

GazeForensics: DeepFake Detection via Gaze-guided Spatial Inconsistency Learning

Qinlin He^a, Chunlei Peng^{a,*}, Decheng Liu^a, Nannan Wang^b, Xinbo Gao^c

^a*State Key Laboratory of Integrated Services Networks, School of Cyber Engineering, Xidian University, Xi'an 710071, Shaanxi, P. R. China*

^b*State Key Laboratory of Integrated Services Networks, School of Telecommunications Engineering, Xidian University, Xi'an 710071, Shaanxi, P. R. China*

^c*Chongqing Key Laboratory of Image Cognition, Chongqing University of Posts and Telecommunications, Chongqing 400065, P. R. China*

Abstract

DeepFake detection is pivotal in personal privacy and public safety. With the iterative advancement of DeepFake techniques, high-quality forged videos and images are becoming increasingly deceptive. Prior research has seen numerous attempts by scholars to incorporate biometric features into the field of DeepFake detection. However, traditional biometric-based approaches tend to segregate biometric features from general ones and freeze the biometric feature extractor. These approaches resulted in the exclusion of valuable general features, potentially leading to a performance decline and, consequently, a failure to fully exploit the potential of biometric information in assisting DeepFake detection. Moreover, insufficient attention has been dedicated to scrutinizing gaze authenticity within the realm of DeepFake detection in recent years. In this paper, we introduce *GazeForensics*, an innovative DeepFake detection method that utilizes gaze representation obtained from a 3D gaze estimation model to regularize the corresponding representation within our DeepFake detection model, while concurrently integrating general fea-

*Corresponding author

Email addresses: qinlin@stu.xidian.edu.cn (Qinlin He), clpeng@xidian.edu.cn (Chunlei Peng), dchliu@xidian.edu.cn (Decheng Liu), nnwang@xidian.edu.cn (Nannan Wang), gaoxb@cqupt.edu.cn (Xinbo Gao)

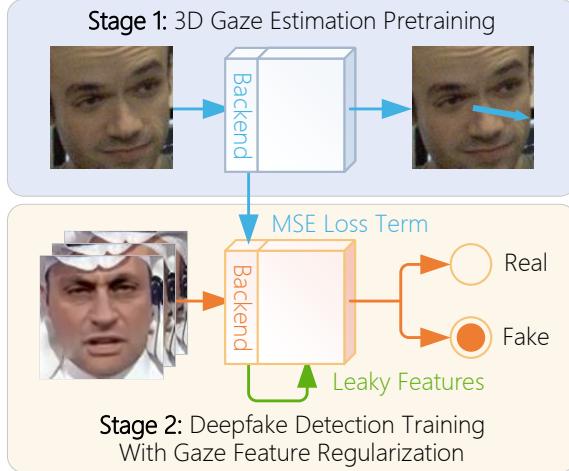


Figure 1: By regulating the corresponding representation vector output by DeepFake detection backend with gaze feature vector while leaving a certain amount of features unconstrained, our DeepFake detection model can achieve a significant improvement in accuracy and robustness.

tures to further enhance the performance of our model. Experiment results reveal that our proposed *GazeForensics* outperforms the current state-of-the-art methods.

Keywords: DeepFake Detection, Attention Machanism, Gaze Estimation

1. Introduction

In recent years, the evolution of facial manipulation technology has given rise to a formidable adversary. These techniques simplified the process of creating extraordinarily convincing face forgeries, amplifying the inherent challenges faced by traditional forgery detection methods. Deepfakes, along with feature editing, lip syncing, and so forth, with their capacity to seamlessly blend fabricated elements with genuine footage, have become potent tools for deception and manipulation, threatening both individuals and the whole society. Whether they are used to impersonate individuals, disseminate false narratives, or manipulate public perception, DeepFake technology poses a multifaceted threat that demands tailored solutions. In recognition of

this urgent need, the development of precise and robust DeepFake detection algorithms is paramount.

Previously proposed DeepFake detection methods can be categorized into two groups based on whether they incorporate biometric features or not. Approaches that do not utilize biometric features [1–6] are designed to mine general cues that are helpful to DeepFake detection. These methods usually demonstrate superior robustness with respect to input data, as most of them don't require additional biometric feature extractors or other preprocessing steps. However, their explainability is slightly compromised compared to models that rely on biometric features since the exact evidence found by these models remains uncertain. In contrast, existing methods that employ biometric features [7–11] typically utilize frozen pre-trained modules to extract desired biometric features as input for subsequent model components. These methods involve the selection of specific information relevant to certain biometric features, leading to the exclusion of general cues unrelated to the chosen biometric features.

To combine the aforementioned advantages of both biometric-based methods and non-biometric-based methods, we present our innovative gaze-based DeepFake detection method dubbed *GazeForensics*, which utilizes gaze features to provide guidance and regularization for our DeepFake detection model. Figure 1 depicts the basic concept of our proposed model. Observing that existing biometric-based DeepFake detection methods may be limited by incomplete data representations due to static backends and exclusion of general cues, we improved the training scheme and benefited from preserving general features besides biometric ones, balancing the biometric features and general features in a quantifiable way. Our approach is evaluated on the following datasets: FaceForensics++[12], Celeb-DF[13], and WildDeepfake[14]. The experimental results on both FaceForensics++ and WildDeepfake datasets demonstrate our superiority over current state-of-the-art (SOTA) approaches.

The contributions of our proposed approach can be summarized as fol-

lows:

1. We integrated DeepFake detection with 3D gaze estimation, filling this gap in the field of biometric-based DeepFake detection. This integration endows our model with the ability to discern forgery videos by distinguishing spatial inconsistencies within eye regions from a gaze perspective between given frames, improving the explainability, accuracy, and robustness of our model to some extent.
2. We proposed an innovative biometric feature integration strategy by introducing Mean Square Error (MSE) and leaky features fusion to regularize our DeepFake detection model. This feature fusion strategy not only provides a quantifiable method to balance general features and biometric ones, but also enhances the robustness and flexibility in handling challenging datasets.
3. Experimental results demonstrate that our proposed approach outperforms current SOTA models in both FaceForensics++ and WildDeepfake datasets and exhibits better explainability.

The structure of this paper is as follows: Section 1 provides a brief introduction to the background and our proposed method. Section 2 presents frameworks and works associated with our proposed approach. In Section 3, we delve into the details of *GazeForensics* framework. Section 4 showcases experimental results and their analysis. Finally, Section 5 concludes and summarizes our work.

2. Related Work

In this section, we will briefly overview the classical and recent methods for gaze estimation or DeepFake detection relevant to our work.

2.1. Gaze Estimation

Gaze estimation is an important research area that has applications in various fields, including psychology, medicine [15–17] , and human-computer interaction [18–21] . The ability to accurately estimate where a person is looking can provide valuable insights into their cognitive processes and behavior. Unlike gaze following [22–24] and 2D gaze estimation [18, 19] the task of 3D gaze estimation does not involve identifying the object or location that a person is looking at. Instead, it aims to derive the direction of a person’s gaze from images or image sequences. Zhang *et al.* have proposed and improved methods for 3D gaze estimation based on monocular images [25, 26] , offering insights into combining local eye features with overall head pose features for subsequent research. In a similar vein, Cheng *et al.* observed the phenomenon of "two-eye asymmetry" and introduced the ARE-Net [27] to fully exploit the role of binocular information in gaze estimation. However, the need to detect eye positions and employ additional modules to generate latent vectors describing head pose hindered the progress of research and application on gaze estimation. Consequently, approaches utilizing the full-face region as direct input gained attention, leading to the emergence of numerous full-face gaze estimation methods. Inspired by Kafka *et al.* ’s work [18] , Zhang *et al.* proposed a full-face gaze estimation method based on spatial weights mechanism [28] . Kellnhofer *et al.* made significant contributions by optimizing bidirectional LSTM capsules [29] using a pinball regression loss, facilitating 3D gaze estimation for continuous image sequences [30] . Abdelrahman *et al.* applied a linear combination of regression and classification losses separately for each angle, and their framework exhibited exceptional accuracy in single-image gaze estimation [31] .

2.2. DeepFake Detection

The increasing threat posed by advancements in facial manipulation technology [32–37] , alongside the growing utilization of Generative Adversarial Networks (GANs) [38–42] in the realm of DeepFake, has garnered mounting

attention from both the populace and researchers. Existing methods for detecting DeepFakes can broadly be categorized into two categories based on whether they rely on biometric features.

A wide spectrum of techniques has been employed in the realm of DeepFake detection methods that do not rely on biometric features. The work of Güera *et al.* [1] and the method proposed by Sabir *et al.* [2] both utilized a Recurrent Neural Network (RNN) to process features extracted by Convolutional Neural Networks (CNNs) in each frame, aiming to identify temporal inconsistencies in forged videos. To mine both spatial and temporal inconsistency within DeepFake videos, Gu *et al.* proposed Spatial-Temporal Inconsistency Learning (STIL). Approaching DeepFake detection from a contrastive learning perspective also showed great potential. Sun *et al.* employed two distinct modules to identify the associations and disparities between frames and, based on this, introduced the Dual Contrastive Learning (DCL) architecture [3]. Gu *et al.* took a different approach by delving into local and global contrast separately, presenting the Hierarchical Contrastive Inconsistency Learning (HCIL) framework [4]. Concerning content generated through GANs, prior research has analyzed image frequency domains using Deep Neural Networks (DNNs), leading to the development of numerous DeepFake detection methods based on image spectrum [5, 43–47]. In addition to the aforementioned studies and methods derived from hand-crafted features in the early stages [48, 49], several researchers have sought to enhance model performance by applying increasingly complex DNNs to the task of DeepFake detection [6, 50–53].

With regard to the DeepFake detection methods that incorporate biometric features, they are explicitly designed for the identification and validation of certain biometric characteristics, such as blinking [7, 54], heartbeat [10], mouth movements [9], or eye movements [8, 11, 55]. These methods employ well-designed frameworks to discern counterfeit content based on an in-depth understanding of these biometric features. Both the studies conducted by Li *et al.* [7] and Jung *et al.* [54] employed the feature of eye blink-

ing as a clue for detecting DeepFake videos. In the former scenario, a CNN was employed to extract distinctive features from eye region images. Subsequently, Long Short-Term Memory (LSTM) cells [56] were used to model temporal dependencies within the feature sequences. In contrast, the latter study entailed a statistical analysis of the blink intervals within the samples, which was conducted in comparison to the data available in their reference database. DeepRhythm [10] employed remote visual photoplethysmography to monitor the periodic changes of facial skin color caused by heartbeat, which is considered a distinctive feature that cannot be replicated by frame-by-frame generated DeepFake videos in their assumption. LipForensics [9] distinguished DeepFake contents by freezing and fine-tuning modules after pre-training them on the visual speech recognition task, focusing on mining high-level semantic irregularities in mouth movements. Wang *et al.* concatenated blink sequences with gaze vector sequences [55], emphasizing both binocular blinking and the consistency of binocular movements. Li *et al.* defined four features related to eye movements and fed the extracted features into a support vector machine to identify DeepFake videos [8]. Demir *et al.* conducted an in-depth analysis of differences between DeepFake and genuine videos regarding eye appearance, motion patterns, binocular consistency, and other features. They extracted these features and used them as inputs for subsequent binary classification DNN model [11], offering valuable insights for future researchers on distinguishing deepfake videos based on eye-related characteristics.

Many methods employing biometric features heavily rely on frozen pre-trained backends for feature extraction [8, 9, 55] , resulting in a notable loss of general information. Consequently, there is a demand for a DeepFake detection method that preserves a wide range of general information while capturing essential detailed biometric features.

3. Proposed Approach

In this section, we present *GazeForensics*, a DeepFake detection method that integrates gaze-related features with general features to enhance the model’s accuracy and robustness. Our approach primarily focuses on discerning spatial inconsistencies within eye regions between arbitrary different frames. This is achieved by incorporating regularization and expansion techniques into the representation vectors extracted by our model’s backend, allowing for the quantification of gaze information’s significance in the model’s decision-making process.

3.1. Overview

In *GazeForensics*, we aim to develop an accurate and robust DeepFake detection model. Our approach centers on the meticulous analysis of spatial inconsistency inherent in gaze-related features across disparate frames. Moreover, we harness supplementary general features to augment the model accuracy even further. We hypothesize that DeepFake videos, particularly those generated using frame-by-frame techniques, exhibit discernible discrepancies in preserving biometric attributes within the ocular regions [57] such as iris color, reflection properties, and eye shape. These artifacts are deemed retrievable in gaze feature vector output by a 3D gaze estimation backend, offering valuable insights into the detection of potential manipulations.

In contrast to prior biometric-based DeepFake detection methods that are dedicated to preserving the model’s reliance on specific biometric features by freezing pre-trained backend modules [8, 9, 55], we encourage our DeepFake detection backend to learn actively to avoid reduced representational capacity stemming from potential covariate shift [58]. To be specific, we introduced an MSE loss term to regularize the DeepFake detection backend with the gaze-related features extracted by the frozen CNN backend from a pre-trained 3D gaze estimation model. This additional MSE loss term serves as a guide, enabling the Deepfake detection backend to incorporate both DeepFake-detection-related features and detailed gaze-related features within

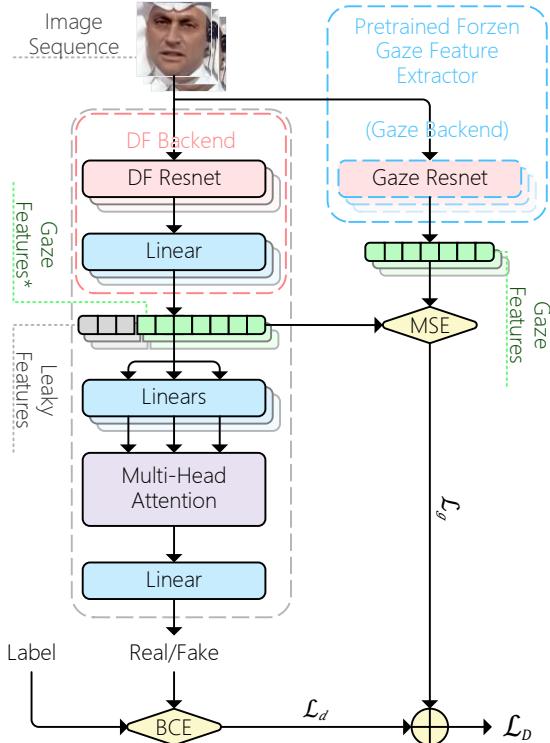


Figure 2: The overview of *GazeForensics*. Our framework utilizes a frozen pre-trained gaze feature extractor to extract gaze-related features, which are then employed to regularize the DeepFake detection backend through an MSE loss term. Simultaneously, a certain amount of features are allowed to bypass this regularization process, offering both general and supplementary features. In this context, the symbols \oplus , BCE, MSE, and Gaze Features* represent scalar addition, binary cross-entropy, mean squared error, and the gaze features estimated by the DeepFake detection backend, respectively.

an integrated representation vector. This integration is achieved as the representation extracted by the MSE-constrained Deepfake detection backend is influenced by both the MSE gaze constraint and the Binary Cross-Entropy (BCE) DeepFake detection loss. Therefore, this representation, denoted as \mathbf{r}_d , is an integration of the gaze representation (\mathbf{r}_g) and a latent DeepFake-detection-related cue representation (\mathbf{r}_{ld}):

$$\mathbf{r}_d = \mathbf{r}_g + \mathbf{r}_{ld}$$

To further enhance the incorporation of supplementary general features in assisting Deepfake detection alongside gaze-related features, we expanded the dimension of the layer that is constrained by the MSE loss term. This extension vector is allowed to escape the MSE constraint, granting it the freedom to learn any feature. Our expectation is that these additional dimensions should serve as a representation of supplementary features that \mathbf{r}_{ld} cannot accommodate.

As illustrated in Figure 2, our proposed DeepFake detection framework comprises a CNN backend for DeepFake detection, a frozen CNN backend pre-trained on 3D gaze estimation, a multi-head attention module [59], and several linear layers. To be specific, we employed Resnet-18 [50] as both our DeepFake detection backend and gaze estimation backend. Rather than using RNNs, we chose the attention mechanism, which enables simultaneous comparison over all frames. This choice is grounded in the advantages that the attention mechanism offers when conducting spatial feature comparisons in DeepFake detection. It's worth noticing that the gaze estimation backend is only used in the training phase, which is designed to transfer its knowledge of extracting detailed gaze-related features to the DeepFake detection backend by the MSE constraint.

3.2. Formulations

Let the dataset used for DeepFake detection be denoted as D_d , and it is defined as $\{(x_d^i, y_d^i)\}_{i=1}^{N_d}$. Here, N_d represents the total number of video

clips in the dataset, where each element x_d^i refers to an individual video clip, and y_d^i indicates whether the corresponding video clip is genuine or fake. Correspondingly, let the gaze estimation dataset, Gaze360 dataset [30] to be specific, be denoted as D_g , and it is defined as $\{(x_g^i, y_g^i)\}_{i=1}^{N_g}$. In this context, N_g represents the overall size of the gaze estimation dataset, where each element x_g^i corresponds to a full face image, and y_g^i comprises the yaw and pitch angles representing the direction of gaze.

We utilize the loss function denoted as CLS , which was formulated by Ahmed *et al.* [31], to optimize our 3D gaze estimation model. The expression for this loss function is defined as follows:

$$\mathcal{L}_G = \frac{1}{N_g} \sum_{i=1}^{N_g} CLS(y_g^i, f_g(x_g^i, \theta_{gb}, \theta_{go}))$$

where the 3D gaze estimation model f_g is parameterized by two sets of parameters: the backend parameters θ_{gb} and other parameters θ_{go} . This model can be decomposed into two constituent parts: f_{gb} , the backend modules of the gaze estimation model, and f_{go} , the subsequent portion of the network. Their relationship can be expressed as $f_g = f_{go} \circ f_{gb}$. Similarly, the DeepFake detection model f_d can be represented as $f_d = f_{do} \circ f_{db}$, where f_{db} and f_{do} represent their corresponding counterparts.

To account for additional features that are not captured by the MSE gaze constraint during the training process of DeepFake detection, we define the representation extracted by the frozen pre-trained 3D gaze estimation backend as $\mathbf{r}_g^i \in \mathbb{R}^\eta$. This representation is equivalent to the expression $f_{gb}(x_d^i, \theta_{gb})$. Likewise, the representation obtained from the DeepFake detection backend is denoted as $\mathbf{r}_d^i \in \mathbb{R}^{\eta+\lambda}$ and is defined by $f_{db}(x_d^i, \theta_{db})$. To clarify, we consider λ as a hyperparameter that governs the number of unconstrained dimensions, allowing the quantification of the additional features.

Since we aim to integrate DeepFake cues into the gaze feature vector, we establish the total loss function \mathcal{L}_D for the DeepFake detection training stage. This loss function comprises two distinct components, denoted as \mathcal{L}_d

and \mathcal{L}_g respectively:

1. **DeepFake Detection Component (\mathcal{L}_d)**: This component quantifies the disparity between the model's predictions and the actual labels. \mathcal{L}_d encapsulates the primary objective of identifying DeepFake content. The representation for \mathcal{L}_d is shown as follows:

$$\mathcal{L}_d = \frac{1}{N_d} \sum_{i=1}^{N_d} BCE(y_d^i, f_d(x_d^i))$$

where BCE is defined as:

$$BCE(y, \hat{y}) = -[y \cdot \log(\hat{y}) + (1 - y) \cdot \log(1 - \hat{y})]$$

2. **Gaze Constraint Component (\mathcal{L}_g)**: The gaze constraint component employs the MSE loss and serves as an auxiliary loss term. It leverages the representation of gaze features to guide and constrain the feature extraction process within DeepFake detection. This enhances the model's sensitivity to subtle ocular inconsistencies indicative of DeepFakes. The mathematical expression for \mathcal{L}_g is as follows:

$$\mathcal{L}_g = \frac{1}{N_d} \sum_{i=1}^{N_d} MSE(f_{gb}(x_d^i), \mathbf{r}|_{d_1:n}^i)$$

where MSE is defined as:

$$MSE(y, \hat{y}) = \frac{1}{N} \sum_0^N (y - \hat{y})^2$$

In the formulation of the overall loss function \mathcal{L}_D , we introduce a scaling factor denoted as μ . This factor serves to modulate the relative influence of the gaze-constraint loss term \mathcal{L}_g . By adjusting μ , we can precisely control the model's sensitivity to gaze-related features and the impact of gaze constraint

on the overall performance of DeepFake detection. The overall loss \mathcal{L}_D is defined as:

$$\mathcal{L}_D = \mathcal{L}_d + \mu \cdot \mathcal{L}_g$$

4. Experiments and Results

In this section, we present a comprehensive evaluation of *GazeForensics* on publicly available DeepFake detection datasets. We initiate our evaluation by scrutinizing the impact of two pivotal parameters on model performance, which are μ and λ . Analytical experiments are conducted with the WildDeepfake dataset, allowing us to gain insights into the impact on our model when changing the μ and λ . Subsequently, we proceed to assess the effectiveness of our proposed approach on several datasets, including FaceForensics++, CelebDF, and WildDeepfake. We also compared our method against a spectrum of classical and recent DeepFake detection methods to provide a comprehensive perspective on its performance in diverse contexts. Finally, to substantiate the efficacy of our proposed approach, we conduct a series of ablation experiments, systematically investigating the contributions of various components.

4.1. Datasets

For our experimental evaluations, we employed three publicly accessible DeepFake detection datasets, each with distinct characteristics:

1. **FaceForensics++** [12]: The FaceForensics++ (FF++) dataset serves as an established benchmark for evaluating various DeepFake detection methods. It encompasses a wide range of forgery techniques, including Deepfakes[32] (DF), Face2Face[34] (F2F), FaceSwap[33] (FS), NeuralTextures[36] (NT), and so on. However, in our experiment, we followed the prevalent practice used by researchers working with the FF++ dataset by exclusively using DeepFake videos generated from the aforementioned four methods, alongside authentic videos. The dataset

itself comprises 1000 real videos and 1000 manipulated videos for each manipulation type.

2. **Celeb-DF** [13]: The Celeb-DF dataset offers a collection of high-quality videos. It selects celebrity interview videos from YouTube and employs an improved DeepFake synthesis algorithm for face manipulation. This dataset includes 590 authentic videos alongside 5639 DeepFake videos. Additionally, there are 300 authentic videos collected from YouTube that don't have corresponding DeepFake counterparts.
3. **WildDeepfake** [14]: The WildDeepfake (Wild-DF) dataset presents a significant challenge with its diverse, internet-sourced content. It includes video samples of various qualities, undisclosed forgery techniques, and some videos with partial manipulation, making the ground truth labels less conclusive. This dataset consists of 3805 authentic video sequences and 3509 manipulated video sequences. Notably, the sequences in this dataset might represent different segments clipped from the same source video.

For 3D gaze estimation pre-training, we utilized the Gaze360[30] dataset. This dataset is distinguished by a broad spectrum of gaze and head poses, a variety of indoor and outdoor capture environments, and a diverse range of characters. These attributes make it well-suited for our pre-training purposes. The angular annotations of yaw and pitch indicate which direction the subject is looking in 3D space with respect to the camera origin. These data are measured using the standard AprilTag-based procedure [60]. The dataset is intended to be used for developing and evaluating gaze-tracking models that can estimate 3D gaze direction accurately in unconstrained images.

4.2. Experimental Settings

We implemented *GazeForensics* using the PyTorch framework, conducting all experiments on a single RTX 4080 GPU.

Pre-training: To bolster the robustness of the gaze estimation backend, we introduced a preliminary step involving 3D gaze estimation before commencing the DeepFake detection training. For our gaze estimation model, we adopted L2CS-Net[31]. During the training of this model, we applied additional data augmentation techniques that were not originally included in the reference paper. These augmentations encompassed resolution randomization, color jitter, as well as random cropping and rotation. Resolution randomization deliberately reduced the training data resolution to random values. These augmentations were vital not only for addressing video quality variations in some DeepFake detection datasets but also for mitigating potential covariate shifts when transitioning to the DeepFake detection dataset. Following the refinement of our gaze estimation model to handle diverse image contents effectively, we utilized its backend in the subsequent DeepFake detection training phase.

Preprocess: We employed RetinaFace[61] for facial alignment in the datasets that require image cropping. This cropping operation aims to achieve a uniform aspect ratio for all images and eliminate extraneous background information. Subsequently, we divided each original video into sequences of 14 frames, resizing them to a standardized resolution of 224*224 pixels before inputting them into our models. Notably, in the DeepFake detection training stage, the only applied data argumentation is a random horizontal flip since we mainly focus on model design. Additionally, our dataset partitioning followed the recommended guidelines and specifications provided in the dataset papers and associated documentation.

Training: The model underwent 50 epochs of training with the AdamW optimizer employing a weight decay coefficient of 0.01. A maximum learning rate of 7e-5 was selected, and the OneCycleLR scheduler was used to introduce warm-up phases and control the learning rate during training. Throughout the training, preprocessed frames were jointly processed by both the *GazeForensics* model and the frozen backend of L2CS-Net. To shift the model’s focus from gaze-related features to DeepFake detection in later

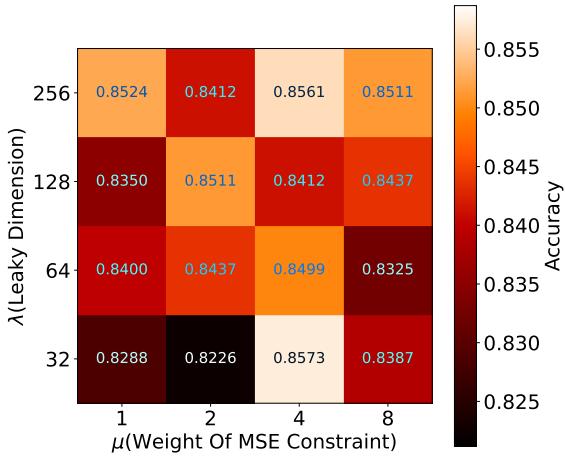


Figure 3: Grid search result on Wild-DF dataset, accuracy on the test set are reported.

stages, the hyperparameter μ was gradually reduced. After each epoch, μ was multiplied by a factor of 0.912, eventually reaching 0.01 times its initial value after 50 epochs. This dynamic reduction of μ ensured that the DeepFake detection backend prioritized the DeepFake detection loss \mathcal{L}_d in the overall loss function \mathcal{L}_D . This parameter schedule facilitated a balanced transition from extracting gaze-related representations to enhancing DeepFake detection capabilities within the DeepFake detection backend. Notably, this scheduling did not lead to rebounds in the value of \mathcal{L}_g during our experiments, signifying a refined balance and adaptation as previously mentioned.

Parameters: We selected crucial hyperparameters, μ and λ , with grid search. To expedite the parameter search process within time and computational constraints, we predefined a range of potential values for each parameter. For μ , we considered a range of four potential values: 1.0, 2.0, 4.0, and 8.0. Similarly, for λ , we confined our search to four predefined values: 32, 64, 128, and 256. To ensure result precision and highlight the varying impacts of different parameter choices, we conducted grid search experiments on the challenging WildDeepfake dataset. The experimental configuration aligns with the settings mentioned earlier in this subsection. Figure 3 illustrates the outcomes, with different μ and λ combinations showing significant

Method	FaceForensics++ c23 (HQ)				
	DF	F2F	FS	NT	Avg.
Xception [51]	0.9893	0.9893	0.9964	0.9500	0.9813
C3D [63]	0.9286	0.8857	0.9179	0.8964	0.9072
I3D [62]	0.9286	0.9286	0.9643	0.9036	0.9313
LSTM [56]	0.9964	0.9929	0.9821	0.9393	0.9777
TEI [64]	0.9786	0.9714	0.9750	0.9429	0.9670
DeepRhythm [10]	0.9870	0.9890	0.9780	-	-
S-MIL [65]	0.9857	0.9929	0.9929	0.9571	0.9822
S-MIL-T [65]	0.9964	0.9964	1.0	0.9429	0.9839
STIL [66]	0.9964	0.9928	1.0	0.9536	0.9857
ISTVT [67]	0.9964	0.9964	1.0	0.9676	0.9901
GazeForensics(Ours)	0.9893	1.0	1.0	0.9964	0.9964

Table 1: Comparison on high-quality FaceForensics++ dataset. Accuracy is reported.

variability in accuracy. The observed accuracy ranged from a minimum of 0.8226 to a maximum of 0.8573. Notably, the parameter pair $\mu = 4.0$ and $\lambda = 32$ outperformed most other combinations in terms of test set accuracy. Therefore, these values were adopted as the default configuration for the subsequent experiments.

4.3. Baseline Comparison

In this subsection, we will compare our proposed method with recent or classical DeepFake detection approaches to demonstrate the superiority of our approach, including Xception [51], LSTM [56], I3D [62], C3D [63], TEI [64], ADDNet-3D [14], F3-Net [5], S-MIL [65], STIL [66], DeepRhythm [10], ISTVT [67].

Comparasion on FF++ dataset. We conducted comprehensive experiments on four sub-datasets of the FF++ dataset, encompassing two different video quality levels, Low Quality (LQ, c23) and High Quality (HQ, c40). Table 1 and Table 2 provide an overview of the accuracy of both previous methods and our proposed *GazeForensics* on these datasets. In general,

Method	FaceForensics++ c40 (LQ)				
	DF	F2F	FS	NT	Avg.
Xception [51]	0.9678	0.9107	0.9464	0.8714	0.9241
C3D [63]	0.8929	0.8286	0.8786	0.8714	0.8679
I3D [62]	0.9107	0.8643	0.9143	0.7857	0.8688
LSTM [56]	0.9643	0.8821	0.9429	0.8821	0.9179
TEI [64]	0.9500	0.9107	0.9464	0.9036	0.9277
F3-Net [5]	0.9862	0.9584	0.9723	0.8601	0.9443
S-MIL [65]	0.9679	0.9143	0.9464	0.8857	0.9286
S-MIL-T [65]	0.9714	0.9107	0.9607	0.8679	0.9277
STIL [66]	0.9821	0.9214	0.9714	0.9178	0.9482
ISTVT [67]	0.9893	0.9607	0.9750	0.9214	0.9616
GazeForensics(Ours)	0.9786	0.9964	0.9964	1.0	0.9929

Table 2: Comparison on low-quality FaceForensics++ dataset. Accuracy is reported.

GazeForensics achieves SOTA results for three types of forgery methods: Face2Face [34], FaceSwap [33], and NeuralTexture [36] in both LQ and HQ settings. Notably, our method excels in detecting manipulations created by NeuralTexture, demonstrating a remarkable 7.86% increase in accuracy on the low-quality NerualTexture subset compared to previous SOTA methods. From the perspective of video quality, unlike previous methods that experienced significant accuracy degradation when transitioning from HQ to LQ, our approach maintains consistent accuracy levels across the F2F, FS, and NT manipulation types. This consistency showcases the superior adaptability of our method and underscores the robustness of our proposed framework concerning video quality. This may be attributed to the incorporation of random-resolution data augmentation during the pre-training of the gaze estimation backend and the integration of gaze-related information in the feature extraction. However, when compared to previous methods that consider Deepfakes as the easiest manipulation type to detect, *GazeForensics* exhibits a slightly inferior performance, with a decrease of 1.07% in detecting

Method	Celeb-DF	Wild-DF
Xception [51]	0.9944	0.8325
I3D [62]	0.9923	0.6269
ADDNet-3D [14]	-	0.6550
LSTM [56]	0.9573	-
F3-Net [5]	0.9595	0.8066
S-MIL [65]	0.9923	-
S-MIL-T [65]	0.9884	-
STIL [66]	0.9961	0.8462
ISTVT [67]	0.9981	-
GazeForensics(Ours)	0.9942	0.8573

Table 3: Comparison on Celeb-DF and WildDeepfake datasets. Accuracy is reported.

these manipulations.

Comparasion on other datasets. We have also conducted experiments on the Celeb-DF and WildDeepfake datasets to demonstrate the superiority of *GazeForensics*. As shown in Table 3, our proposed method outperforms previous SOTA methods on the Wild-DF dataset. Given that the Wild-DF dataset is notably challenging and complex, the superior accuracy of our approach demonstrates the effectiveness of our unique combination of gaze constraint and general feature fusion in enhancing the model’s generalizability and performance in complex scenarios. For the Celeb-DF dataset, which contains negative samples generated using advanced DeepFake forgery techniques, *GazeForensics* leveraged the dataset’s high visual quality and sophisticated forgery methods. This allowed for a detailed examination of spatial inconsistencies within the eye regions across image sequences. As a result, **GazeForensics** significantly narrowed the performance gap with previous methods to just 0.39% on the Celeb-DF dataset. This gap is notably smaller, especially when compared to the DeepFakes presented in the FF++ dataset.

Configuration	Accuracy
frozen backend, $\lambda = 0$	0.6960
frozen backend, $\lambda = 32$	0.8362
$\mu = 0, \lambda = 32$	0.8362
$\mu = 4, \lambda = 0$	0.8511
$\mu = 4, \lambda = 32$	0.8573

Table 4: Comparison of different key component configurations. Accuracy is reported.

4.4. Ablation Study

In this subsection, we conduct an investigation into the effectiveness of the MSE gaze constraint and the utilization of leaky features through a series of meticulously designed experiments.

Occlusion Sensitivity Visualization: In our endeavor to assess the impact of incorporating gaze constraint on the decision-making process of the model, we conducted an experiment to visualize and compare the occlusion sensitivity of both *GazeForensics* and the Xception [51]. A randomly selected subset of test examples from the FF++ low-quality dataset underwent occlusion sensitivity measurement. Specifically, we introduced a grey patch to evaluate the models’ responses under various occluded positions. The occlusion sensitivity visualization results are presented in Figure 4. Upon visual inspection, it becomes evident that the occlusion sensitivity results observed in the Xception model exhibit no discernible pattern. In contrast, *GazeForensics* places a strong emphasis on scrutinizing the eye regions to identify manipulation cues. Furthermore, it also demonstrates capability in utilizing more generalized cues in some samples, thereby expanding its versatility. These visualizations offer compelling evidence that gaze constraint significantly reshapes the foundational decision-making process in the DeepFake detection model, setting it apart from traditional methods. Consequently, this augmentation enhances the model’s explainability.

Effectiveness of Key Components: To evaluate the efficacy of the MSE gaze constraint and leaky feature fusion, we conducted experiments



Figure 4: Occlusion sensitivity samples measured with Xception and *GazeForensics* on FF++ dataset

using the Wild-DF dataset with several distinct configurations. The accuracy for different configurations is presented in Table 4. refers to the direct copying of gaze-related feature representations extracted by the pre-trained gaze backend ($f_{gb}(x_d^i)$) to the corresponding elements within the representation extracted by the DeepFake detection backend, denoted as $\mathbf{r}|_{d_1:n}^i$. This table reveals that our proposed MSE gaze constraint and leaky feature fusion significantly enhance the model’s accuracy. Notably, enabling leaky feature fusion with the frozen backend results in a significant 14.02% increase in accuracy, which underscores the limitations and deficiencies of exclusively relying on features extracted by a frozen backend trained on different datasets. Additionally, the 1.49% improvement observed when enabling the MSE gaze constraint further accentuates the model’s enhancement through the integration of gaze-related features.

5. Conclusion

In this paper, we introduced the *GazeForensics* framework, a novel approach that amalgamates gaze-related features with DeepFake detection, directing the model’s attention to eye regions in the identification of face forgeries. Furthermore, we have presented an innovative biometric feature integration approach, incorporating MSE constraint and leaky feature fusion into the representation extracted by the CNN backend, resulting in a significant improvement in model accuracy and robustness. Our approach has achieved SOTA performance in the FaceForensic++ and WildDeepfake datasets, while concurrently exhibiting commendable accuracy in the Celeb-DF dataset. The regularization method we have proposed has broader applicability, as it can theoretically be adapted to most biometric-based DeepFake detection methods. We consider our work to be a significant advancement in the ongoing effort to combat DeepFake manipulation using biometric features.

References

- [1] D. Güera, E. J. Delp, Deepfake video detection using recurrent neural networks, in: 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 2018, pp. 1–6. doi:10.1109/AVSS.2018.8639163.
- [2] E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, P. Natarajan, Recurrent convolutional strategies for face manipulation detection in videos, *Interfaces (GUI)* 3 (1) 80–87.
- [3] K. Sun, T. Yao, S. Chen, S. Ding, J. Li, R. Ji, Dual contrastive learning for general face forgery detection, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36, 2022, pp. 2316–2324.
- [4] Z. Gu, T. Yao, Y. Chen, S. Ding, L. Ma, Hierarchical contrastive inconsistency learning for deepfake video detection, in: S. Avidan, G. Bros-

- tow, M. Cissé, G. M. Farinella, T. Hassner (Eds.), Computer Vision – ECCV 2022, Springer Nature Switzerland, Cham, 2022, pp. 596–613.
- [5] Y. Qian, G. Yin, L. Sheng, Z. Chen, J. Shao, Thinking in frequency: Face forgery detection by mining frequency-aware clues, in: European conference on computer vision, Springer, 2020, pp. 86–103.
 - [6] D. Afchar, V. Nozick, J. Yamagishi, I. Echizen, Mesonet: a compact facial video forgery detection network, in: 2018 IEEE international workshop on information forensics and security (WIFS), IEEE, 2018, pp. 1–7.
 - [7] Y. Li, M.-C. Chang, S. Lyu, In ictu oculi: Exposing ai created fake videos by detecting eye blinking, in: 2018 IEEE International Workshop on Information Forensics and Security (WIFS), 2018, pp. 1–7. doi: [10.1109/WIFS.2018.8630787](https://doi.org/10.1109/WIFS.2018.8630787).
 - [8] M. Li, B. Liu, Y. Hu, Y. Wang, Exposing deepfake videos by tracking eye movements, in: 2020 25th International Conference on Pattern Recognition (ICPR), 2021, pp. 5184–5189. doi: [10.1109/ICPR48806.2021.9413139](https://doi.org/10.1109/ICPR48806.2021.9413139).
 - [9] A. Haliassos, K. Vougioukas, S. Petridis, M. 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 (CVPR), 2021, pp. 5039–5049.
 - [10] H. Qi, Q. Guo, F. Juefei-Xu, X. Xie, L. Ma, W. Feng, Y. Liu, J. Zhao, Deeprhythm: Exposing deepfakes with attentional visual heartbeat rhythms, in: Proceedings of the 28th ACM International Conference on Multimedia, MM '20, Association for Computing Machinery, New York, NY, USA, 2020, p. 4318–4327. doi: [10.1145/3394171.3413707](https://doi.org/10.1145/3394171.3413707)
URL <https://doi.org/10.1145/3394171.3413707>

- [11] I. Demir, U. A. Ciftci, Where do deep fakes look? synthetic face detection via gaze tracking, in: ACM Symposium on Eye Tracking Research and Applications, ETRA '21 Full Papers, Association for Computing Machinery, New York, NY, USA, 2021. doi:10.1145/3448017.3457387.
URL <https://doi.org/10.1145/3448017.3457387>
- [12] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, M. Niesser, Faceforensics++: Learning to detect manipulated facial images, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019.
- [13] Y. Li, X. Yang, P. Sun, H. Qi, S. Lyu, Celeb-df: A large-scale challenging dataset for deepfake forensics, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.
- [14] B. Zi, M. Chang, J. Chen, X. Ma, Y.-G. Jiang, Wilddeepfake: A challenging real-world dataset for deepfake detection, in: Proceedings of the 28th ACM International Conference on Multimedia, MM '20, Association for Computing Machinery, New York, NY, USA, 2020, p. 2382–2390. doi:10.1145/3394171.3413769.
URL <https://doi.org/10.1145/3394171.3413769>
- [15] G. E. Raptis, C. Katsini, M. Belk, C. Fidas, G. Samaras, N. Avouris, Using eye gaze data and visual activities to infer human cognitive styles: method and feasibility studies, in: proceedings of the 25th conference on user modeling, Adaptation and Personalization, 2017, pp. 164–173.
- [16] J. Kerr-Gaffney, A. Harrison, K. Tchanturia, Eye-tracking research in eating disorders: A systematic review, International Journal of Eating Disorders 52 (1) (2019) 3–27.
- [17] S. De Silva, S. Dayarathna, G. Ariyarathne, D. Meedeniya, S. Jayarathna, A. M. Michalek, Computational decision support system for

adhd identification, International Journal of Automation and Computing 18 (2) (2021) 233–255.

- [18] K. Kafka, A. Khosla, P. Kellnhofer, H. Kannan, S. Bhandarkar, W. Matuskik, A. Torralba, Eye tracking for everyone, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2176–2184.
- [19] J. He, K. Pham, N. Valliappan, P. Xu, C. Roberts, D. Lagun, V. Navalpakkam, On-device few-shot personalization for real-time gaze estimation, in: Proceedings of the IEEE/CVF international conference on computer vision workshops, 2019, pp. 0–0.
- [20] P. Majaranta, A. Bulling, Eye tracking and eye-based human–computer interaction, in: Advances in physiological computing, Springer, 2014, pp. 39–65.
- [21] X. Zhang, Y. Sugano, A. Bulling, Evaluation of appearance-based methods and implications for gaze-based applications, in: Proceedings of the 2019 CHI conference on human factors in computing systems, 2019, pp. 1–13.
- [22] A. Recasens*, A. Khosla*, C. Vondrick, A. Torralba, Where are they looking?, in: Advances in Neural Information Processing Systems (NIPS), 2015, * indicates equal contribution.
- [23] Y. Li, W. Shen, Z. Gao, Y. Zhu, G. Zhai, G. Guo, Looking here or there? gaze following in 360-degree images, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 3742–3751.
- [24] A. Recasens, C. Vondrick, A. Khosla, A. Torralba, Following gaze in video, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1435–1443.

- [25] X. Zhang, Y. Sugano, M. Fritz, A. Bulling, Appearance-based gaze estimation in the wild, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 4511–4520.
- [26] X. Zhang, Y. Sugano, M. Fritz, A. Bulling, Mpigaze: Real-world dataset and deep appearance-based gaze estimation, *IEEE transactions on pattern analysis and machine intelligence* 41 (1) (2017) 162–175.
- [27] Y. Cheng, F. Lu, X. Zhang, Appearance-based gaze estimation via evaluation-guided asymmetric regression, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 100–115.
- [28] X. Zhang, Y. Sugano, M. Fritz, A. Bulling, It’s written all over your face: Full-face appearance-based gaze estimation, in: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 2017, pp. 51–60.
- [29] A. Graves, S. Fernández, J. Schmidhuber, Bidirectional lstm networks for improved phoneme classification and recognition, in: International conference on artificial neural networks, Springer, 2005, pp. 799–804.
- [30] P. Kellnhofer, A. Recasens, S. Stent, W. Matusik, A. Torralba, Gaze360: Physically unconstrained gaze estimation in the wild, in: Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 6912–6921.
- [31] A. A. Abdelrahman, T. Hempel, A. Khalifa, A. Al-Hamadi, L2csnet: fine-grained gaze estimation in unconstrained environments, arXiv preprint arXiv:2203.03339.
- [32] Deepfakes, <https://github.com/deepfakes/faceswap>.
- [33] Faceswap, <https://github.com/MarekKowalski/FaceSwap>.
- [34] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, M. Niessner, Face2face: Real-time face capture and reenactment of rgb videos, in:

Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

- [35] L. Li, J. Bao, H. Yang, D. Chen, F. Wen, Faceshifter: Towards high fidelity and occlusion aware face swapping, arXiv preprint arXiv:1912.13457.
- [36] J. Thies, M. Zollhöfer, M. Nießner, Deferred neural rendering: Image synthesis using neural textures, Acm Transactions on Graphics (TOG) 38 (4) (2019) 1–12.
- [37] J. Thies, M. Zollhöfer, M. Nießner, L. Valgaerts, M. Stamminger, C. Theobalt, Real-time expression transfer for facial reenactment., ACM Trans. Graph. 34 (6) (2015) 183–1.
- [38] T. Karras, S. Laine, T. Aila, A style-based generator architecture for generative adversarial networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.
- [39] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, T. Aila, Analyzing and improving the image quality of stylegan, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.
- [40] Y. Nirkin, Y. Keller, T. Hassner, Fsgan: Subject agnostic face swapping and reenactment, in: Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 7184–7193.
- [41] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, J. Choo, Stargan: Unified generative adversarial networks for multi-domain image-to-image translation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.
- [42] T. Karras, T. Aila, S. Laine, J. Lehtinen, Progressive growing of gans for improved quality, stability, and variation, arXiv preprint arXiv:1710.10196.

- [43] O. Giudice, L. Guarnera, S. Battiato, Fighting deepfakes by detecting gan dct anomalies, *Journal of Imaging* 7 (8) (2021) 128.
- [44] X. Zhang, S. Karaman, S.-F. Chang, Detecting and simulating artifacts in gan fake images, in: 2019 IEEE international workshop on information forensics and security (WIFS), IEEE, 2019, pp. 1–6.
- [45] K. Chandrasegaran, N.-T. Tran, N.-M. Cheung, A closer look at fourier spectrum discrepancies for cnn-generated images detection, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 7200–7209.
- [46] T. Dzanic, K. Shah, F. Witherden, Fourier spectrum discrepancies in deep network generated images, *Advances in neural information processing systems* 33 (2020) 3022–3032.
- [47] D. Liu, Z. Zheng, C. Peng, Y. Wang, N. Wang, X. Gao, Hierarchical forgery classifier on multi-modality face forgery clues, *IEEE Transactions on Multimedia*.
- [48] X. Pan, X. Zhang, S. Lyu, Exposing image splicing with inconsistent local noise variances, in: 2012 IEEE International conference on computational photography (ICCP), IEEE, 2012, pp. 1–10.
- [49] P. Buchana, I. Cazan, M. Diaz-Granados, F. Juefei-Xu, M. Savvides, Simultaneous forgery identification and localization in paintings using advanced correlation filters, in: 2016 IEEE International Conference on Image Processing (ICIP), IEEE, 2016, pp. 146–150.
- [50] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
- [51] F. Chollet, Xception: Deep learning with depthwise separable convolutions, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1251–1258.

- [52] D. Cozzolino, G. Poggi, L. Verdoliva, Recasting residual-based local descriptors as convolutional neural networks: an application to image forgery detection, in: Proceedings of the 5th ACM workshop on information hiding and multimedia security, 2017, pp. 159–164.
- [53] H. H. Nguyen, J. Yamagishi, I. Echizen, Use of a capsule network to detect fake images and videos, arXiv preprint arXiv:1910.12467.
- [54] T. Jung, S. Kim, K. Kim, Deepvision: Deepfakes detection using human eye blinking pattern, IEEE Access 8 (2020) 83144–83154. doi:10.1109/ACCESS.2020.2988660.
- [55] W. Wang, Z. Wang, G. Wang, Q. Zou, Deepfake video detection exploiting binocular synchronization, in: E. Pimenidis, P. Angelov, C. Jayne, A. Papaleonidas, M. Aydin (Eds.), Artificial Neural Networks and Machine Learning – ICANN 2022, Springer Nature Switzerland, Cham, 2022, pp. 101–112.
- [56] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural computation 9 (8) (1997) 1735–1780.
- [57] F. Matern, C. Riess, M. Stamminger, Exploiting visual artifacts to expose deepfakes and face manipulations, in: 2019 IEEE Winter Applications of Computer Vision Workshops (WACVW), 2019, pp. 83–92. doi:10.1109/WACVW.2019.00020.
- [58] J. Quinonero-Candela, M. Sugiyama, A. Schwaighofer, N. D. Lawrence, Dataset shift in machine learning, 2008.
- [59] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30.
- [60] J. Wang, E. Olson, Apriltag 2: Efficient and robust fiducial detection, in: 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2016, pp. 4193–4198.

- [61] J. Deng, J. Guo, E. Ververas, I. Kotsia, S. Zafeiriou, Retinaface: Single-shot multi-level face localisation in the wild, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 5203–5212.
- [62] J. Carreira, A. Zisserman, Quo vadis, action recognition? a new model and the kinetics dataset, in: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6299–6308.
- [63] D. Tran, L. D. Bourdev, R. Fergus, L. Torresani, M. Paluri, C3d: generic features for video analysis, CoRR, abs/1412.0767 2 (7) (2014) 8.
- [64] Z. Liu, D. Luo, Y. Wang, L. Wang, Y. Tai, C. Wang, J. Li, F. Huang, T. Lu, Teinet: Towards an efficient architecture for video recognition, in: Proceedings of the AAAI conference on artificial intelligence, Vol. 34, 2020, pp. 11669–11676.
- [65] X. Li, Y. Lang, Y. Chen, X. Mao, Y. He, S. Wang, H. Xue, Q. Lu, Sharp multiple instance learning for deepfake video detection, in: Proceedings of the 28th ACM international conference on multimedia, 2020, pp. 1864–1872.
- [66] Z. Gu, Y. Chen, T. Yao, S. Ding, J. Li, F. Huang, L. Ma, Spatiotemporal inconsistency learning for deepfake video detection, in: Proceedings of the 29th ACM international conference on multimedia, 2021, pp. 3473–3481.
- [67] C. Zhao, C. Wang, G. Hu, H. Chen, C. Liu, J. Tang, Istvt: interpretable spatial-temporal video transformer for deepfake detection, IEEE Transactions on Information Forensics and Security 18 (2023) 1335–1348.