch. The depth branch, shown in Fig. 4, leverages two complementary depth representations. It uses the original pretrained DAv2 encoder to preserve generic geometric priors, and the adapted (forgery-aware) encoder to capture forgery-aware responses. Combining both allows the model to retain a stable representation of facial geometry while becoming sensitive to manipulation-induced deviations. The branch is organized into a Depth Embedding Extraction (DEE) block followed by a Depth Branch Prediction (DBP) block.

- **The DEE Block:** Let $E^{\text{orig}}(\cdot)$ and $E^{\text{ft}}(\cdot)$ denote the frozen original and adapted DAv2 encoders. Rather than using the depth regression heads, we extract intermediate transformer features directly from the underlying DINOv2-ViT-L backbone encoder. For the given input $\mathbf{x}_{\text{dep}}$, each encoder produces four intermediate outputs from layers $i \in \{4, 11, 17, 23\}$, yielding patch tokens $\mathbf{Z}_i \in \mathbb{R}^{B\times N\times C}$ with $N = 37 \times 37$ and $C = 1024$ for both the original $E^{\text{orig}}(\cdot)$ and adapted $E^{\text{ft}}(\cdot)$ encoder. We discard the class tokens and reshape the patch embeddings into spatial feature maps $\widehat{\mathbf{Z}}_i \in \mathbb{R}^{B\times C\times 37\times 37}$. To retain multi-level transformer evidence, we concatenate the four stages along the channel dimension, resulting in two depth-derived representations:

$$\begin{aligned}\mathbf{F}^{\text{orig}} &= [\widehat{\mathbf{Z}}_1^{\text{orig}}; \widehat{\mathbf{Z}}_2^{\text{orig}}; \widehat{\mathbf{Z}}_3^{\text{orig}}; \widehat{\mathbf{Z}}_4^{\text{orig}}] \in \mathbb{R}^{B\times 4096\times 37\times 37},\\ \mathbf{F}^{\text{ft}} &= [\widehat{\mathbf{Z}}_1^{\text{ft}}; \widehat{\mathbf{Z}}_2^{\text{ft}}; \widehat{\mathbf{Z}}_3^{\text{ft}}; \widehat{\mathbf{Z}}_4^{\text{ft}}] \in \mathbb{R}^{B\times 4096\times 37\times 37}.\end{aligned} \tag{8}$$

The two tensors $\mathbf{F}^{\text{orig}}$ and $\mathbf{F}^{\text{ft}}$ are fused by a lightweight classification head. Each stream is first projected independently into a common lower-dimensional space through a 1×1 convolution, followed by Group Normalization, GELU activation, and spatial dropout:

$$\begin{aligned}\mathbf{A} &= \phi_{\text{orig}}(\mathbf{F}^{\text{orig}}) \in \mathbb{R}^{B\times 256\times 37\times 37},\\ \mathbf{B} &= \phi_{\text{ft}}(\mathbf{F}^{\text{ft}}) \in \mathbb{R}^{B\times 256\times 37\times 37}.\end{aligned} \tag{9}$$

To reduce sensitivity to activation magnitude and emphasize structural differences, we apply ℓ_2 normalization across channels at each spatial location. i.e.:

$$\overline{\mathbf{A}}_{b,:,u,v} = \frac{\mathbf{A}_{b,:,u,v}}{\|\mathbf{A}_{b,:,u,v}\|_2}, \quad \overline{\mathbf{B}}_{b,:,u,v} = \frac{\mathbf{B}_{b,:,u,v}}{\|\mathbf{B}_{b,:,u,v}\|_2}. \tag{10}$$

Instead of symmetrically concatenating both streams, we center the fusion on the forgery-aware representation and augment it with an explicit disagreement signal computed relative to the original encoder:

$$\Delta = |\overline{\mathbf{B}} - \overline{\mathbf{A}}| \in \mathbb{R}^{B\times 256\times 37\times 37}, \tag{11}$$

$$\mathbf{U}_0 = [\overline{\mathbf{B}}; \Delta] \in \mathbb{R}^{B\times 512\times 37\times 37}, \tag{12}$$

which produces the final representation of the DEE block.

- **The DBP Block:** In the DBP block, the fused tensor $\mathbf{U}_0$ is refined through a sequence of lightweight channel and spatial mixing blocks, consisting of pointwise and depthwise convolutions with group normalization, GELU activations and spatial dropout. This progressively reduces the representation to $\mathbf{U} \in \mathbb{R}^{B\times 128\times 37\times 37}$ while preserving spatial structure. Next, we additionally compress the information content in $\mathbf{U}$ with a combination of global-average and max pooling

$$\mathbf{z}_{\text{dep}} = \frac{1}{2}\,\text{GAP}(\mathbf{U}) + \frac{1}{2}\,\text{GMP}(\mathbf{U}) \in \mathbb{R}^{B\times 128}, \tag{13}$$

to capture both global context and salient activations. The pooled descriptor is then classified using a compact MLP to produce the depth-branch logits, i.e.:

$$\mathbf{o}_{\text{dep}} = \text{MLP}_{\text{dep}}(\mathbf{z}_{\text{dep}}) \in \mathbb{R}^{B\times 2}. \tag{14}$$

In our implementation, the MLP_{dep} classifier consists of a LayerNorm, a linear layer $128 \to 64$, a GELU activation, dropout, and a final linear layer $64 \to 2$.

The RGB Branch. In parallel to the depth branch, FADepth includes an RGB appearance branch that captures complementary visual cues. The branch operates on the same face crop, resized to $\mathbf{x}_{\text{rgb}} \in \mathbb{R}^{3\times 380\times 380}$, and is implemented with an EfficientNet-B4 backbone pretrained on pseudo-manipulated face images (following [29]) to provide task-relevant appearance priors. The face crop is processed through the standard EfficientNet-B4 pipeline, followed by global average pooling, dropout, and a classifier:

$$\mathbf{o}_{\text{rgb}} = E_{\text{rgb}}(\mathbf{x}_{\text{rgb}}) \in \mathbb{R}^{B\times 2}, \tag{15}$$

where the logits $\mathbf{o}_{\text{rgb}}$ provide appearance-based evidence complementary to the geometry-sensitive cues from the depth branch, described in the previous sections.

Confidence-Aware Fusion. The final prediction is obtained by combining the depth and RGB logits in a confidence-aware manner. Each branch is first calibrated using a learnable positive temperature,

$$\tau_{\text{rgb}} = \text{softplus}(\theta_{\text{rgb}}) + \epsilon, \quad \tau_{\text{dep}} = \text{softplus}(\theta_{\text{dep}}) + \epsilon, \tag{16}$$

yielding calibrated logits $\widetilde{\mathbf{o}}_{\text{rgb}} = \mathbf{o}_{\text{rgb}}/\tau_{\text{rgb}}$ and $\widetilde{\mathbf{o}}_{\text{dep}} = \mathbf{o}_{\text{dep}}/\tau_{\text{dep}}$. From the calibrated logits, we compute

branch probabilities $\mathbf{p}_{\text{rgb}} = \text{softmax}(\widetilde{\mathbf{o}}_{\text{rgb}})$, and $\mathbf{p}_{\text{dep}} = \text{softmax}(\widetilde{\mathbf{o}}_{\text{dep}})$, and estimate their reliability via normalized negative entropy,

$$r = 1 - \frac{\mathcal{H}(\mathbf{p})}{\log 2}, \quad \mathcal{H}(\mathbf{p}) = -\sum_{c} p_c \log p_c, \tag{17}$$

so that more confident predictions receive higher weights.

To model a global preference between modalities, we introduce learnable prior logits β_{rgb} and β_{dep}, yielding prior branch weights $[\pi_{\text{rgb}}, \pi_{\text{dep}}] = \text{softmax}([\beta_{\text{rgb}}, \beta_{\text{dep}}])$. The final sample-wise weights are then defined as

$$\widehat{w}_{\text{rgb}} = \pi_{\text{rgb}}(\rho + r_{\text{rgb}}), \quad \widehat{w}_{\text{dep}} = \pi_{\text{dep}}(\rho + r_{\text{dep}}), \tag{18}$$

followed by normalization

$$w_i = \frac{\widehat{w}_i}{\widehat{w}_{\text{rgb}} + \widehat{w}_{\text{dep}} + \epsilon}, \quad i \in \{\text{rgb}, \text{dep}\}, \tag{19}$$

where $\rho = 0.10$ prevents either branch from collapsing to zero contribution. The final fused logits are ultimately computed as

$$\mathbf{o} = w_{\text{rgb}}\widetilde{\mathbf{o}}_{\text{rgb}} + w_{\text{dep}}\widetilde{\mathbf{o}}_{\text{dep}}. \tag{20}$$

The final prediction is obtained through a softmax over $\mathbf{o}$.

Training Objective. Training is driven by the fused classification loss, complemented by auxiliary branch losses:

$$L_{\text{II}} = L_{\text{fuse}} + 0.35L_{\text{dep}} + 0.10L_{\text{rgb}}, \tag{21}$$

with L_{fuse} as the dominant term. During training, all components of the detector are optimized jointly, while the depth encoders remain frozen. The auxiliary losses stabilize optimization and encourage both branches to learn complementary yet discriminative representations. Optimization details are provided in the next section.

4. Experimental Setup

Experimental Datasets. We evaluate the proposed FADepth detector on six widely used standard deepfake video datasets: FaceForensics++ (FF++) [27], Celeb-DF v2 (CDF) [18], Deepfake Detection Challenge Preview Dataset (DFDCP) [7], DeepFake Detection Challenge (DFDC) [6], Face Forensics in the Wild (FFIW) [42], and Celeb-DF v3 (CDF++) [19]. We report results on the official test splits of all datasets. To assess generalization, we adopt a cross-dataset protocol in which FADepth is trained without ever seeing actual deepfakes, exclusively with pseudo-manipulated samples generated from the pristine training split of FF++, following [29], and without any fine-tuning on the target datasets. The model is then evaluated on the official test splits of all datasets.

Performance Scoring. Although the proposed detector operates on individual face crops, we evaluate performance at the video level. For each video, we sample 32 frames, detect and crop faces, then run the model on the resulting face crops. When multiple faces are present in a frame, we assign the frame score as the maximum score among the detected faces. Video scores are then obtained by averaging across the sampled frames. We report the Area Under the ROC Curve (AUC) for each dataset. To summarize cross-dataset performance with a single scalar value, we also compute the macro-average AUC, i.e.: $\text{mAUC} = \frac{1}{|\mathcal{D}|}\sum_{d\in\mathcal{D}} \text{AUC}(d)$, where $\mathcal{D}$ denotes the set of evaluation datasets. Since different baselines report results on different subsets of datasets, for each baseline b that reports on a subset $\mathcal{D}_b \subseteq \mathcal{D}$, we additionally compute

$$\text{mAUC}_b^{(\text{ours})} = \frac{1}{|\mathcal{D}_b|}\sum_{d\in\mathcal{D}_b} \text{AUC}^{(\text{ours})}(d), \tag{22}$$

with $\Delta_b = \text{mAUC}_b^{(\text{ours})} - \text{mAUC}(b)$, which enables fair subset-matched comparison across competing methods.

Baselines. We compare against a broad and diverse set of 24 deepfake detectors from the generalization-oriented groups discussed in Section 2, i.e., (1) data-centric and training-strategy methods [4, 15, 16, 25, 29, 36, 37, 41], (2) manipulation-agnostic / representation-oriented detectors [3, 11, 22, 28, 30, 38, 40], and (3) cue-driven / specialized-evidence methods [5, 9, 10, 12, 21, 23, 42]. In addition, we include the two most closely related depth-aware detectors, DEM-3-DFD [20] and DeI-DFD [34]. To ensure a fair comparison, we report results from the published papers and use the same test protocol to evaluate FADepth. When a method does not report results for a particular test dataset, the corresponding entry is treated as unavailable and excluded from the subset-average computation.

Implementation Details. Training is performed in two stages. In Stage I, we adapt the DAv2 ViT-L model [39] using pseudo-manipulated samples generated from the pristine FF++ training split following [29]. The resulting model serves as the adapted depth encoder. In Stage II, we train FADepth by combining the frozen original and adapted DAv2 encoders with an RGB branch based on EfficientNet-B4 pretrained on pseudo-manipulated data. The depth and RGB branches operate at 518×518 and 380×380 resolution, respectively. We first train the classification heads and fusion module, followed by a short fine-tuning phase of the full model. Model selection is based on validation performance. Full training details with hyperparameter settings are provided in the supplementary material and can also be found in the publicly released source code.

Computational Complexity. All models are implemented in PyTorch and trained/evaluated on an NVIDIA A100-SXM4-80GB GPU. At the default two-branch input configuration (518×518 for depth and 380×380 for RGB), the proposed detector comprises 690.4M parameters and requires 1.02×10^{12} MACs (2.04×10^{12} FLOPs) per image. The average inference time is 162.65 ms/image on average.

Table 1. **Comparison with the state-of-the-art.** Reported are AUC scores (%) on the six test datasets. Methods are grouped according to the categories discussed in the Related Work. Since different baselines report results on different subsets of datasets, **Avg. AUC** denotes the average over the datasets reported by each baseline, while **mAUC** denotes the average of our method recomputed on the identical subset. Δ is the corresponding subset-matched improvement. For FADepth, mAUC corresponds to the average over all six datasets.

	Method	Venue	FF	CDFv2	DFDC	DFDCP	FFIW	CDF++	Avg. AUC (↑)	mAUC (↑)	Δ
Data-centric & training	PCL+I2G [41]	ICCV′21	99.07	90.03	67.52	74.37	–	–	82.75	88.89	6.14
	CORE [25]	CVPRW′22	–	80.90	72.10	72.00	71.00	74.90	74.18	86.51	12.33
	SBI [29]	CVPR′22	**99.64**	93.18	72.42	86.15	84.83	73.40	84.94	88.70	3.76
	AltFreezing [36]	CVPR′23	98.60	89.50	–	–	–	–	94.05	96.42	2.37
	LSDA [37]	CVPR′24	–	83.00	73.60	81.50	–	72.70	77.70	85.62	7.92
	ProDet [4]	NeurIPS′24	–	92.50	–	–	–	73.60	83.05	89.89	6.84
	MSBA-CLIP [16]	arXiv′26	–	86.10	76.37	80.38	–	–	80.95	85.30	4.35
	ProDet+TSRL [15]	arXiv′26	–	90.70	70.60	84.50	–	–	81.93	85.30	3.37
Manip.-agnostic / repr.-oriented	RECCE [3]	CVPR′22	99.32	68.71	69.06	–	–	–	79.03	89.25	10.22
	UCF [38]	ICCV′23	–	82.40	80.50	–	–	76.10	79.67	84.90	5.23
	IID [11]	CVPR′23	–	83.80	**81.23**	–	–	75.60	80.21	84.90	4.69
	CFM [22]	TIFS′24	–	89.70	70.60	80.20	83.10	76.50	80.02	86.51	6.49
	FDML [40]	Neurocomp.′24	99.55	73.10	–	–	–	–	86.33	96.42	10.09
	LESB [30]	WACV′25	–	93.13	71.98	–	83.01	–	82.71	86.05	3.34
	RFFR [28]	PR′25	–	88.98	67.84	–	–	–	78.41	84.06	5.65
Cue-driven / special. evidence	SRM [23]	CVPR′21	–	84.00	69.50	72.80	79.40	79.40	77.02	86.51	9.49
	SPSL [21]	CVPR′21	–	79.90	72.40	77.00	79.40	74.40	76.62	86.51	9.89
	FFIW [42]	CVPR′21	99.30	78.30	–	74.10	–	–	83.90	93.54	9.64
	LipForensics [9]	CVPR′21	97.10	82.40	73.50	–	–	–	84.33	89.25	4.92
	Mover [10]	arXiv′23	–	87.10	–	78.90	–	–	83.00	90.50	7.50
	Style latent flows [5]	CVPR′24	–	89.00	–	–	–	–	89.00	93.19	4.19
	BSF [12]	ICCV′25	–	89.70	75.20	–	–	–	82.45	84.06	1.61
Depth-based	DEM-3-DFD [20]	NN′23	–	72.30	–	–	–	–	72.30	93.19	20.89
	DeI-DFD [34]	arXiv′24	98.33	83.35	–	–	–	–	90.84	96.42	5.58
	FADepth (Ours)	–	**99.64**	**93.19**	74.92	**87.80**	**90.04**	**86.58**	–	**88.70**	–

5. Results and Discussions

5.1. Comparison with the State-of-the-Art

In the first series of experiments we compare FADepth against a cross-section of state-of-the-art deepfake detectors from the literature. Experiments are conducted in cross-dataset settings and the results are reported in Table 1.

Quantitative Evaluation. Since different baselines report performance on different subsets, we provide both their reported averages and subset-matched comparisons using mAUC and Δ. The proposed FADepth detector achieves an overall mAUC of 88.70 across all six benchmarks and achieves the best performance on five out of the six datasets. In particular, it matches the top result on FF (99.64), slightly improves over prior work on CDFv2 (93.19), and establishes new best results on DFDCP, FFIW, and CDF++. Gains are most pronounced on the challenging out-of-domain datasets FFIW and CDF++, with improvements of $+5.21$ and $+7.18$ points. While DFDC remains the only dataset where our method is not top-ranked, it achieves the strongest overall balance across benchmarks.

Comparison Across Detector Groups. Across all baseline detector groups, the proposed method yields consistent positive improvements in subset-matched mAUC. The performance gains range from $+2.37$ to $+12.33$ over data-centric approaches, from $+3.34$ to $+10.22$ over manipulation-agnostic methods, and from $+1.61$ to $+9.89$ over cue-driven detectors. Notably, within the most closely related depth-based category, our method improves by $+20.89$ over DEM-3-DFD and $+5.58$ over DeI-DFD. These results show that the proposed depth-centric formulation is not merely competitive within the depth-aware setting, but also compares favorably with strong RGB-only, temporal, and representation-oriented detectors.

5.2. Ablation Studies

In the next series of experiments, we analyze the key components of the proposed approach and assess their individual contributions through a series of ablation studies.

Modality Configuration and Fusion Design. We first explore the impact of different modality configurations and fusion strategies, including depth-only, RGB-only, and combined models. From the results in Table 2, we observe that the adapted depth branch alone already performs well, slightly outperforming the RGB-only baseline, confirming that the proposed depth adaptation produces informative forgery cues. Adding the original depth encoder does not improve performance in isolation, but becomes beneficial when combined with RGB evidence. Combining depth and RGB consistently yields the highest results, with our dual-depth + RGB design achieving the best performance. Compared to alternative fusion strategies, the confidence-aware fusion performs best, indicating that modality-specific decisions are more effective than early feature-level integration. Overall, the results suggest that the gains stem from the combination of adapted depth features, preserved geo-

Table 2. **Ablation of modality configuration and fusion strategy.** Reported are AUC scores (%) on the six evaluation datasets and the average over all six test datasets. *Adapted depth only* uses features from the adapted DAv2 encoder only, *dual-depth only* combines the adapted and original DAv2 encoders, and the fusion variants combine depth and RGB evidence using different fusion mechanisms.

	Setting	FF	CDFv2	DFDC	DFDCP	FFIW	CDF++	Avg. AUC (↑)
Single-branch	Adapted depth only	99.22	81.63	75.50	86.16	90.37	83.26	86.02
	Dual-depth only	99.45	81.59	74.51	87.21	88.76	82.48	85.67
	RGB-only (EfficientNet-B4)	99.57	93.98	72.65	86.56	86.04	74.81	85.60
Fusion variants	Adapted depth + RGB (late fusion)	99.58	89.81	75.65	88.60	91.19	84.76	88.27
	Dual-depth + RGB (feature fusion)	99.68	92.98	75.37	87.53	90.52	83.50	88.26
	Proposed dual-depth + RGB	99.64	93.19	74.92	87.80	90.04	86.58	**88.70**

Table 3. **Alternative depth-output designs.** AUC (%). The adapted model is converted into a 1-channel depth-map generator and a separate Xception classifier is trained on the resulting maps. We compare direct adapted-depth outputs and residual depth maps against the proposed FADepth feature-space model.

Setting	FF	CDFv2	DFDC	DFDCP	FFIW	CDF++	Avg
Adapted depth	98.56	82.61	72.53	**88.40**	84.65	85.09	85.31
Residual depth	98.82	80.76	**77.39**	87.20	88.81	81.23	85.70
FADepth (ours)	**99.64**	**93.19**	74.92	87.80	**90.04**	**86.58**	**88.70**

metric priors, and complementary RGB evidence.

Alternative Depth-Output Designs. Next, we evaluate whether the adapted model can be used via its rendered depth outputs by training a separate Xception-based classifier on either the predicted depth maps or their residuals, i.e., the difference between the depth predictions of the original and forgery-aware depth-prediction model. As can be seen from Table 3, both variants underperform compared to our feature-based FADepth models, with the residual-based design performing slightly better. However, they remain more than 3 AUC points below the proposed detector, indicating that the key signal lies in the internal depth representation rather than in the rendered depth maps.

Synthetic Supervisory Target Construction. In the final ablation, we compare different strategies for constructing the synthetic depth targets used to adapt DAv2 in Figure 5. The desired behavior of these strategies is twofold: on authentic images, the adapted model should remain close to the base DAv2 prediction, while on manip- ulated images it should produce a clear and spatially localized deviation inside the manipulated region. The proposed strategy in Fig. 5(a), which replaces the manipulated region with a mean-anchored plateau while preserving the surrounding structure, best satisfies these requirements. It leaves real-image predictions largely unchanged and produces a stable, localized response on manipulated inputs.

Alternative designs either destabilize the representation or yield weaker signals. Forcing manipulated regions outside the natural depth range (b) leads to globally corrupted predictions, indicating that the model departs too far from the pretrained depth manifold. Reducing the magnitude of the range violation (c) partially alleviates this but still produces saturated and poorly localized responses. The constant-plateau strategy (d) is more stable, but yields weaker and less consistent deviations due to its lack of image adaptivity. Overall, these results show that the proposed range-preserving, image-adaptive target construction is crucial for maintaining geometric consistency while inducing reliable forgery-sensitive responses.

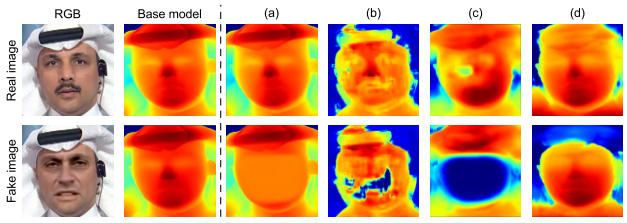

Figure 5. **Qualitative comparison of synthetic supervisory target strategies.** For one real and one manipulated FF++ sample, we compare the base DAv2 prediction with the outputs of models adapted using four target constructions. The proposed mean-anchored strategy in (a) best preserves the real prediction while inducing a clear and localized response on the manipulated region.

6. Conclusion

In this paper, we introduced a new approach to generalizable deepfake detection based on learning forgery signals in pretrained depth representations. By adapting a monocular depth foundation model with synthetic depth targets, we encourage manipulated facial regions to induce detectable responses in the learned representation. Building on this idea, we developed FADepth, a depth-centric detector that combines forgery-aware depth features with complementary RGB evidence. Experiments on six benchmark datasets demonstrated strong cross-dataset generalization and highly competitive performance compared to the state-of-the-art.

Limitations. Despite the strong results, FADepth remains computationally demanding due to the use of large depth backbones and a two-branch design. However, this increased complexity is an acceptable trade-off for the substantial gains in cross-dataset generalization and robustness.

Ethics Statement. Deepfake detection can help reduce harms related to fraud, impersonation, and misinformation, but it should not be viewed as fail-safe solutions. We encourage the use of such methods as part of broader human-centered and procedural safeguards, and stress that continued advances in generative models require ongoing evaluation of detector robustness and failure modes.

Acknowledgments

This work was supported in part by the ARIS Research Programmes P2–0250 and P2–0214, and the ARIS projects J2-50065 DeepFake DAD and General DeepFake.

References

[1] B. Batagelj, A. Kronovšek, V. Štruc, and P. Peer. Robust cross-dataset deepfake detection with multitask self-supervised learning. *ICT Express*, pages 1–5, 2025.

[2] M. Brodarič, M. Ivanovska, D. K. Jain, P. Peer, and V. Štruc. Hcsi-net: Hierarchical cross-stream interaction for generalizable deepfake detection. In *Proceedings of the International Workshop on Biometrics and Forensics (IWBF)*, pages 1–6, 2026.

[3] J. Cao, C. Ma, T. Yao, S. Chen, S. Ding, and X. Yang. End-to-end reconstruction-classification learning for face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4113–4122, 2022.

[4] J. Cheng, Z. Yan, Y. Zhang, Y. Luo, Z. Wang, and C. Li. Can we leave deepfake data behind in training deepfake detector? *NIPS*, 2024.

[5] J. Choi, T. Kim, Y. Jeong, S. Baek, and J. 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.

[6] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. C. Ferrer. The deepfake detection challenge (dfdc) dataset. *arXiv preprint:2006.07397*, 2020.

[7] B. Dolhansky, R. Howes, B. Pflaum, N. Baram, and C. C. Ferrer. The deepfake detection challenge (dfdc) preview dataset. *arXiv preprint:1910.08854*, 2019.

[8] L. Dragar, P. Rot, P. Peer, V. Štruc, and B. Batagelj. W-tdl: Window-based temporal deepfake localization. In *Proceedings of the 2nd International Workshop on Multimodal and Responsible Affective Computing (MRAC '24), Proceedings of the 32nd ACM International Conference on Multimedia (MM'24)*. ACM, 2024.

[9] A. Haliassos, K. Vougioukas, S. Petridis, and M. Pantic. Lips don't lie: A generalisable and robust approach to face forgery detection. In *IEEE/CVF CVPR*, 2021.

[10] J. Hu, X. Liao, D. Gao, S. Tsutsui, Q. Wang, Z. Qin, and M. Z. Shou. Mover: Mask and recovery based facial part consistency aware method for deepfake video detection. *arXiv preprint arXiv:2303.01740*, 2023.

[11] B. Huang, Z. Wang, J. Yang, J. Ai, Q. Zou, Q. Wang, and D. Ye. Implicit identity driven deepfake face swapping detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4490–4499, 2023.

[12] T. Kim, J. Choi, Y. Jeong, H. Noh, J. Yoo, S. Baek, and J. Choi. Beyond spatial frequency: Pixel-wise temporal frequency-based deepfake video detection. *IEEE/CVF ICCV*, 2025.

[13] N. Larue, N.-S. Vu, V. Struc, P. Peer, and V. Christophides. Seeable: Soft discrepancies and bounded contrastive learning for exposing deepfakes. In *IEEE/CVF ICCV*, 2023.

[14] N. Larue, V. Štruc, P. Peer, and N.-S. Vu. Learning the manifold of authenticity: Hybrid-curvature representation learning for generalizable deepfake detection. *IEEE Access*, pages 1–14, 2026.

[15] Z. Lei, Z. Wang, J. Cheng, B. Huang, Y. Yang, Z. Han, C. Liang, and D. Ye. Tutor-student reinforcement learning: A dynamic curriculum for robust deepfake detection. *arXiv preprint arXiv:2603.24139*, 2026.

[16] J. Li, J. Tong, and P. Wu. Detecting deepfakes with multivariate soft blending and clip-based image-text alignment. *arXiv preprint arXiv:2602.15903*, 2026.

[17] L. Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wen, and B. Guo. Face x-ray for more general face forgery detection. In *CVPR*, 2020.

[18] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *IEEE/CVF CVPR*, 2020.

[19] Y. Li, D. Zhu, X. Cui, and S. Lyu. Celeb-df++: A large-scale challenging video deepfake benchmark for generalizable forensics. *arXiv preprint arXiv:2507.18015*, 2025.

[20] B. Liang, Z. Wang, B. Huang, Q. Zou, Q. Wang, and J. Liang. Depth map guided triplet network for deepfake face detection. *Neural Networks*, 159:34–42, 2023.

[21] H. Liu, X. Li, W. Zhou, Y. Chen, Y. He, H. Xue, W. Zhang, and N. Yu. Spatial-phase shallow learning: rethinking face forgery detection in frequency domain. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 772–781, 2021.

[22] A. Luo, C. Kong, J. Huang, Y. Hu, X. Kang, and A. C. Kot. Beyond the prior forgery knowledge: Mining critical clues for general face forgery detection. *IEEE Transactions on Information Forensics and Security*, 19:1168–1182, 2024.

[23] Y. Luo, Y. Zhang, J. Yan, and W. Liu. Generalizing face forgery detection with high-frequency features. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16317–16326, 2021.

[24] L. Maiano, L. Papa, K. Vocaj, and I. Amerini. Depthfake: a depth-based strategy for detecting deepfake videos. In *International Conference on Pattern Recognition*, pages 17–31. Springer, 2022.

[25] Y. Ni, D. Meng, C. Yu, C. Quan, D. Ren, and Y. Zhao. Core: Consistent representation learning for face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12–21, 2022.

[26] G. Pei, J. Zhang, M. Hu, Z. Zhang, C. Wang, Y. Wu, G. Zhai, J. Yang, and D. Tao. Deepfake generation and detection: A benchmark and survey. *ACM Computing Surveys*, 2024.

[27] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner. Faceforensics++: Learning to detect manipulated facial images. In *IEEE/CVF ICCV*, 2019.

[28] L. Shi, J. Zhang, Z. Ji, J. Bai, and S. Shan. Real face foundation representation learning for generalized deepfake detection. *Pattern Recognition*, 161:111299, 2025.

[29] K. Shiohara and T. Yamasaki. Detecting deepfakes with self-blended images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18720–18729, 2022.

[30] E. Soltandoost, R. Plesh, S. Schuckers, P. Peer, and V. Struc. Extracting local information from global representations for interpretable deepfake detection. In *Proceedings of IEEE/CFV Winter Conference on Applications in Computer Vision - Workshops (WACV-W) 2025*, pages 1–11, Tucson, USA, 2025.

[31] R. Tolosana, C. Rathgeb, R. Vera-Rodriguez, C. Busch, L. Verdilova, S. Lyu, H. H. Nguyen, J. Yamagishi, I. Echizen, P. Rot, K. Grm, V. Štruc, A. Datcheva, Z. Akhtar, S. Romero-Tapiador, J. Fierrez, A. Morales, J. Ortega-Garcia, E. Kindt, C. Jasserand, T. Kalvet, and M. Tiits. Future trends in digital face manipulation and detection. In C. Rathgeb, R. Tolosana, R. Vera-Rodriguez, and C. Busch, editors, *Handbook of Digital Face Manipulation and Detection*, pages 463–482. 2022.

[32] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia. Deepfakes and beyond: A survey of face manipulation and fake detection. *Information fusion*, 64:131–148, 2020.

[33] H. Wang, M. Li, S. Li, Z. Qian, and X. Zhang. Exploring depth information for face manipulation detection. *arXiv preprint arXiv:2212.14230*, 2022.

[34] H. Wang, S. Li, J. He, Z. Qian, X. Zhang, and S. Fan. Exploring depth information for detecting manipulated face videos. *arXiv preprint arXiv:2411.18572*, 2024.

[35] T. Wang, X. Liao, K. P. Chow, X. Lin, and Y. Wang. Deepfake detection: A comprehensive survey from the reliability perspective. *ACM Computing Surveys*, 57(3):1–35, 2024.

[36] Z. Wang, J. Bao, W. Zhou, W. Wang, and H. Li. Altfreezing for more general video face forgery detection. In *IEEE/CVF CVPR*, 2023.

[37] Z. Yan, Y. Luo, S. Lyu, Q. Liu, and B. Wu. Transcending forgery specificity with latent space augmentation for generalizable deepfake detection. In *IEEE/CVF CVPR*, 2024.

[38] Z. Yan, Y. Zhang, Y. Fan, and B. Wu. Ucf: Uncovering common features for generalizable deepfake detection. In *ICCV*, 2023.

[39] L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, and H. Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2024.

[40] M. Yu, H. Li, J. Yang, X. Li, S. Li, and J. Zhang. Fdml: Feature disentangling and multi-view learning for face forgery detection. *Neurocomputing*, 2024.

[41] T. Zhao, X. Xu, M. Xu, H. Ding, Y. Xiong, and W. Xia. Learning self-consistency for deepfake detection. In *IEEE/CVF ICCV*, 2021.

[42] T. Zhou, W. Wang, Z. Liang, and J. Shen. Face forensics in the wild. In *IEEE/CVF CVPR*, 2021.