Extracting Local Information from Global Representations for Interpretable Deepfake Detection[*]

Elahe Soltandoost[1], Richard Plesh[2], Stephanie Schuckers[3], Peter Peer[4], Vitomir Štruc[4]
[1] University of Padova, Via VIII Febbraio 2, 35122 Padova PD, Italy
[2]Clarkson University, Potsdam, NY 13699 USA
[3]University of North Carolina Charlotte, Charlotte, North Carolina, USA
[4]University of Ljubljana, 1000 Ljubljana, Slovenia

Abstract

The detection of deepfakes has become increasingly challenging due to the sophistication of manipulation techniques that produce highly convincing fake videos. Traditional detection methods often lack transparency and provide limited insight into their decision-making processes. To address these challenges, we propose in this paper a Locally-Explainable Self-Blended (LESB) DeepFake detector that in addition to the final fake-vs-real classification decision also provides information, on which local facial region (i.e., eyes, mouth or nose) contributed the most to the decision process. At the heart of the detector is a novel Local Feature Discovery (LFD) technique that can be applied to the embedding space of pretrained DeepFake detectors and allows identifying embedding space directions that encode variations in the appearance of local facial features. We demonstrate the merits of the proposed LFD technique and LESB detector in comprehensive experiments on four popular datasets, i.e., Celeb-DF, DeepFake Detection Challenge, Face Forensics in the Wild and FaceForensics++, and show that the proposed detector is not only competitive in comparison to strong baselines, but also exhibits enhanced transparency in the decision-making process by providing insights on the contribution of local face parts in the final detection decision.

1. Introduction

The emergence of DeepFakes (i.e., hyper-realistic AI-generated imagery often used for malicious purposes) has recently become a significant concern, as the tools for creating falsified multimedia content have evolved and are also increasingly accessible [5, 42, 44]. Contemporary DeepFakes are nearly indistinguishable from authentic media, making it difficult for humans to discern what is real from what is fake, thereby raising ethical, security, and trust issues [28,48,60]. The rapid advancements in artificial intelligence (AI) and machine learning (ML) has, on the one hand, facilitated this rise, but also made it possible to develop automatic detection techniques that are critical for combating the threat posed by DeepFake generation technology [52].

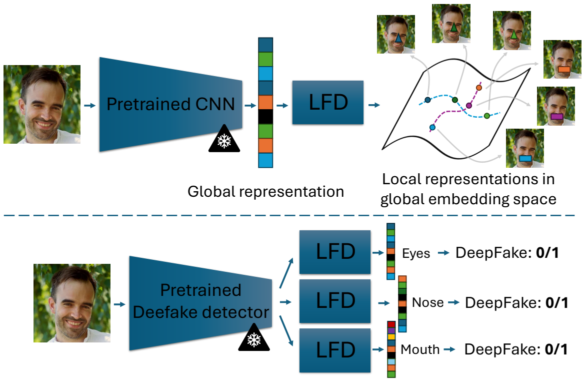
Pretrained CNN
Global representation
LFD
Local representations in global embedding space
Pretrained Deefake detector
LFD
LFD
LFD
Eyes → DeepFake: **0/1**
Nose → DeepFake: **0/1**
Mouth → DeepFake: **0/1**

Figure 1. **Illustration of the main contributions of this work.** We introduce a *Local Feature Discovery* (LFD) technique that allows us to identify feature space directions in the embedding space of pretrained CNNs that correspond to variations in spatially local features, despite the fact the embedding space encodes global image information (top part). We then train multiple LFD models to discover different local feature spaces, e.g., for the mouth, nose, and eyes in case of facial images, and apply those to a pretrained Deepfake detector to *improve its interpretability* (bottom part).

State-of-the-art deepfake detection methods today still mostly rely on discriminative models based on, e.g., Convolutional Neural Networks (CNNs), that aim to classify images into one of two categories: real or fake [25,26]. While such models typically achieve impressive performance on datasets with known DeepFake techniques, they often struggle with generalization, particularly when applied to unseen manipulation techniques. This issue of generalizability led researchers to explore alternative approaches, including self-supervised learning and one-class models. Chen *et al.*

[*]Supported by ARIS through P2-0250, P2-0214 and J2-50065.

[10], for example, introduced a self-supervised method that enhances model sensitivity and generalizability by dynamically generating challenging forgery examples. Similarly, Lee *et al.* [61] proposed a method that leveraged one-class domain generalization and frequency domain processing to address the challenge of unseen manipulation methods. In addition to these one-class methods, other recently applied strategies for improving generalization included the design of generalizable image representations [21], advanced representation-learning approaches [33], anomaly detection methodology [34] and other similar strategies [56].

Despite these advancements, the majority of existing DeepFake detectors still function as "*black boxes*" with complex input-output mappings that are challenging to interpret, causing transparency issues. This is problematic for multiple reasons: (i) first, such detectors are not complying with privacy laws, such as the General Data Protection Regulation (GDPR) [18], which explicitly requires automated-decision making systems that may impact humans to be interpretable; (ii) second, the opaqueness of existing detectors makes it challenging to confirm the correctness of the results and leaves the door open for potentially malicious model tempering, and (iii) finally, human examiners that are often required to validate the results of automatic detectors are provided only limited information about the detection result, making is difficult to objectively validate the detection results. The complexity and opacity of these models, hence, make it critical to design explainability mechanisms that help making the predictions on modern DeepFake detectors more interpretable and transparent [49].

To address this gap, we present in this paper a novel approach that aids in explaining the decisions of pretrained DeepFake detectors by highlighting local facial regions that contribute most to a given decision. In other words, instead of providing only a *fake-vs-real* decision, our approach allows emphasizing whether regions, such as the eyes, nose or the mouth, were the most important for the detection result, thus offer better insight into the decision process. At the core of the approach is a novel **Local Feature Discovery** (LFD) technique that can be applied to the embedding space of pretrained DeepFake detectors and that is able to identify embedding space directions that correspond to variations in spatially local facial features (e.g., the eyes, nose or the mouth), as illustrated in the top part of Figure 1. The proposed LFD technique allows us to analyze the appearance of local facial features within the typically global embedding space of contemporary DeepFake detectors and, in turn, to design detection models that base their decisions on local facial parts rather than global appearance, as shown at the bottom of Figure 1. Thus, by extracting local information from the typically global internal representations of DeepFake detection models, we are able to infer cues about the importance of the individual regions and provide additional output for the detector next to the binary fake-vs-real decision. To demonstrate the feasibility of the such an approach, we incorporate the proposed technique in a state-of-the-art Deepfake detection model and illustrate its merits in comprehensive experiments on multiple publicly available dataset, i.e., Celeb-DF (CDF) [36], DeepFake Detection Challenge (DFDC) [16], Face Forensics in the Wild (FFIW) [66] and FaceForensics++ (FF++) [53].

In summary, the main contributions of this paper are:

- We propose a **Local Feature Discovery** (LFD) technique that allows us to analyze spatially local data variations within the global embedding space of pretrained DeepFake detectors that originally exploit global facial appearance. Through rigorous experiments, we show that this approach provides additional insights into the detection process, thus improving the transparency of the results.
- We incorporate LFD into a state-of-the-art (SOTA) self-supervised approach that relies on so-called self-blended images [57] to produce simulated pseudo deepfake data for DeepFake detector training. The integration leads to a **Locally Explainable Self-Blended** detector (LESB) that matches the performance of the original approach but improves its interpretability and transparency.
- We evaluate the proposed techniques in experiments on multiple popular datasets and demonstrate the benefits of modeling local image variations in the global embedding space of DeepFake detectors for interpretability purposes.

2. Related Work

In this section, we review the existing literature on deepfake detection and explainable AI (xAI) to provide the necessary background for our research. For a more comprehensive coverage of these areas, the reader is referred to some of the exceptional surveys on this topic, e.g., [17, 44, 54, 60].

DeepFake Detection Methods. Early DeepFake detectors were mostly based on known shortcomings of deepfake generation approaches and strived to detect explicit traces of the generation process [1, 3, 14, 20, 62]. Later techniques focused on learning-based methods, where discriminative models [55, 59] over spatial and frequency domain features were trained to distinguish between pristine and falsified imagery [13, 25, 38, 51, 55]. While these methods significantly improved performance, they still face generalization problems and, hence, lack adaptability to new datasets [26].

To address these challenges, researchers started looking increasingly towards self-supervised learning and one-class anomaly detection techniques [23, 30, 58] that are capable of learning from authentic/pristine/real data only, which in turn, significantly improved the generalization of DeepFake detection models and applicability of modern detectors in cross-dataset (cross-deepfake) settings [2, 10, 19, 27, 33, 35, 37, 45, 57, 64]. Specifically noteworthy here are techniques,

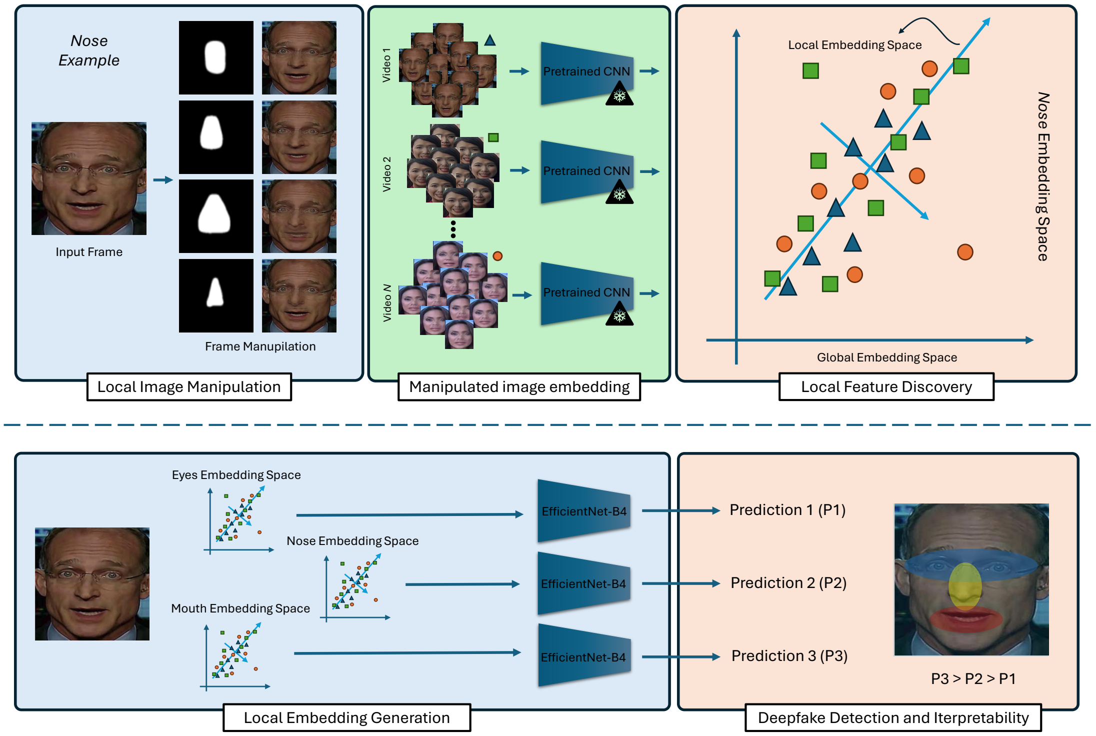

Figure 2. **Overview of the proposed Local Feature Discovery (LFD) approach (top) and its application for explainable DeepFake detection (bottom).** In the first step, masks for the targeted face part are extracted (illustrated for the nose), and different augmentations are then applied within these regions to generate images with spatially local variations. These variations are then modeled using Principal Component Analysis (PCA) to generate a local subspace within the global embedding space of the pretrained CNN used. The learned PCA transforms of different face parts are then applied in the inference stage to generate local embeddings, based on which a fake-vs-real decision is taken. Because multiple predictors are used, the outputs can be analyzed to facilitate interpretability.

such as Face X-Ray [34], Self-Blended Images [57] or SeeABLE [33] that rely on blending procedures.

While CNN-based models have improved deepfake detection accuracy, their black-box nature raises concerns about trust and interpretability. This has prompted the rise of Explainable AI (xAI) techniques, which seek to provide more transparent insights into model decision-making.

Explainable AI (xAI). Most relevant to the research presented in this paper are two groups of xAI techniques, namely, (i) **attribution approaches** and (ii) **feature-space exploration** methods. The primary objective of attribution approaches is to identify important regions or pixels within an image that are important for model predictions. Methods that produce saliency maps that emphasize important image regions for the decision process [31, 43] are common examples of techniques from this group and the Local Interpretable Model-Agnostic Explanations (LIME) approach from [6, 8, 40] is a SOTA example specifically designed to highlight significant areas in deepfake detection models.

Embedding- and feature-space exploration approaches, on the other hand, enhance interpretability by analyzing data representations within a model's embedding space [32, 50]. Certain embedding exploration techniques, such as those discussed by [24, 46, 47], are capable of decoding features, such as stance or gender. However, they struggle with accurately detecting intricate local details, particularly when alterations are subtle. Methods that try to explore and understand prototypical representations in the models' feature space also face challenges, as prototypes can be overly similar, making it difficult to differentiate among them and understand their contributions effectively [4, 44].

While numerous xAI techniques have been proposed to improve the interpretability of deepfake detectors, many of these approaches focus on broad features or global overall patterns. Such methods provide valuable insights, but often fail to examine specific facial regions in detail [24, 46, 47]. Unlike these methods, we develop in this paper a Local Feature Discovery (LFD) technique that emphasizes localized facial features and their contribution to the detection process. Deepfakes often alter specific parts of a face while leaving others unchanged, making broad methods prone to missing subtle signs of manipulation. Our approach, through the proposed LFD technique, is designed to scrutinize these specific features, enhancing the transparency and interpretability of deepfake detectors.

3. Methodology

In this section, we present the two main contributions of this work, that is, (i) the novel *Local Feature Discovery (LFD)* technique that allows to model spatially local facial variations in the global embedding space of pretrained DeepFake detectors, and (ii) the *Locally Explainable Self-Blended detector (LESB)* that incorporates the LFD approach into a state-of-the-art (SOTA) self-supervised DeepFake detector to make it explainable.

3.1. Local Feature Discovery (LFD)

The proposed Local Feature Discovery techniques, illustrated at the top of Figure 2, aims to identify a subspace that corresponds to variations in local facial features within the embedding space of a pretrained DeepFake detection model that in general encodes global facial appearances. This is achieved by inducing image transformations/augmentations within spatially local and semantically meaningful image areas that correspond to prominent features, such as the eyes, nose or the mouth. The top part of Figure 2 illustrates this process for the nose region. Once the variations are generated, a dedicated PCA model is learned for each facial part that defines a subspace within the global embedding space, in which local facial-part variations are encoded. This approach allows the model to isolate specific local features in the embedding space of a pretrained DeepFake detector and, in turn, analyze how local features contribute to the fake-vs-real decision at the model's output.

The Local Feature Discovery approach consists of several steps, including: (i) local face-mask generation, (ii) data augmentation, and (iii) PCA modelling. Details on these steps are given below.

Local Face-Mask Generation. The first step in the LFD pipeline is the generation of the local face-masks that correspond to prominent facial regions. To this end, we first apply the RetinaFace detector [15] to detect the facial region in all input images, and then adopt the face alignment model from Dlib with a custom shape predictor that extracts 81 landmarks from the face, extending the standard 68 landmarks with 13 additional points on the forehead [12], as shown on the left side of Figure 3.

Given the detected landmarks, we then define binary masks that correspond to three distinct facial regions, i.e., the eyes, the nose and the mouth. Each binary mask is then augmented using different shape transformations. In Figure 3, we illustrate this process for the nose region, but the process equally applies to other face areas. Here, the first image of each row shows the basic nose shape extracted based on the detected landmarks from a given input image, while the remaining masks correspond to perturbed regions with variations in overall shape and position. The illustrated perturbation process allows us to better model local face variations that are critical for designing an informative subspace that can model local face variations within the existing embedding spaces of pretrained neural networks.

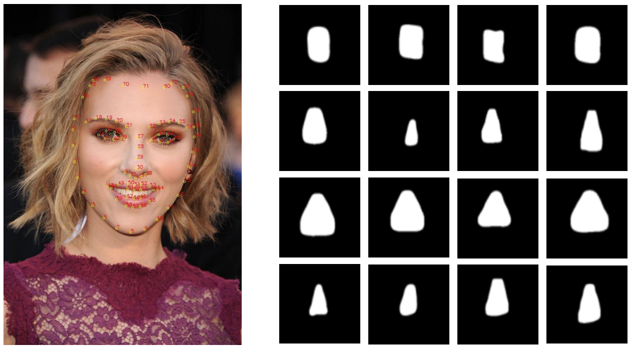

Figure 3. **Local face-mask generation.** We use the Dlib landmarking model to generate an 81 point markup of the facial region, as shown on the left [12]. Based on the detected landmarks, we then define local binary masks that correspond to the facial region-of-interest. In the 4×4 grid on the right, each row shows the basic nose masks extracted from a given input image in the first column, followed by variations in shapes, dimensions in positions in the remaining columns.

Data Augmentation. In the second step, we employ various data augmentation strategies to introduce a rich set of appearance variations into the facial images. Let x denote a facial image from a given dataset, let m be one of the binary masks, illustrated in Figure 3, smoothed with a Gaussian filter, and let ψ denote a selected augmentation technique from the following groups:

- **Color transforms:** the first group of image augmentation techniques includes global color transformations, such as random shifts in RGB channels, hue, saturation, value, brightness, and contrast. These color transform, when applied to the input images, induce various color distortions, similar to the ones introduced by the DeepFake generation models.
- **Frequency transforms:** the second group of augmentation techniques includes frequency-domain transformations, such as downsampling or sharpening, and are used to model image variations that impact the frequency characteristics of the facial images.
- **Geometric transforms:** the last group of transform includes geometric transformations. Here, the input face images are first zero-padded in all directions, randomly translated or scaled and then center-cropped to produce output images with slight geometric perturbations.

To produce an augmented image with perturbed local spatial regions x_a, we randomly apply the transformations from the three above groups to the input image to produce a globally distorted result, i.e., $x_d = \psi(x)$ and then blend the result

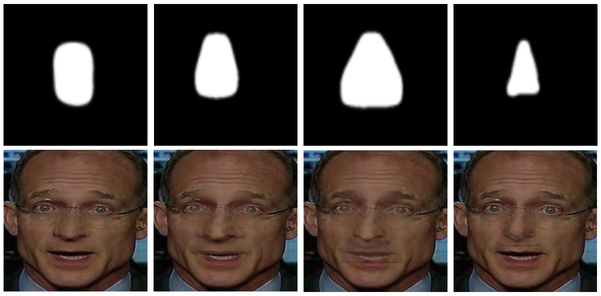

Figure 4. **Examples of facial images after blending with the manipulated masks.** The top row shows different blending masks, while the bottom row illustrates some of the local nose variations that are induced for learning the LDF transform within the nose regions. A similar process is also used for other facial parts.

with the original source image x as follows:

$$x_a = (1 - m) \odot x + m \odot x_d, \tag{1}$$

where $\odot$ stands for the Hadamard product. An illustrative example of the effect of these augmentations when manipulating the nose region is shown in Figure 4. Note how the augmented images, x_a, correspond in most pixel values and due to the blending process differ only in the local variations around the targeted facial region. This augmentation procedure, hence, allows us to produce facial images with spatially constrained appearance variations that allow us to probe the embedding space of deep learning models and, consequently, model the local variations using principal component analysis (PCA).

Local Subspace Modelling with PCA. To model local face part variations, we use a dataset of facial images $\mathcal{D} = \{x^{(i)}\}_{i=1}^{N}$, and for each of the N images in $\mathcal{D}$, we apply the augmentation procedure described in the previous section. If we denote the number of augmentations of each image from $\mathcal{D}$ as M, we can define an augmented dataset $\mathcal{D}_a$ as follows: $\mathcal{D}_a = \{x_a^{(i,j)}|_{j=1:M}^{i=1:N}\}$. Using a pretrained CNN model ϕ, we can then project the augmented images into the embedding space of ϕ, leading to a set of d-dimensional embeddings $\mathcal{E} = \{e_a^{(i,j)}|_{j=1:M}^{i=1:N}\} \in \mathbb{R}^d$. As noted above, these embeddings correspond to set of input images that are identical in the majority of pixel values, except for the local face variations induced by our data augmentation process.

To model these local variations, we first remove information that corresponds to the global input image characteristics. Thus, we compute input-image conditional means in the embedding space, i.e.:

$$\mu_j = \frac{1}{M} \sum_{k=1}^{M} e_a^{(i,k)}. \tag{2}$$

Next, we remove these means from the embeddings to ensure that only local face part variations are captured in the corresponding scatter, i.e.:

$$E = [e_a^{(i,j)} - \mu_j]\Big|_{j=1:M}^{i=1:N}. \tag{3}$$

Finally, we center the embeddings around a global mean $\bar{E}$, compute the covariance matrix, i.e.:

$$\Delta = \frac{1}{N(N-1)}(E - \bar{E})(E - \bar{E})^T, \tag{4}$$

and then solve the PCA-defined eigenproblem:

$$\Delta V = \Lambda V, \tag{5}$$

where the leading d' eigenvectors of the problem, i.e., $V = [v_1, v_2, \ldots, v_{d'}] \in \mathbb{R}^{d \times d'}$, with $d' \leq d$, jointly constitute the local subspace that captures variations in face-part appearance in the embedding space of ψ. This subspace can, hence, be analyzed to make face-part based inferences about the input image and represents the basic component of the proposed Local Feature Discovery (LFD) approach.

Local Subspace Projection. To analyze images based on local face characteristics, it is necessary to project the given input face into the subspace identified by LFD, as illustrated in the top part of Figure 1. Formally, this can be described as follows. Let x represent a face image and let the corresponding embedding computed through the pretrained CNN model ψ be denoted as e, i.e., $e = \psi(x)$. If we assume that V stands the subspace encoding local face part variations, then the centered version of the embedding e, i.e., e_c can be projected onto the subspace as:

$$e_{loc} = V^T e_c, \tag{6}$$

where $e_{loc} \in \mathbb{R}^{d'}$ is the embedding in the identified local (e.g., nose, mouth or eye) subspace.

3.2. Locally-Explainable Self-Blended Detector

Next, we incorporate the proposed LFD approach into a state-of-the-art DeepFake detector to make its decision process more transparent. Specifically, we extend the Self-Blended Images (SBI) from [57] with the LFD technique, as depicted in the bottom part of Figure 1

Prerequisites. SBI represents a powerful approach towards learning DeepFake detectors that performs well across different datasets and DeepFake types and generalizes well to unseen face-manipulation techniques. It falls into the group of detectors that learn only from pristine images, while never observing an actual DeepFake during the training process. The main idea of SBI is to generate so-called *pseudo-fakes* from pristine images by simulating common artifacts that are typically present in a wide range of falsified images. These include color distortions, frequency and compression artifacts, and blending-induced degradations. A discriminative model (e.g., a CNN-based classifier) is then trained to distinguish between pristine facial images and the manipulated images with the added artifacts. The main feature distinguishing SBI from other conceptually similar techniques lies in the fact that the target and source face to be used in the simulated blending procedure are in fact identical, hence, the name Self-Blended Images. This self-blending leads to subtle manipulation traces in the pseudo-fake images, forcing the classifier to learn generalizable decision boundaries and, in turn, a DeepFake detector that generalizes well across different data characteristics. Further details on the SBI model are available in [57].

Incorporating LFD into SBI. In the next step, we integrate LFD into a pretrained SBI model. Specifically, we use the pretrained EfficientNet-B4 [63] from the SBI GitHub repository and extend it to analyze local face parts instead of global facial appearances. The utilized EfficientNet-B4 detector was trained with pristine facial images and pseudo-fakes produced by SBI and was shown in [57] to lead to competitive DeepFake detection results on various datasets.

Let the mapping function that maps a given input image x into the embedding space of the EfficientNet-B4 detector be denoted as χ and let the LFD transforms for the nose, mouth and eye region be denoted as V_{nose}, V_{mouth} and V_{eyes}, respectively. The embeddings, encoding the local face parts, can then be computed as follows:

$$e_k = V_k^T(\chi(x)), \tag{7}$$

where $k \in \{nose, mouth, eyes\}$ and $e_k \in \mathbb{R}^{d'}$.

Given a dataset of pristine facial images $\{x_i\}_{i=1}^N$ and corresponding pseudo-fakes generated by SBI, we can learn a classifier (in the form of a simple Multi-Layer Perceptron (MLP)) over the respective local embeddings from Eq. (7). At run-time, each MLP then outputs a prediction on whether the input image is a DeepFake or a pristine face, as also illustrated in the lower part of Figure 2.

The Locally-Explainable Self-Blended Detector. While the number of facial parts considered is, in general, not limited, we use three facial parts in this paper to design the Locally-Explainable Self-Blended Detector (LESB). Assume that an MLP model was trained for each of the three facial parts of interest, i.e., the eyes, nose and mouth. Each MLP then outputs a separate prediction on whether the input image is a DeepFake or not. The prediction scores, i.e., p_1, p_2 and p_3, are defined on the range $[0, 1]$ with higher scores corresponding to more confident decision towards the input being a DeepFake. By comparing the three scores and associating them with the relevant facial regions, it is possible to interpret, which region contributed stronger towards the final decision, which is based on a simple sum of the individual predictions. This idea is illustrated through the toy example in the lower right of Figure 2.

4. Experiments and Results

In this section, we present the experiments to evaluate the characteristics of the LFD techniques and the performance and interpretability of the LESB DeepFake detector.

4.1. Experimental Setup

Datasets. For the experiments, we select four popular DeepFake datasets, i.e., Celeb-DF (CDF) [36], DeepFake Detection Challenge (DFDC) [16], Face Forensics in the Wild (FFIW) [66] and FaceForensics++ (FF++) [53]. High-level details on the datasets are given below, i.e.:

- **Celeb-DF (CDF) [36]** consists of 590 real and $5,639$ deepfake videos, featuring celebrities. The DeepFakes in this dataset were produced with a highly sophisticated DeepFake generation approach, making effective detection a challenging task.
- **DeepFake Detection Challenge (DFDC) [16]** comprises over $128,000$ video sequences with more than $100,000$ deepfakes and is widely acknowledged as one of the most challenging dataset available.
- **Face Forensics in the Wild (FFIW) [66]** consists of $10,000$ high-quality videos, each containing an average of three human faces per frame. FFIW is designed to evaluate DeepFake detectors in real-world, multi-person scenarios and includes deepfakes generated by a domain-adversarial quality assessment network.
- **FaceForensics++ (FF++) [53]** comprises 1000 videos that are designated for: training (720 videos), validation (140 videos) and testing (140 videos). The DeepFakes in FF++ were generated with four different techniques, i.e., DeepFake (DF), FaceSwap (FS), Face2Face (F2F) and Natural Textures (NT).

Performance Measures. To evaluate the performance of the DeepFake detectors, we utilize the *Area Under the Curve (AUC)*, similar to prior work [2,7,10,27,29,33,35,37,45,64]. AUC scores quantify Receiver Operating Characteristics (ROC) curves using a single (threshold-independent) number and are, therefore, a convenient and popular performance measure for evaluating DeepFake detectors.

Baselines. When evaluating DeepFake detection performance, we compare LESB to multiple state-of-the-art techniques, including DSP-FWA [35], LRL [11], FRDM [39], PCL + I2G [64], Two-branch [41], DAM [67], LipForensics [22], FTCN [65] and the original SBI [57] approach, implemented with an EfficientNet-B4 (EFNB4) backbone to ensure a fair comparison.

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

In the first experiment, we compare the performance of the proposed models with competing state-of-the-art detectods across different datasets. We evaluate both, our local

Method	**CDF**	**DFDC**	**FFIW**
DSP-FWA [35]	69.30	-	-
LRL [11]	78.26	-	-
FRDM [39]	79.40	-	-
PCL + I2G [64]	90.03	67.52	-
Two-branch [41]	76.65	-	-
DAM [67]	75.30	-	-
LipForensics [22]	82.40	-	-
FTCN [65]	86.90	71.00	74.47
EFNB4 + SBI [57]	93.18	72.42	84.83
LFD-Eyes (Ours)	92.63	71.29	80.56
LFD-Nose (Ours)	93.22	70.74	83.21
LFD-Mouth (Ours)	93.85	71.12	83.30
LESB (Ours)	93.13	71.98	83.01

Table 1. **Comparison of AUC scores (in %) across different datasets.** We compare detectors trained over local face-part subspaces, denoted as LFD-*Part*, the proposed LESB detector, and multiple state-of-the-art baselines.

DeepFake detectors that learn an MLP classifier in the loca face-part subspace (denoted as LFD-Eyes, LFD-Nose and LFD-Mouth), as well as the combined LESB detector that considers the outputs of all three local models. For this experiment, all models are trained on the FF++ dataset and tested in cross-dataset settings on CDF, DFDC and FFIW.

From the reported results in Table 1, we observe that the local LFD-basedd models perform quite similarly to the original SBI approach and maintain its strong performance, despite being based on local face-part variations. Similarly, the combined LESB model also yield comparable performance with convincing results on all three datasets. When compared to other DeepFake detectors, we see that all LFD-based models, the LESB detector and the SBI approach considerably outperform all other competitors on the CDF and FFIW datasets, while also being competitive on DFDC.

4.3. Alternative LESB Designs

Next, we explore alternative designs for the LESB detector. For the results reported in the previous section, LESB was implemented by combining the predictions from the local nose, eye and mouth models through a simple sum. Since, this may not be optimal, we investigate various fusion strategies in this section, including a maximum score, minimum score, median score, geometric mean and harmonic mean fusion strategies. Additionally, we also consider a decision-level scheme using majority voting.

From the results reported in Table 2, we observe that different combinations of the local LFD-based models yield comparable performances, with small performance differences being mostly visible on individual datasets. The only exception herre is the voting scheme, which led to significantly worse results on all four considered datasets. Among the score-level strategies, no fusion scheme is evidently superior to any other, suggesting that the sum rule is a viable choice for the implementation of LESB. This implementation is, therefore, also used in all following analyses of the proposed detector.

Method	**FFIW**	**CDF**	**FF++**	**DFDC**	**Average**
EFNB4 + SBI (Shiohara et al. 2022)	84.83	93.18	99.59	72.42	87.55
LESB (Sum/Arithmetic Mean)	83.01	93.13	99.54	71.98	86.92
LESB - Maximum score	82.01	93.09	99.61	71.44	86.54
LESB - Minimum score	84.11	93.35	99.46	71.21	87.03
LESB - Median score	83.43	93.38	99.55	72.08	87.11
LESB - Geometric Mean	83.46	93.38	99.54	71.82	87.05
LESB - Harmonic Mean	83.49	93.37	99.53	71.51	86.87
LESB - Majority Voting	70.00	82.52	97.59	60.33	77.61

Table 2. **Analysis of alternative LESB designs.** The table shows AUC scores (in %) for the LESB detector implemented with various fusion strategies.

4.4. Visualization of Model Focus Using Heatmaps

Next, we investigate the characteristics of the developed local LFD-based detectors. Specifically, we are interested in the facial regions the models focus on when making decisions. To this end, we use GradCAM++ [9] and visualize image areas the contribute most to the activations of the local MLP classifiers. The goal of this experiments is to understand models behavior on one end, and to validate the design of the LFD-based models on the other.

In Figure 5, we shows heatmaps for an example face images for thethe "Eyes", "Nose", and "Mouth" models compared to a baseline SBI model trained on the entire face. The heatmaps reveal that our region-specific models exhibit focused attention on prominent facial areas such as the eyes, nose, and mouth, and, thus conduct DeepFake detection by analyzing relevant local facial features. Conversely, the baseline SBI model, denote as "Face", shows relatively dispersed attention, indicating less effective focus on potentially informative local face regions.

4.5. Explaining Decisions with LESB

As demonstrated in the previous two sections, our LESB detector exhibits comparable performance as the original SBI approach. However, its added value lies in the capability of providing additional insight into the decision process by illustrating which facial regions contribute most to the detection result. In this section, we demonstrate this capability through several qualitative examples. Specifically, we generate heatmaps for the eye, nose and mouth regions based on the predictions of the local MLP classifier that are jointly used in the proposed LESB DeepFake detector.

Interpreting Decisions by Type. In Figure 6, we show various face examples that resulted in different model predictions, including True Positives (TP), False Negatives (FN), True Negatives (TN) and False Positives (FP), where DeepFakes are considered to constitute the positive class.

As can be seen from the TP results, LESB bases it de-

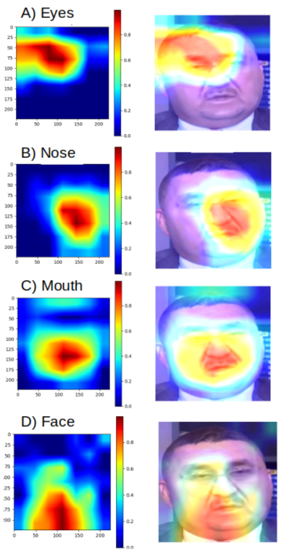

Figure 5. **Comparison of GradCAM++ heatmaps.** We compare the local LFD-based models, marked as Eyes, Nose, and Mouth, to the original global SBI model, denoted as Face, on an example face image. The heatmaps highlight facial areas that contribute the most to the activations of the corresponding classifiers/detectors. Note how the local models focus on prominent facial features, whereas the global SBI model exhibits more dispersed focus that covers a comparably larger areas of the face.

tection results on prominent face areas, marking the nose, mouth, and eyes in blue (high confidence) for DeepFakes. Nevertheless, it overlooks small manipulations, as seen in the absence of blue in different regions in the FN results. With the TN results (i.e., pristine images) most regions are designated as not being altered/falsified as seen in the red and light-green colored areas and with the FP results. LESB incorrectly assigns high confidence scores (in blue) to all facial areas, revealing the sources of error in these decisions.

Interpreting Decisions across Datasets. In Figure 7, we show the contribution of local face regions to the decision process across different datasets and focus only on manipulate images, i.e., DeepFakes. Based on the presented results, we can make some interesting observations, i.e.:

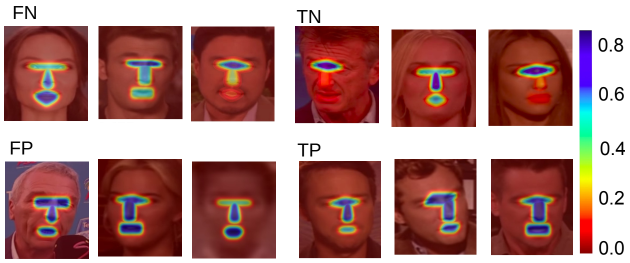

Figure 6. **Contribution of local regions to the decision process.** The examples present highlighted local face regions, where the color corresponds to the importance of the region for the detector prediction, in accordance with the color scale on the right.

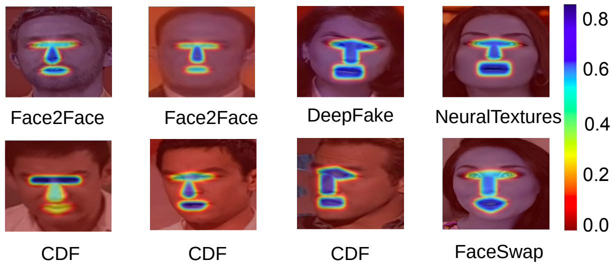

Figure 7. **Contribution of face regions to the decision procedure across datasets.** Blue areas indicate regions of higher confidence, while yellow areas correspond to lower confidence regions.

- **Face2Face:** On Face2Face, our visualizations show a strong focus on the nose and sometimes on the mouth regions. This suggests that manipulations in these regions are more detectable, possibly due to artifacts introduced during the face-swapping process.
- **DeepFake:** Here, our approach assigns equal confidence to the eyes, nose and mouth regions. The model's emphasis on these areas might be due to more pronounced alterations in the central facial features found in this dataset.
- **NeuralTextures:** On NeuralTextures, our models significantly highlights the nose and mouth, and to a lesser extent the eye region. This pattern reflects the specific manipulation techniques used in this dataset, which may cause more detectable distortions in these areas.
- **FaceSwap:** The heatmaps for FaceSwap show the highest confidence for the mouth, with notable attention also on the eyes and nose. This suggests that FaceSwap manipulations particularly affect dynamic facial areas, e.g., the mouth, making these regions crucial for detection.
- **CDF:** The heatmaps for CDF show a varied distribution of attention across different facial regions. Specifically, LESB adapts its focus based on specific manipulations in each image. For instance, some images show high confidence for the nose, while others highlight the mouth or eyes. This variability indicates that LESB is sensitive to a wide range of manipulations and can adjust its detection focus based on the unique characteristics of each image.

5. Conclusions

We presented a novel Locally-Explainable Self-Blended (LESB) DeepFake detector that in addition to yielding highly competitive results when compared to the state-of-the-art, also offers insights into the decision process by highlighting specific facial regions that contribute the most the detection decision. The detector was evaluated across multiple datasets and was shown to exhibit better transparency than competing models when making decisions.

Ethics Statement. Our research on deepfake detection prioritizes ethical responsibility by enhancing transparency and interpretability through explainable artificial intelligence (xAI), fostering trust in AI systems for combating misinformation. However, we acknowledge the potential negative ethical impacts, such as the misuse of explainable deepfake detection methods to evade detection, and emphasize the need for continuous monitoring and responsible deployment to mitigate such risks.

References

[1] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. MesoNet: a compact facial video forgery detection network. In *IEEE International Workshop on Information Forensics and Security*, dec 2018. 2

[2] Samet Akcay, Amir Atapour-Abarghouei, and Toby P Breckon. Ganomaly: Semi-supervised anomaly detection via adversarial training. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III 14*, pages 622–637. Springer, 2019. 2, 6

[3] Irene Amerini, Leonardo Galteri, Roberto Caldelli, and Alberto Del Bimbo. Deepfake video detection through optical flow based cnn. In *CVPR-W*, 2019. 2

[4] Merel de Leeuw den Bouter, Javier Lloret Pardo, Zeno Geradts, and Marcel Worring. Protoexplorer: Interpretable forensic analysis of deepfake videos using prototype exploration and refinement. *Information Visualization*, 23(3):239–257, 2024. 3

[5] Fadi Boutros, Vitomir Struc, Julian Fierrez, and Naser Damer. Synthetic data for face recognition: Current state and future prospects. *Image and Vision Computing*, 135:104688, 2023. 1

[6] Aidan Boyd, Patrick Tinsley, Kevin W Bowyer, and Adam Czajka. Cyborg: Blending human saliency into the loss improves deep learning-based synthetic face detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6108–6117, 2023. 3

[7] Marko Brodarič, Vitomir Štruc, and Peter Peer. Cross-dataset deepfake detection: evaluating the generalization capabilities of modern deepfake detectors. In *Proceedings of the 27th Computer Vision Winter Workshop (CVWW 2024)*, pages 47–56, 2024. 6

[8] Giuseppe Cartella, Vittorio Cuculo, Marcella Cornia, Marco Papasidero, Federico Ruozzi, Rita Cucchiara, et al. Pixels of faith: Exploiting visual saliency to detect religious image manipulation. In *Proceedings of the European Conference on Computer Vision Workshops*, 2024. 3

[9] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 839–847. IEEE, 2018. 7

[10] Liang Chen, Yong Zhang, Yibing Song, Lingqiao Liu, and Jue Wang. Self-supervised learning of adversarial example: Towards good generalizations for deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18710–18719, 2022. 2, 6

[11] Shen Chen, Taiping Yao, Yang Chen, Shouhong Ding, Jilin Li, and Rongrong Ji. Local relation learning for face forgery detection. In *AAAI*, volume 35, pages 1081–1088, 2021. 6, 7

[12] CodeNiko. 81 facial landmarks shape predictor. `https://github.com/codeniko/shape_predictor_81_face_landmarks`. Accessed: 2021-11-13. 4

[13] A. Das and L. Sebastian. A comparative analysis and study of a fast parallel cnn based deepfake video detection model with feature selection (fpcdfm). In *2023 Advanced Computing and Communication Technologies for High Performance Applications (ACCTHPA)*, pages 1–9. IEEE, 2023. 2

[14] Sowmen Das, Selim Seferbekov, Arup Datta, Md Islam, Md Amin, et al. Towards solving the deepfake problem: An analysis on improving deepfake detection using dynamic face augmentation. In *ICCV*, pages 3776–3785, 2021. 2

[15] Jiankang Deng, Jia Guo, Yandong Wen, Zhifeng Li, and Wei Liu. Retinaface: Single-stage dense face localization in the wild. *arXiv preprint arXiv:1905.00641*, 2019. 4

[16] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) dataset. *arXiv preprint arXiv:2006.07397*, 2020. 2, 6

[17] Rudresh Dwivedi, Devam Dave, Het Naik, Smiti Singhal, Rana Omer, Pankesh Patel, Bin Qian, Zhenyu Wen, Tejal Shah, Graham Morgan, et al. Explainable ai (xai): Core ideas, techniques, and solutions. *ACM Computing Surveys*, 55(9):1–33, 2023. 2

[18] European Union. General data protection regulation (gdpr). `https://gdpr-info.eu/`. Accessed: August 29, 2024. 2

[19] Jiazhi Guan, Hang Zhou, Mingming Gong, Youjian Zhao, Errui Ding, and Jingdong Wang. Detecting deepfake by creating spatio-temporal regularity disruption. *arXiv preprint arXiv:2207.10402*, 2022. 2

[20] Hui Guo, Shu Hu, Xin Wang, Ming-Ching Chang, and Siwei Lyu. Eyes tell all: Irregular pupil shapes reveal gan-generated faces. In *ICASSP*, pages 2904–2908, 2022. 2

[21] Zhiqing Guo, Zhenhong Jia, Liejun Wang, Dewang Wang, Gaobo Yang, and Nikola Kasabov. Constructing new backbone networks via space-frequency interactive convolution for deepfake detection. *IEEE Transactions on Information Forensics and Security*, 2023. 2

[22] Alexandros Haliassos, Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Lips don't lie: A generalisable and robust approach to face forgery detection. In *CVPR*, pages 5039–5049, 2021. 6, 7

[23] Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. Using self-supervised learning can improve model robustness and uncertainty. *Advances in neural information processing systems*, 32, 2019. 2

[24] Matthew Q Hill, Connor J Parde, Carlos D Castillo, Y Ivette Colon, Rajeev Ranjan, Jun-Cheng Chen, Volker Blanz, and Alice J O'Toole. Deep convolutional neural networks in the face of caricature. *Nature Machine Intelligence*, 1(11):522–529, 2019. 3

[25] Chih-Chung Hsu, Yu-Xiang Zhuang, and Chia-Yi Lee. Deep fake image detection based on pairwise learning. *Applied Sciences*, 10(2):370, 2020. 1, 2

[26] N. S. Ivanov, A. V. Arzhskov, and V. G. Ivanenko. Combining deep learning and super-resolution algorithms for deep fake detection. In *IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus)*, pages 326–328. IEEE, 2020. 1, 2

[27] Marija Ivanovska and Vitomir Štruc. A comparative study on discriminative and one–class learning models for deepfake detection. 2021. 2, 6

[28] Marija Ivanovska and Vitomir Struc. On the vulnerability of deepfake detectors to attacks generated by denoising diffusion models. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1051–1060, 2024. 1

[29] Marija Ivanovska and Vitomir Štruc. Y-gan: Learning dual data representations for anomaly detection in images. *Expert Systems with Applications*, 248:123410, 2024. 6

[30] Loic Jezequel, Ngoc-Son Vu, Jean Beaudet, and Aymeric Histace. Efficient anomaly detection using self-supervised multi-cue tasks. *IEEE Transactions on Image Processing*, 32:807–821, 2023. 2

[31] Thrupthi Ann John, Vineeth N Balasubramanian, and CV Jawahar. Canonical saliency maps: Decoding deep face models. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 3(4):561–572, 2021. 3

[32] Janez Križaj, Richard O Plesh, Mahesh Banavar, Stephanie Schuckers, and Vitomir Štruc. Deep face decoder: towards understanding the embedding space of convolutional networks through visual reconstruction of deep face templates. *Engineering applications of artificial intelligence*, 132:107941, 2024. 3

[33] Nicolas Larue, Ngoc-Son Vu, Vitomir Struc, Peter Peer, and Vassilis Christophides. Seeable: Soft discrepancies and bounded contrastive learning for exposing deepfakes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21011–21021, 2023. 2, 3, 6

[34] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face x-ray for more general face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5001–5010, 2020. 2, 3

[35] Yuezun Li and Siwei Lyu. Exposing deepfake videos by detecting face warping artifacts. In *CVPR Workshops*, pages 1–6, 2019. 2, 6, 7

[36] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3207–3216, 2020. 2, 6

[37] Chen Liang, Zhang Yong, Song Yibing, Jue Wang, and Lingqiao Liu. Ost: Improving generalization of deepfake detection via one-shot test-time training. In *NeurIPS*, 2022. 2, 6

[38] Honggu Liu, Xiaodan Li, Wenbo Zhou, Yuefeng Chen, Yuan He, Hui Xue, Weiming Zhang, and Nenghai Yu. Spatial-phase shallow learning: Rethinking face forgery detection in frequency domain. In *CVPR*. IEEE, 2021. 2

[39] Yuchen Luo, Yong Zhang, Junchi Yan, and Wei Liu. Generalizing face forgery detection with high-frequency features. In *CVPR*, pages 16317–16326, 2021. 6, 7

[40] Badhrinarayan Malolan, Ankit Parekh, and Faruk Kazi. Explainable deep-fake detection using visual interpretability methods. In *2020 3rd International conference on Information and Computer Technologies (ICICT)*, pages 289–293. IEEE, 2020. 3

[41] Iacopo Masi, Aditya Killekar, Royston Marian Mascarenhas, Shenoy Pratik Gurudatt, and Wael AbdAlmageed. Two-branch recurrent network for isolating deepfakes in videos. In *ECCV*, pages 667–684, 2020. 6, 7

[42] Blaž Meden, Peter Rot, Philipp Terhörst, Naser Damer, Arjan Kuijper, Walter J Scheirer, Arun Ross, Peter Peer, and Vitomir Štruc. Privacy–enhancing face biometrics: A comprehensive survey. *IEEE Transactions on Information Forensics and Security*, 16:4147–4183, 2021. 1

[43] Domingo Mery and Bernardita Morris. On black-box explanation for face verification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3418–3427, 2022. 3

[44] Rami Mubarak, Tariq Alsboui, Omar Alshaikh, Isa Inuwa-Dute, Saad Khan, and Simon Parkinson. A survey on the detection and impacts of deepfakes in visual, audio, and textual formats. *IEEE Access*, 2023. 1, 2, 3

[45] Poojan Oza and Vishal M Patel. One-class convolutional neural network. *IEEE Signal Processing Letters*, 26(2):277–281, 2018. 2, 6

[46] Alice J O'Toole, Carlos D Castillo, Connor J Parde, Matthew Q Hill, and Rama Chellappa. Face space representations in deep convolutional neural networks. *Trends in cognitive sciences*, 22(9):794–809, 2018. 3

[47] Connor J Parde, Y Ivette Colón, Matthew Q Hill, Carlos D Castillo, Prithviraj Dhar, and Alice J O'Toole. Closing the gap between single-unit and neural population codes: Insights from deep learning in face recognition. *Journal of vision*, 21(8):15–15, 2021. 3

[48] Martin Pernuš, Vitomir Štruc, and Simon Dobrišek. Maskfacegan: High resolution face editing with masked gan latent code optimization. *IEEE Transactions on Image Processing*, 2023. 1

[49] Richard Plesh, Janez Krizaj, Keivan Bahmani, Mahesh Banavar, Vitomir Štruc, and Stephanie Schuckers. Discovering interpretable feature directions in the embedding space of face recognition models, 2024. 2

[50] Richard Plesh, Janez Križaj, Keivan Bahmani, Mahesh Banavar, Vitomir Štruc, and Stephanie Schuckers. Discovering interpretable feature directions in the embedding space of face recognition models. In *IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10, 2024. 3

[51] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *ECCV*, 2020. 2

[52] Mansi Rehaan, Nirmal Kaur, and Staffy Kingra. Face manipulated deepfake generation and recognition approaches: A survey. *Smart Science*, 12(1):53–73, 2024. 1

[53] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics: A large-scale video dataset for forgery detection in human faces. *arXiv preprint arXiv:1803.09179*, 2018. 2, 6

[54] Waddah Saeed and Christian Omlin. Explainable ai (xai): A systematic meta-survey of current challenges and future opportunities. *Knowledge-Based Systems*, 263:110273, 2023. 2

[55] Alper Sengur et al. Investigation of comparison on modified cnn techniques to classify fake face in deepfake videos. *Journal of Computer and Communications*, 6(6):21–33, 2018. 2

[56] Jia Wen Seow, Mei Kuan Lim, Raphaël CW Phan, and Joseph K Liu. A comprehensive overview of deepfake: Generation, detection, datasets, and opportunities. *Neurocomputing*, 513:351–371, 2022. 2

[57] K. Shiohara and T. Yamasaki. Detecting deepfakes with self-blended images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1816–1826, 2022. 2, 3, 5, 6, 7

[58] Jihoon Tack, Sangwoo Mo, Jongheon Jeong, and Jinwoo Shin. Csi: Novelty detection via contrastive learning on distributionally shifted instances. *Advances in neural information processing systems*, 33:11839–11852, 2020. 2

[59] S. Tariq, S. Lee, H. Kim, Y. Shin, and S. S. Woo. Detecting both machine and human created fake face images in the wild. In *Proceedings of the 2nd International Workshop on Multimedia Privacy and Security*, 2018. 2

[60] Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, and Javier Ortega-Garcia. Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion*, 64:131–148, 2020. 1, 2

[61] Pengxiang Xu, Zhiyuan Ma, Xue Mei, and Jie Shen. Detecting facial manipulated images via one-class domain generalization. *Multimedia Systems*, 30(33), 2024. 2

[62] Xin Yang, Yuezun Li, and Siwei Lyu. Exposing deep fakes using inconsistent head poses. In *ICASSP*, 2019. 2

[63] P. Zhang, L. Yang, and D. Li. Efficientnet-b4-ranger: A novel method for greenhouse cucumber disease recognition under natural complex environment. *Computers and Electronics in Agriculture*, 176:105652, 2020. 6

[64] Tianchen Zhao, Xiang Xu, Mingze Xu, Hui Ding, Yuanjun Xiong, and Wei Xia. Learning self-consistency for deepfake detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15023–15033, 2021. 2, 6, 7

[65] Yinglin Zheng, Jianmin Bao, Dong Chen, Ming Zeng, and Fang Wen. Exploring temporal coherence for more general video face forgery detection. In *ICCV*, pages 15044–15054, 2021. 6, 7

[66] Tianfei Zhou, Wenguan Wang, Zhiyuan Liang, and Jianbing Shen. Face forensics in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5778–5788, June 2021. 2, 6

[67] Tianfei Zhou, Wenguan Wang, Zhiyuan Liang, and Jianbing Shen. Face forensics in the wild. In *CVPR*, 2021. 6, 7