Diffusion Deepfake

Anonymous CVPR submission

Paper ID *****

Abstract

Recent progress in generative AI, primarily through diffusion models, presents significant challenges for real-world deepfake detection. The increased realism in image details, diverse content, and widespread accessibility to the general public complicates the identification of these sophisticated deepfakes. Acknowledging the urgency to address the vulnerability of current deepfake detectors to this evolving threat, our paper introduces two extensive deepfake datasets generated by state-of-the-art diffusion models as other datasets are less diverse and low in quality. Our extensive experiments also showed that our dataset is more challenging compared to the other face deepfake datasets. Our strategic dataset creation not only challenge the deepfake detectors but also sets a new benchmark for more evaluation. Our comprehensive evaluation reveals the struggle of existing detection methods, often optimized for specific image domains and manipulations, to effectively adapt to the intricate nature of diffusion deepfakes, limiting their practical utility. To address this critical issue, we investigate the impact of enhancing training data diversity on representative detection methods. This involves expanding the diversity of both manipulation techniques and image domains. Our findings underscore that increasing training data diversity results in improved generalizability. Moreover, we propose a novel momentum difficulty boosting strategy to tackle the additional challenge posed by training data heterogeneity. This strategy dynamically assigns appropriate sample weights based on learning difficulty, enhancing the model's adaptability to both easy and challenging samples. Extensive experiments on both existing and newly proposed benchmarks demonstrate that our model optimization approach surpasses prior alternatives significantly. Code and data will be available.

1. Introduction

As more aspects of human life move into the digital realm, advancements in deepfake technology, particularly in generative AI like diffusion models [54], have produced highly realistic images, especially faces, which are almost indistinguishable to untrained human eyes. The misuse of deepfake technology poses increasing risks, including misinformation, political manipulation, privacy breaches, fraud, and cyber threats [28].

Diffusion-based deepfakes differ significantly from earlier techniques in three main aspects. Firstly, they exhibit **high-quality** by generating face images with realistic details, eliminating defects like edge or smear effects, and correcting abnormal biometric features such as asymmetric eyes/ears. Secondly, diffusion models showcase **diversity** in their outputs, creating face images across various contexts and domains due to extensive training on large datasets like LAION-5B, containing billions of real-world photos from diverse online sources [41]. Lastly, the **accessibility** of diffusion-based deepfakes extends to users with varying skill levels, transforming the creation process from a highly skilled task to an easy procedure. Even amateurs can produce convincing forgeries by generative models e.g., Stablility Diffusion [3] and MidJourney [2].

The rapid progress in deepfake creation technologies, fueled by diffusion models, has outpaced deepfake detection research in adapting to emerging challenges. Firstly, the lack of dedicated deepfake datasets for state-of-the-art diffusion models is evident. Widely used datasets like FF++ [39] and CelebDF [26] were assembled years ago using outdated facial manipulation techniques. The absence of a standardized diffusion-based benchmark impedes comprehensive assessment of deepfake detection models.

Secondly, existing research on deepfake detection often neglects the crucial issue of generalization. Many studies operate in controlled environments, training models on specific domains and manipulations and subsequently testing them on images from the same source. However, this approach falters when confronted with diffusion-generated deepfake images that span diverse domains and contents. Recent studies [10, 53] highlight the struggle of deepfake detectors to generalize to unseen manipulations or unfamiliar domains. Attempts to tackle this challenge, such as domain adaptation or transfer learning [6], have yielded suboptimal performance.

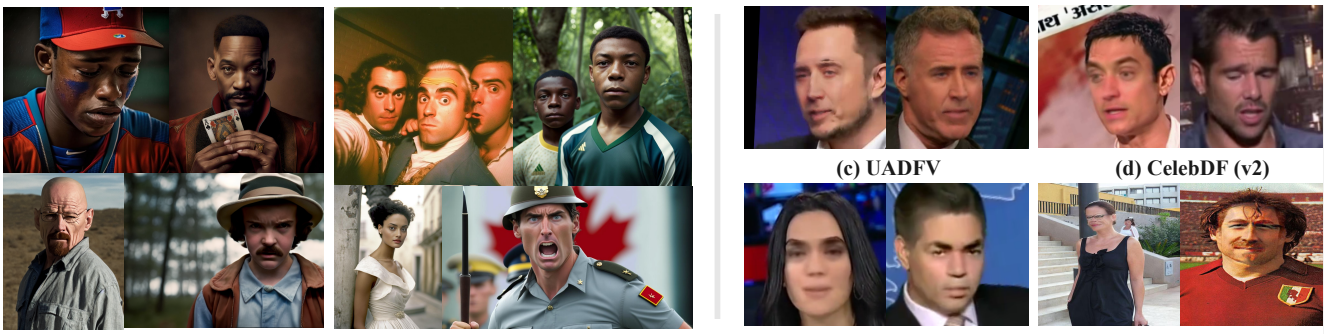

Figure 1. Our proposed diffusion deepfake datasets **(a-b)** are featured with more realistic and faithful facial details and diverse background contents compared to the previous **(c-f)**.

To address the identified problems, this paper presents two new deepfake detection benchmarks that utilize advanced diffusion models, namely ***DiffusionDB-Face*** and ***JourneyDB-Face***. These benchmarks encompass a wide range of content, incorporating diverse elements like head poses, facial attributes, photo styles, and realistic appearances. We expect these datasets to stimulate advancements in the identification of deepfakes generated through diffusion techniques. Our thorough assessment of these benchmarks indicates that the majority of current deepfake detectors, trained in constrained conditions, struggle to adapt to the evolving array of visual content generation methods, exemplified by diffusion models.

To enhance generalized deepfake detection, we advocate expanding the training data in terms of both scale and diversity. This approach is inspired by [32, 35] that underscores the effectiveness of employing simple objective functions on extensive and diverse image datasets to achieve robust visual representations. In our initial pursuit of generalized deepfake detection, we suggest training a detector on an inclusive dataset covering a broad spectrum of deepfake generation techniques and image domains.

Acknowledging the varying complexities associated with different types of deepfakes, ranging from basic graphics-based face swaps to more intricate samples generated by diffusion models, we propose a novel momentum difficulty boosting strategy. This involves dynamically assigning different weights to samples based on their difficulties, thereby facilitating the model's adaptability to both straightforward and challenging deepfake samples.

This work contributes: (1) **Novel benchmarks**: We introduce two large-scale benchmarks, namely ***DiffusionDB-Face*** and ***JourneyDB-Face***, for deepfake detection. These benchmarks, designed to align with the rapid progress in generative AI models, offer a substantially increased number of high-quality face images with more diversity of images along with additional text description metadata. This surpasses the capabilities of previous benchmarks, creating notable challenges for existing detection models. Table 2 summarises the comparison between the conventional datasets and our proposed dataset. (2) **Generalizability assessment**: We extensively evaluate the generalizability of existing deepfake detection models on our new benchmarks. Operating under a challenging cross-domain scenario, our analysis uncovers the undesirable sensitivity of current models to domain shifts. This sensitivity often leads to a significant decline in performance. (3) **A novel generic training strategy for generation heterogeneity**: We show that our *momentum difficulty boosting* on datasets featuring diverse sources of deepfake generation methods markedly improves deepfake detection performance.

2. Related Work

DeepFake Creation and Benchmarks The rise of deepfake technology poses a significant security threat, with the potential for misuse in spreading misinformation and engaging in malicious activities. In response, researchers are actively enhancing deepfake detection models to counter this threat. To evaluate these models, various datasets with diverse deepfake and authentic data from multiple sources have been established.

Earlier prominent deepfake datasets, such as FaceForensics++ [39], UADFV [55] and CelebDF [26], have been instrumental in this endeavor. The FaceForensics++ is created through four facial manipulation methods: FaceSwap [22], Face2Face [45], Deepfake [22] and NeuralTexture [46]. It also provides three compression levels to evaluate detectors under varying compression scenarios. UADFV creates fake face images by splicing face region synthesized using deep neural network into the original image. Nevertheless, these datasets exhibit low visual quality, markedly differing from Deepfake videos disseminated on the internet. Consequently, the CelebDF dataset focuses on achieving superior visual quality through an AutoEncoder-based deepfake synthesis method, including 590 real videos and 5639 synthetic

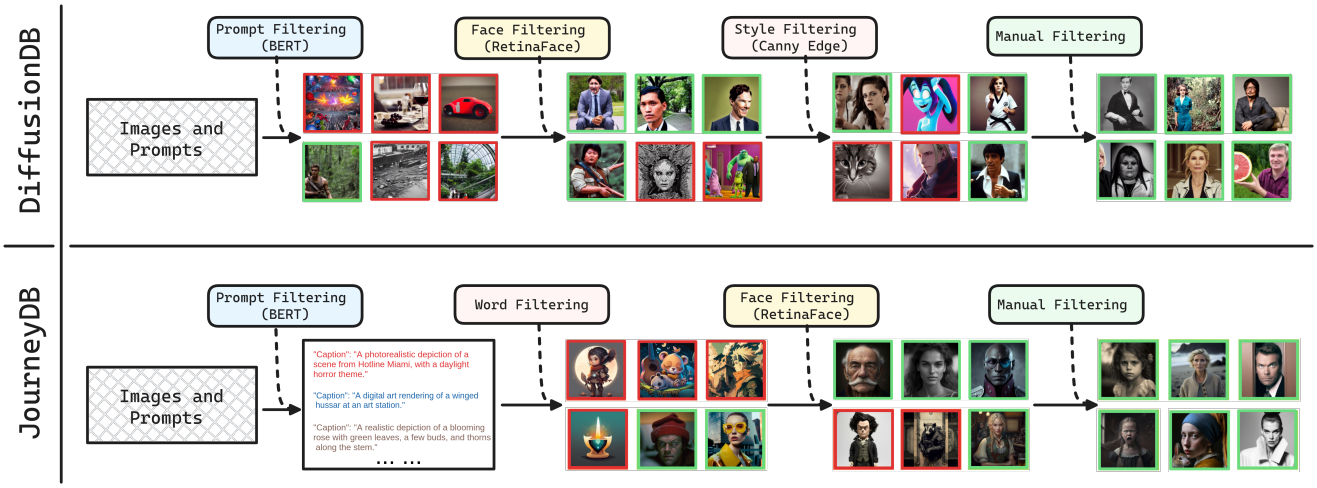

Figure 2. Collection process for the proposed DiffusionDB-Face and JourneyDB-Face datasets. Green border : Images that were kept for the following round; Red Border : Images that were deleted after filtering.

celebrity videos.

With the advancement of generative models, there has been a proliferation of highly realistic Deepfake videos produced by a multitude of GAN variants [14, 34]. However, GAN-based deepfake methods still face limitations, notably the absence of realistic backgrounds in the generated images [8, 29, 51].

Diffusion models [38] have gained widespread attention due to their ability to generate visually plausible content. Ricker *et al*. [37] demonstrated through extensive evaluation experiments that identifying images generated by diffusion models is a more challenging task than recognizing GAN-generated images. In contrast to GANs, deepfake images generated by diffusion models do not exhibit noticeable grid-like artifacts in the frequency domain. Song *et al*. [44] utilized diffusion models to create a synthetic celebrity face dataset, Deepfakeface. They similarly introduced two new tasks to enhance the assessment of detection methods performance. In parallel, we propose two deepfake datasets based on diffusion models: DiffusionDB and JourneyDB. Compared to Deepfakeface, our benchmarks cover a wider range of content and importantly exhibit more significant challenges to existing deepfake detectors (see Tables 4 and 6 in supplementary material). Table 2 summarises three conventional and three diffusion generated benchmark datasets (including our dataset). The table shows that our dataset is bigger than other datasets with more diversity per images and also contains metadata. Supplementary material has more samples from the dataset proving the *diversity* in our dataset which other dataset lacks. By incorporating cutting-edge diffusion models, the deepfake images in these datasets feature diverse elements like head poses, facial features, and image styles while exhibiting a realistic appearance. We expect these datasets to drive progress in detecting deepfake generated by diffusion models.

DeepFake Detection relies on analyzing different feature signals to ascertain the authenticity of an image. Earlier efforts focused on analyzing physiological signals for deepfake detection. Li *et al*. [25] identified the absence of eye blinking as a telltale sign for detecting deepfake videos and showed that distinguishing open and closed eye states could help. Additional efforts have explored features such as head poses [55], speaking-action patterns [4], and the combinations of various physiological signals [9].

Furthermore, many methods involving the search for potential synthetic artifacts and analysis of local features have been proposed. FWA [24] detects deepfakes by simulating facewarping artifacts. Face X-ray [23] predicts the presence of blending boundaries. Zhu *et al*. [57] introduced 3D decomposition into deepfake detection, amplifying subtle local artifacts through facial detail construction and detection. Recent research like DIRE [48] use image reconstruction error as a differentiating factor between real and fake images for detection. Frequency domain cues are also crucial for distinguishing deepfakes. Luo *et al*. [27] highlighted that CNN-based detectors tend to overfit to color textures in cross-database scenarios, suggesting the use of high-frequency noise for face forgery detection.

Data-driven approaches aim to directly learn how to differentiate real images from deepfakes through various strategies, exhibiting better generalization [16, 19, 30, 43, 47]. Capsule [30] pioneers the use of capsule networks in the deepfake detection task. Wang *et al*. [47] emphasized the importance of careful pre- and post-processing and data augmentation to enhance the generalization. Recently,

Guo *et al*. [19] proposed a hierarchical fine-grained formulation to address the diversity of images generated by various forgery methods. By encouraging the model to learn integrated features and inherent hierarchical properties of different forgery attributes, this approach improves deepfake detection representation.

In this work, we emphasize the importance of using heterogeneous training images for extended model generalization. Further, a novel model-agnostic momentum difficulty boosting strategy is introduced for more effective training by dynamically tuning the weights of individual samples during optimization.

3. Diffusion DeepFake Benchmarks

AI-generated content (AIGC) platforms like DALL-E, Stability AI, and Midjourney empower global users to craft detailed, high-quality images from text prompts. Several general large-scale prompt-to-image datasets, e.g. the MidjourneyDB [33] and DiffusionDB [49], have thus been collected by crawling from public sources (e.g. Stable Diffusion and Midjourney Discord servers). Our approach to constructing diffusion-based deepfake datasets involves iterative textual and visual filtering of these general prompt-image datasets. This curation process aims to refine prompts/images progressively, ensuring they exclusively feature high-quality human face images.

3.1. DiffusionDB-Face Construction

We initiated our dataset curation with the DiffusionDB(2M) dataset [49], comprising 2 million images generated by Stable Diffusion, each associated with prompts. To curate a deepfake dataset which only contains high-quality face images, we design an iterative approach with four steps of filtering following a coarse-to-fine progression:

(1) **Prompt Filtering by LLM**:

The goal of this step is to quickly reduce the candidate prompt pool such that only prompts related to human faces are retrieved. Inspired by the outstanding zero-shot capability of large-language models (LLM), we defined a zero-shot classification task to classify the associated prompt of each image into two predefined categories ("human face", "not human face") with the HuggingFace Transformer toolbox [50]. We used a pre-trained language model (BERT-base [13] with 12 transformer blocks, 12 attention heads, 110M parameters) to classify each prompt in the original DiffusionDB to obtain a prediction score for the pre-defined class of "human face" (see Figure 4). We set a threshold value of 0.5 and discarded all the prompts whose prediction were smaller than the thershold. With this approach we successfully removed 95 + % of the original prompts not semantically related to human faces.

(2) **Detection based auto filtering**: With 84,830 prompts remaining after the first step, we employed the state-of-the-art RetinaFace detector [12] on all associated images to selectively retain those featuring human faces. In this phase, we utilized the RetinaFace model with its default configuration, setting the confidence threshold at 0.5. Images where the confidence score meets or exceeds this threshold are retained for subsequent filtering stages. Utilizing RetinaFace, we successfully extracted most images containing faces (see Figure 5b).

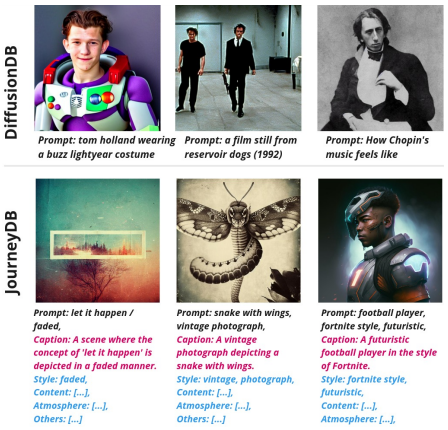

Figure 3. Input metadata along with the corresponding images.

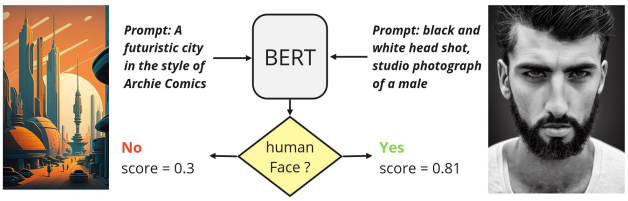

Figure 4. Example of prompt filtering by language model. Note, only text is the input to BERT [13], whilst the associated image is shown for illustration only.

(a) Examples of word filtering for removing anime style images.

(b) Examples of detection based auto filtering using RetinaFace. According to BERT many prompts are highly related to *human faces* but their associated images do not contain realistic human faces

Figure 5. Illustrations of prompt filtering.

Having obtained 39,887 human face images, a quick manual inspection revealed various unrealistic images with distinct artistic styles (e.g., black-and-white / anime / car-

toon / sketch-style faces).

(3) **Edge/color based style filtering**: We adopted two additional steps to further refine our data. (I) We measure the color variance of the original image to identify whether the input image has a too narrow color spectrum; (II) We apply a Canny edge detector to measure the number of edges on the images to identify images with specific drawing styles and animations. Empirically we set the edge threshold at 100 and the color threshold at 200 to determine whether an image is with unrealistic style, and excluded images if their edges exceeded the edge threshold or color variance falls below the color threshold. This step helped to reduce about 50% images from the last round.

(4) **Manual filtering**: In the final step, we conducted a manual annotation process, resulting in a curated dataset of 18,371 high quality realistic human faces, which we refer as DiffusionDB-Face.

3.2. JourneyDB-Face Construction

To retrieve face images from JourneyDB [33] suitable for deepfake detection, we followed the same procedure as in Sec 3.1, with three minor adjustments. (1) To ensure we have enough test images in our deepfake detection benchmark, we ignored the original train / validation / test split provided by JourneyDB.

(2) Since the metadata of JourneyDB also include style prompts (see Figure 3), we thus replaced the edge/color-based style filtering step as in DiffusionDB to an exclusive word filtering on the style prompts to remove images with unrealistic styles such as "Anime Style", e.g. see Figure 5a.

(3) The test partition of the JourneyDB dataset does not come with any metadata, so we direcly applied RetinaFace detector followed by a manual filtering process.

3.3. Data Preprocessing

The basic statistics of DiffusionDB-Face and Journey-Face are shown in Figure 7. We used the Deepface [42] framework to analyze the gender distribution statistics within our dataset. After the acquisition of the datasets, a comprehensive preprocessing pipeline was executed to optimize the data for utilization in deep learning architectures and to facilitate ease of analytical operations. Specifically, we performed a re-examination of each image for facial detection by MTCNN [56]. Some images were further discarded at this stage due to face detection failures or only containing too small faces without enough visual details for the deepfake detection task.

After face detection, the images were uniformly cropped to a resolution of 256×256 pixels, establishing a standard input size.

Finally, the preprocessed dataset with standardized face detection crops has 24,794 and 87,833 deepfake images for propsoed DiffusionDB-Face (DFDB-Face) and JourneyDB-Face (JDB-Face) benchmark respectively. Subsequently, these images were categorized into train / test / val subsets with a $90:5:5$ ratio respectively as shown in Table 3.

Additionally, to evaluate the full classification performance, we have sourced 94,120 authentic face images from the Flickr-Faces-HQ (FFHQ) dataset [21] so that we can measure both the sensitivity and specificity of the deepfake detection methods.

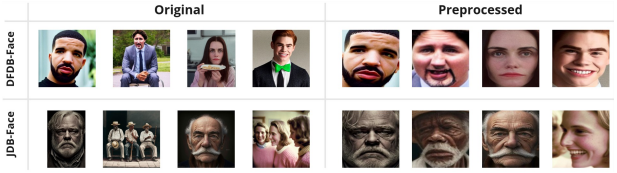

Figure 6. Visualization before and after preprocessing the images.

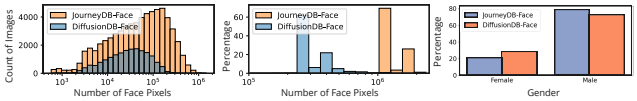

Figure 7. Basic statistics of our datasets.

4. Momentum Difficulty Boosting

The conventional deepfake detection training and evaluation protocol tends to overlook the critical issue of generalization, often yielding inflated detection performance. Specifically, a detector may exhibit impressive results when trained and tested on deepfakes generated from the same source, within a limited range of manipulations and image domains. However, as observed in [11, 53] and corroborated by our subsequent evaluations, these detectors experience a substantial performance drop when applied to deepfakes from different sources/domains. This challenge is particularly pronounced, as demonstrated in Sec 5.1, when detectors are applied to the diverse diffusion-generated deepfakes. To address this limitation, we advocate for a new setting, where the performance of a detector should be benchmarked against multi-source training and test datasets, providing a more comprehensive understanding of its generalizability across various domains.

We begin with a set of K diverse deepfake datasets $\{\mathcal{D}^1, \mathcal{D}^2, \cdots, \mathcal{D}^K\}$. Each dataset $\mathcal{D}^k = \{(\mathbf{x}_i^k, y_i^k)\}_{i=1}^{N_k}, k \in [K]$, comprises N_k images sourced from specific domains and deepfake manipulation methods. For instance, one dataset may include Instagram-style selfies with deepfakes generated using diffusion models. We further denote f_θ the target deepfake detection model parameterized by θ, $\hat{y}_i^k = f_\theta(\mathbf{x}_i^k)$ the model prediction, and $\ell(y_i^k, f_\theta(\mathbf{x}_i^k))$ a general loss function in the context of deepfake detection, e.g. a standard binary

Table 1. Number of images after each round of processing: DiffusionDB and JourneyDB

Dataset	INPUT	Round 1	Round 2	Round 3	Round 4	Preprocessed (Final)
DiffusionDB-Face	2,000,000	84,830	39,887	18,845	15,198	24,794
JourneyDB-Face	4,932,309	238,869	225,759	78,904	61,984	87,833

Table 2. Dataset summary. Top: Conventional datasets; Bottom: Diffusion datasets; V: Video datasets. MF/S : Multiple faces per sample ; Generation: Generation methods.

Source	No. Fake	No. Real	Generation	Metadata	MF/S
FF++ [39] (V)	4,000	977	F2F [45],DF [22],FS [22],NT [46]		
UADFV [55] (V)	49	49	FS [22], DF [22]		
CelebDFv2 [26] (V)	5,639	590	Autoencoder [26]		
DeepFakeFace [44]	3×30,000	30,000	SD [3], IP [3], IF [1]		
JDB-Face (ours)	**87,833**	**94,120**	Midjourney [2]		
DFDB-Face (ours)	**24,794**	**94,120**	SD [3]		

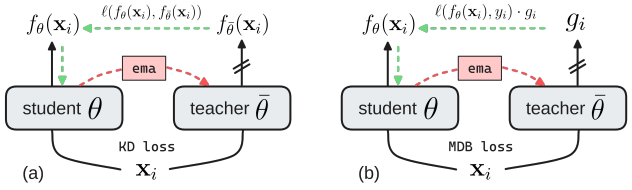

Figure 8. (a) Momentum-based knowledge distillation [7]; (b) Our MDS: only use the 'teacher' network to weight samples by their difficulties.

cross-entropy loss or more advanced loss designs as in [31, 52].

Conventional Setting Existing methods [18, 19, 37] often train deepfake detection models individually on each $\mathcal{D}^k$, and evaluate each trained model using the corresponding test set $\bar{\mathcal{D}}^k$, where images are sampled from the same source. Formally, they attempt to optimize the objective:

$$\min_{\theta_k} \mathbb{E}_{x_i^k \in \mathcal{D}^k} \left[\ell(y_i^k, f_{\theta_k}(\mathbf{x}_i^k)\right] + \lambda \mathbf{R}(\theta_k), \tag{1}$$

where the first and second term correspond to the empirical loss on $\mathcal{D}^k$ and the regularization term, respectively. However, this approach makes the unrealistic assumption that the image domains and manipulation methods are known during deployment, suffering from significant performance drop when facing domain and forgery type shifts.

Proposed Setting Instead of employing domain and manipulation-specific models, our objective is to train a single model f_θ agnostic to data source k. We combine images from $\{\mathcal{D}^k\}_{k=1}^K$ into a heterogeneous dataset denoted as $\mathcal{D}_H = \{(\mathbf{x}_i, y_i, k_i)\}_{i=1}^{\sum_k |\mathcal{D}_k|}$. As shown in our later experiments in Sec 5.1, directly training on such a mixed dataset did not translate to good cross-dataset performance, due to additional challenges imposed by diverse training samples with various level of difficulty.

Momentum Difficulty Boosting We thus propose to employ a boosting function to ease the training with data heterogeneity. This function regulates the importance of examples, assigning more weights to the difficult ones. Specifically, $g_i = g(\mathbf{x}_i, y_i, \theta)$ quantifies the instantaneous instance difficulty of sample $\mathbf{x}_i$, considering the under-optimized model parameters θ. Our revised optimization objective thus becomes

$$\min_{\theta} \mathbb{E}_{\mathbf{x}_i \in \mathcal{D}_H} \left[g_i \times \ell(y_i, f_\theta(\mathbf{x}_i)\right] + \lambda \mathbf{R}(\theta). \tag{2}$$

We proposed a simple yet effective strategy, momentum difficulty boosting (MDB), to calculate the sample-wise difficulty scores. Specifically, we maintain a momentum moving-average of the detector, $\bar\theta$, and use it to calculate sample difficulties on-the-fly by measuring the cross-entropy between the momentum network's prediction and the data samples' ground truths. Formally, we define the sample-wise difficulty score as

$$g(\mathbf{x}_i, y_i, \theta) = CE(y_i, f_{\bar\theta}(\mathbf{x}_i)), \tag{3}$$

where $\bar\theta$ slowly tracks the detector's parameter θ by the momentum updating rule: $\bar\theta = m\bar\theta + (1-m)\theta$.

The momentum update's benefit lies in mitigating the substantial variance in predicted sample difficulty scores, thereby enhancing training stability. Our approach shares conceptual similarities with knowledge distillation [5, 7, 20], with the difference that instead of directly distilling knowledge from a teacher network $\bar\theta$, we leverage it as a guiding function to adjust the training data distribution by assigning different weight to each sample based on their difficulty levels. During training, the sample weights g_i are decided by $\bar\theta$ based on in Eq. (3), where both the weights and $\bar\theta$ are updated dynamically at each mini-batch (see Figure 8). To prevent domination of certain samples with exceptionally high difficulty scores, we re-scale the sample weights

Table 3. Data split per dataset.

Dataset	Train		Test		Validation	
	Real	Fake	Real	Fake	Real	Fake
CelebDF V2 [26]	35,469	160,595	1,971	8,922	1,971	8,922
FF ++ [39]	17,847	102,755	990	5,711	993	5,711
UADFV [55]	1,393	1,371	77	78	76	77
Deepfakeface [44]	27,000	81,00	1,500	4,500	1,500	4,500
JDB-Face (ours)	82,440	78,757	4,581	4,375	4,580	4,376
DFDB-Face (ours)	82,440	22,331	4,581	1,241	4,580	1,241

in each mini-batch to fall within the range $[1, C]$, where C denotes the capped maximum sample weight.

5. Experiments

5.1. Evaluation of off-the-shelf models

We first produce a comprehensive evaluation of a range of existing pre-trained deepfake detectors on the generalization capability to understand how their performance degrade when tested on deepfake images from different sources/domains than training, especially on the newly collected diffusion-based deepfakes from our DiffusionDB-Face and JourneyDB-Face datasets.

Datasets We consider three conventional datasets and three diffusion-based datasets in our evaluation also summarized in Table 2. (1) *FaceForensics++ (FF++)* [39] consists of 1,000 video clips designed for digital forensics. It encompasses four facial modification techniques, including Face2Face [45], Deepfakes[22], FaceSwap[22], and NeuralTextures[46]. This dataset contains 977 YouTube videos, each featuring front-facing, easily trackable faces. (2) *CelebDFv2* [26], includes genuine YouTube videos and synthesized deepfake videos. In its first version, there are 408 genuine videos and 795 deepfake videos, covering diverse characteristics like ethnicity, age, and gender. The second version extends the dataset with 590 genuine videos and 5,639 deepfake videos obtained from online sources, further increasing data diversity. (3) *UADFV* [25] includes 49 genuine videos collected from the internet and then manipulated by [22] to generate deepfakes. (4) *Deepfakeface* [44] includes 90,000 fake images from three different generation methods i.e. StableDiffusionv1.5 [3], Inpainting [3] and InsightFace [1], along with 30,000 real images. (5-6) Our *DiffusionDB-Face* and *JourneyDB-Face* include various diffusion-based deepfakes generated by two art generative AI providers, Stability AI and MidJourney. The dataset collection process are detailed in Sec 3.1 and 3.2. (7) Fake-CelebA [48] was formed using four diffusion generation methods (a) SD-v2 [38] (42,000 images), (b) IF [40] (1,000 images), (c) DALLE-2 [36] (500 images), (d) Midjourney [2] (100 images), along with 42,000 real images. Train/test/validaiton split of the datasets is summarised in Table 3.

Competitors We consider seven pre-trained deepfake detection models. Specifically, (1) *HiFi Net* [19] is a fine-grained deepfake detector based on multi-branch feature extraction and hierarchical forgery predictions, trained on a customised dataset with a taxonomy of image forgery types ranging from CNN-based manipulations to image editing. (2) *SBIs* [43] is trained on FF++ with a novel image blending method reproducing common forgery artifacts, e.g., blending boundaries and statistical inconsistencies. (3) *CADDM* [15] is trained on FF++ with a constraint to mitigate the effect of identity leakage whilst performing deepfake detection. We used its EfficientNet-b4 variant in our evaluation. (4) *CNNDet* [47] trains a ResNet50 model on a customized dataset of deepfakes solely generated by ProGAN [17]. (5) *DSP-FWA* [24] is a deepfake detector specifically aiming to detect the warping artifacts of the deepfake creation process, trained with real images collected from Internet and a customized algorithm to generate negative data with warping effects. We used its SPP-Net varriant in our evaluation. (6) *Capsule* [30] is a Capsule network-based deepfake detector trained with the FF++ dataset. (7) *DIRE* [48] is a diffusion model generated deepfake detection method where a novel image representation is introduced to measure the error between input images.

Setting We followed the evaluation protocol proposed in [53]. For FF++ dataset, we have considered it as a unified dataset rather than separating it into four different parts with seperate manipulations. All the images from each dataset were preprocessed and cropped into a size of 256×256. The video datasets were sampled into frames, i.e. we took 32 frames per video after detecting the frames that included faces. Specifically, 19,830/114,213 (real/fake) video frames are sampled for FF++, 1,548/1524 for UADFV and 39,411/178,439 for CelebDFv2. All the listed deepfake detectors were evaluated with their officially released pre-trained weights and directly applied to the test splits of each dataset without further finetuning. We adopt three metrics for evaluation, including AUC (area under the ROC curve), EER (equal error rate), and ACC (accuracy).

Results As shown in Table 4, we have made the following

Table 4. Evaluation performance of off-the-shelf DeepFake detectors on conventional deepfake datasets (FF++, CelebDFv2, UADFV) and diffusion deepfake datasets (Deepfakeface, DFDB-Face, JDB-Face, Fake CelebA). Highest accuracy in **bold**.

(a) Conventional deepfake datasets (FF++, CelebDFv2, UADFV)

Model	**FF++**			**CelebDFv2**			**UADFV**		
Metric	AUC	EER	ACC	AUC	EER	ACC	AUC	EER	ACC
HiFi Net	0.60	0.41	**0.58**	0.60	0.41	**0.58**	0.60	0.45	0.54
SBIs	0.58	0.43	0.56	0.51	0.72	**0.67**	0.51	0.74	0.50
CADDM	0.50	0.48	0.52	0.50	0.50	0.50	0.56	0.46	**0.53**
CNNDet	0.76	0.29	**0.71**	0.54	0.46	0.53	0.53	0.41	0.58
DSP-FWA	0.54	0.61	0.33	0.66	0.40	0.51	0.48	0.51	0.48
Capsule	0.80	0.26	**0.73**	0.61	0.43	0.56	0.79	0.29	0.71
DIRE	0.11	0.91	0.22	0.14	0.90	0.21	0.22	0.85	0.27

(b) Diffusion deepfake datasets (Deepfakeface, DFDB-Face, JDB-Face, Fake CelebA)

Model	**FF++**			**CelebDFv2**			**UADFV**			**Fake CelebA**		
Metric	AUC	EER	ACC	AUC	EER	ACC	AUC	EER	ACC	AUC	EER	ACC
HiFi Net	0.57	0.45	0.45	0.52	0.66	0.51	0.45	0.68	0.40	0.51	0.55	0.49
SBIs	0.51	0.61	0.50	0.25	0.89	0.30	0.41	0.82	0.49	0.57	0.45	0.54
CADDM	0.51	0.49	0.50	0.48	0.70	0.47	0.52	0.73	0.52	0.51	0.68	0.48
CNNDet	0.61	0.41	0.58	0.53	0.49	0.52	0.44	0.75	0.45	0.40	0.62	0.58
DSP-FWA	0.50	0.88	0.40	0.52	0.51	**0.54**	0.52	0.47	0.53	0.38	0.57	0.42
Capsule	0.49	0.49	0.50	0.48	0.57	0.46	0.45	0.56	0.46	0.49	0.68	0.50
DIRE	0.38	0.76	0.55	0.62	0.45	0.71	0.42	0.54	0.51	0.68	0.34	**0.72**

observations:

(1) All pre-trained detectors exhibit pronounced generalization issues when tested on deepfakes originating from different sources or domains. For instance, the Capsule model [30], trained on the FF++ dataset, achieved a high AUC of 0.80 on the same dataset. However, its AUC dropped to 0.61 on CelebDFv2 generated by a different deepfake method. On the three diffusion-based deepfake datasets, its performance further degraded, with AUC decreasing to 0.49, 0.48, and 0.45 for Deepfakeface, DiffusionDB-Face, and JourneyDB-Face, respectively. On the other hand DIRE [48] has performed comparatively better with Fake CelebA but still not upto the mark due to the absence of the same domain's dataset in the training. This observation strongly highlights the generalization issue of existing deepfake detectors, impeding their practical utility in real-world scenarios where deepfakes can emerge from diverse sources and domains.

(2) Among all datasets, the diffusion-based ones have proven to be the most challenging for existing deepfake detectors. This is evident in the substantial performance gap between the three conventional datasets and the diffusion ones. Notably, on the proposed DiffusionDB-Face and JourneyDB-Face, all examined detectors (except DIRE) obtain AUC values below 0.55, indicating even worse performance than random guessing. However, even with DIRE detector, we JDB-Face performed worst among all diffusion datasets with 51% accuracy. This suggests that highly realistic facial images generated by the latest diffusion models can easily confuse pretrained deepfake detectors, leading them to be frequently misclassified as real faces and thus remaining undetected.

More evaluations under varying strategies are given in ***Supplementary material.***

6. Conclusion

Diffusion models presents substantial challenges for real-world deepfake detection. This work addresses this urgency by introducing extensive diffusion deepfake datasets and highlighting the limitations of existing detection methods. Our dataset is not only challenging to detect but is highly diverse compared to the present face deepfake datasets. We emphasize the crucial role of enhancing training data diversity on generalizability. Our proposed momentum difficulty boosting strategy, effectively tackles the challenge posed by training data heterogeneity. Extensive experiments show that our approach achieves state-of-the-art performance, surpassing prior alternatives significantly. It has shown high testing accuracy on the totally unknown dataset proving its generalizing ability. This work not only identifies the challenges of diffusion models in deepfake detection but also provides practical solutions, paving the way for more robust and adaptable countermeasures against the evolving threat of latest deepfakes.

References

[1] Insightface. https://github.com/deepinsight/insightface. 6, 7

[2] Midjourney discord server. https://discord.com/invite/midjourney. 1, 6, 7

[3] Stability ai. https://stability.ai/. 1, 6, 7

[4] Shruti Agarwal, Hany Farid, Yuming Gu, Mingming He, Koki Nagano, and Hao Li. Protecting world leaders against deep fakes. In *CVPRW*, page 38, 2019. 3

[5] Sungsoo Ahn, Shell Xu Hu, Andreas Damianou, Neil D. Lawrence, and Zhenwen Dai. Variational information distillation for knowledge transfer. In *CVPR*, 2019. 6

[6] Shivangi Aneja and Matthias Nießner. Generalized zero and few-shot transfer for facial forgery detection. *arXiv preprint arXiv:2006.11863*, 2020. 1

[7] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, pages 9650–9660, 2021. 6

[8] Yunzhuo Chen, Nur Al Hasan Haldar, Naveed Akhtar, and Ajmal Mian. Text-image guided diffusion model for generating deepfake celebrity interactions. *arXiv preprint arXiv:2309.14751*, 2023. 3

[9] Umur Aybars Ciftci, Ilke Demir, and Lijun Yin. Fakecatcher: Detection of synthetic portrait videos using biological signals. *IEEE TPAMI*, 2020. 3

[10] Davide Cozzolino, Justus Thies, Andreas Rössler, Christian Riess, Matthias Nießner, and Luisa Verdoliva. Forensictransfer: Weakly-supervised domain adaptation for forgery detection. *arXiv preprint arXiv:1812.02510*, 2018. 1

[11] Davide Cozzolino, Andreas Rössler, Justus Thies, Matthias Nießner, and Luisa Verdoliva. Id-reveal: Identity-aware deepfake video detection. In *ICCV*, pages 15108–15117, 2021. 5

[12] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *CVPR*, pages 5203–5212, 2020. 4

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 4

[14] 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. 3

[15] Shichao Dong, Jin Wang, Renhe Ji, Jiajun Liang, Haoqiang Fan, and Zheng Ge. Implicit identity leakage: The stumbling block to improving deepfake detection generalization. In *CVPR*, pages 3994–4004, 2023. 7

[16] Shichao Dong, Jin Wang, Renhe Ji, Jiajun Liang, Haoqiang Fan, and Zheng Ge. Implicit identity leakage: The stumbling block to improving deepfake detection generalization. In *CVPR*, pages 3994–4004, 2023. 3

[17] Hongchang Gao, Jian Pei, and Heng Huang. Progan: Network embedding via proximity generative adversarial network. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1308–1316, 2019. 7

[18] Luca Guarnera, Oliver Giudice, and Sebastiano Battiato. Level up the deepfake detection: a method to effectively discriminate images generated by gan architectures and diffusion models. *arXiv preprint arXiv:2303.00608*, 2023. 6

[19] Xiao Guo, Xiaohong Liu, Zhiyuan Ren, Steven Grosz, Iacopo Masi, and Xiaoming Liu. Hierarchical fine-grained image forgery detection and localization. In *CVPR*, pages 3155–3165, 2023. 3, 4, 6, 7

[20] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 6

[21] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 5

[22] Iryna Korshunova, Wenzhe Shi, Joni Dambre, and Lucas Theis. Fast face-swap using convolutional neural networks. In *Proceedings of the IEEE international conference on computer vision*, pages 3677–3685, 2017. 2, 6, 7

[23] 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 *CVPR*, pages 5001–5010, 2020. 3

[24] Yuezun Li and Siwei Lyu. Exposing deepfake videos by detecting face warping artifacts. In *CVPRW*, 2019. 3, 7

[25] Yuezun Li, Ming-Ching Chang, and Siwei Lyu. In ictu oculi: Exposing ai created fake videos by detecting eye blinking. In *2018 IEEE International workshop on information forensics and security (WIFS)*, pages 1–7. IEEE, 2018. 3, 7

[26] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *CVPR*, 2020. 1, 2, 6, 7

[27] Yuchen Luo, Yong Zhang, Junchi Yan, and Wei Liu. Generalizing face forgery detection with high-frequency features. In *CVPR*, pages 16317–16326, 2021. 3

[28] Mekhail Mustak, Joni Salminen, Matti Mäntymäki, Arafat Rahman, and Yogesh K. Dwivedi. Deepfakes: Deceptions, mitigations, and opportunities. *Journal of Business Research*, 154:113368, 2023. 1

[29] Ryota Natsume, Tatsuya Yatagawa, and Shigeo Morishima. Rsgan: face swapping and editing using face and hair representation in latent spaces. In *ACM SIGGRAPH 2018 Posters*, pages 1–2. 2018. 3

[30] Huy H Nguyen, Junichi Yamagishi, and Isao Echizen. Use of a capsule network to detect fake images and videos. *arXiv preprint arXiv:1910.12467*, 2019. 3, 7, 8

[31] Yunsheng Ni, Depu Meng, Changqian Yu, Chengbin Quan, Dongchun Ren, and Youjian Zhao. Core: Consistent representation learning for face forgery detection. In *CVPRW*, pages 12–21, 2022. 6

[32] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2

[33] Junting Pan, Keqiang Sun, Yuying Ge, Hao Li, Haodong Duan, Xiaoshi Wu, Renrui Zhang, Aojun Zhou, Zipeng Qin, Yi Wang, et al. Journeydb: A benchmark for generative image understanding. *arXiv preprint arXiv:2307.00716*, 2023. 4, 5

[34] Haibo Qiu, Baosheng Yu, Dihong Gong, Zhifeng Li, Wei Liu, and Dacheng Tao. Synface: Face recognition with synthetic data. In *ICCV*, pages 10880–10890, 2021. 3

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

[36] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 7

[37] Jonas Ricker, Simon Damm, Thorsten Holz, and Asja Fischer. Towards the detection of diffusion model deepfakes. *arXiv preprint arXiv:2210.14571*, 2022. 3, 6

[38] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 3, 7

[39] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *ICCV*, pages 1–11, 2019. 1, 2, 6, 7

[40] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 7

[41] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models. pages 25278–25294. Curran Associates, Inc., 2022. 1

[42] Sefik Ilkin Serengil and Alper Ozpinar. Hyperextended lightface: A facial attribute analysis framework. In *2021 International Conference on Engineering and Emerging Technologies (ICEET)*, pages 1–4, 2021. 5

[43] Kaede Shiohara and Toshihiko Yamasaki. Detecting deepfakes with self-blended images. In *CVPR*, pages 18720–18729, 2022. 3, 7

[44] Haixu Song, Shiyu Huang, Yinpeng Dong, and Wei-Wei Tu. Robustness and generalizability of deepfake detection: A study with diffusion models. *arXiv preprint arXiv:2309.02218*, 2023. 3, 6, 7

[45] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *CVPR*, pages 2387–2395, 2016. 2, 6, 7

[46] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *Acm Transactions on Graphics (TOG)*, 38(4):1–12, 2019. 2, 6, 7

[47] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot... for now. In *CVPR*, pages 8695–8704, 2020. 3, 7

[48] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. Dire for diffusion-generated image detection. *arXiv preprint arXiv:2303.09295*, 2023. 3, 7, 8

[49] Zijie J. Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. DiffusionDB: A large-scale prompt gallery dataset for text-to-image generative models. *arXiv:2210.14896 [cs]*, 2022. 4

[50] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019. 4

[51] Wayne Wu, Yunxuan Zhang, Cheng Li, Chen Qian, and Chen Change Loy. Reenactgan: Learning to reenact faces via boundary transfer. In *ECCV*, pages 603–619, 2018. 3

[52] Zhiyuan Yan, Yong Zhang, Yanbo Fan, and Baoyuan Wu. Ucf: Uncovering common features for generalizable deepfake detection. *arXiv preprint arXiv:2304.13949*, 2023. 6

[53] Zhiyuan Yan, Yong Zhang, Xinhang Yuan, Siwei Lyu, and Baoyuan Wu. Deepfakebench: A comprehensive benchmark of deepfake detection. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. 1, 5, 7

[54] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 2022. 1

[55] Xin Yang, Yuezun Li, and Siwei Lyu. Exposing deep fakes using inconsistent head poses. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8261–8265. IEEE, 2019. 2, 3, 6, 7

[56] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23 (10):1499–1503, 2016. 5

[57] Xiangyu Zhu, Hao Wang, Hongyan Fei, Zhen Lei, and Stan Z Li. Face forgery detection by 3d decomposition. In *CVPR*, pages 2929–2939, 2021. 3