Towards Accurate and Fair Deepfake Detection

by

Sayantan Das

A thesis submitted to the
Department of Electrical and Computer Engineering
in conformity with the requirements for
the degree of Master of Applied Science

Queen's University
Kingston, Ontario, Canada
March 2024

Abstract

In this thesis, we tackle two problems in the field of deepfake detection. First, we present a novel approach for the detection of deepfake videos using a pair of vision transformers pre-trained by a self-supervised masked autoencoding setup. Our method consists of two distinct components, one of which focuses on learning spatial information from individual RGB frames of the video, while the other learns temporal consistency information from optical flow fields generated from consecutive frames. Unlike most approaches where pre-training is performed on a generic large corpus of images, we show that by pre-training on smaller face-related datasets, strong results can be obtained. We perform various experiments to evaluate the performance of our method on commonly used datasets. Our experiments show that our method sets a new state-of-the-art in various setups including cross-dataset generalization. Subsequently, we introduce FairAlign, a new method to reduce bias and improve fairness in deepfake detection by aligning conditional embedding distributions in a high-dimensional kernel space. Our approach reduces information related to sensitive attributes in the embedding space that could potentially bias the detection process, thus promoting fairness. FairAlign is a versatile plug-and-play loss term compatible with various deepfake detection networks and is capable of enhancing fairness without compromising detection performance. In addition to applying FairAlign for reducing

gender bias, we implement a systematic pipeline for the annotation of skin tones and promotion of fairness in deepfake detection related to this sensitive attribute. Finally, we perform the first comprehensive study toward quantifying and understanding the trade-off between fairness and accuracy in the context of deepfake detection. We use public deepfake datasets to evaluate our method. Through various experiments, we observe that FairAlign outperforms other bias-mitigating methods across various deepfake detection backbones for both gender and skin tone, setting a new state-of-the-art. Moreover, our fairness-accuracy trade-off analysis demonstrates that our approach demonstrates the best overall performance when considering effectiveness in both deepfake detection and reducing bias.

Acknowledgments

This remarkable journey, brimming with discoveries and challenges, would not have been possible without the unwavering support of many. First and foremost, I extend my deepest gratitude to my supervisor, Prof. Ali Etemad, whose persistent patience, insightful guidance, and encouraging presence have been cornerstones of my academic pursuits.

I also wish to express my sincere thanks to my labmates and friends, with a special mention of Mojtaba Kolahdouzi. Their support and camaraderie proved invaluable, offering solace and assistance during the most challenging times.

Above all, my profound appreciation goes to my family, the foundation of my support network. Their unyielding belief in me, particularly through the demanding phases of graduate studies, played a critical role in achieving this significant milestone.

Statement of Originality

The following work described is my own and I hereby certify the intellectual content of this thesis is the product of my own work. To the best of my knowledge, all references and contributions of other individuals have been acknowledged, cited, and sourced appropriately.

Contents

List of Tables

List of Figures

Glossary of Abbreviations

AUC Area Under the Curve. vii, ix, x, 23, 27–29, 31–33, 41, 43, 44, 48, 49

Celeb-A CelebFaces-Attributes. 5, 8, 24–26

CNN Convolutional Neural Network. 9, 11–13, 19

DF Deepfakes. vii, 24, 28

DPD Demographic Parity Difference. viii, 40, 41, 44–46, 48, 49

DPR Demographic Parity Ratio. viii, 40, 41, 44–46

FAUC Fairness-Area-Under-the-Curve. 16

FF Face2Face. vii, 24, 27–29, 31

FF++ FaceForensics++. vii–ix, 2, 5, 6, 8, 11, 15, 23–33, 35, 39, 42, 43, 45, 47, 48, 50, 54

FPR False Positive Rate. 40–44, 48, 49, 51

FS FaceSwap. vii, 24, 27, 28

HM Harmonic Mean. 7, 17, 39, 49, 50

LSTM Long Short-Term Memory. 13

MST Monk Skin Tone. 38

NT NeuralTextures. vii, 24, 27, 28

ROC Receiver Operating Characteristic. ix, 32, 33

TPR True Positive Rate. 41, 43, 44

ViT Vision Transformers. 13, 15

Chapter 1

Introduction

1.1 Background

Deepfakes are synthetic media, such as images, videos, or audio files, created or manipulated with deep learning algorithms. Rapid advancements in generative models have been a critical driver in the evolution of deepfake technology [6]. Historically, creating convincing deepfakes required extensive technical expertise and resources; however, the widespread and public availability of pre-trained generative vision models have significantly lowered these barriers [7, 8]. While deepfake technologies offer diverse applications in fields like the entertainment industry [9, 10], their increased accessibility also brings significant concerns regarding potential misuse. Such technologies could create false evidence, impersonate individuals for fraudulent purposes, or manipulate media to spread misinformation [11]. This is particularly alarming given the current global landscape where digital media play a central role in shaping public opinion [12].

Facial forgery detection, also known as deepfake detection, is a rapidly growing field with important real-world applications [13]. With the recent explosion in the success of sophisticated deep generative models [14, 15, 16], it has become increasingly

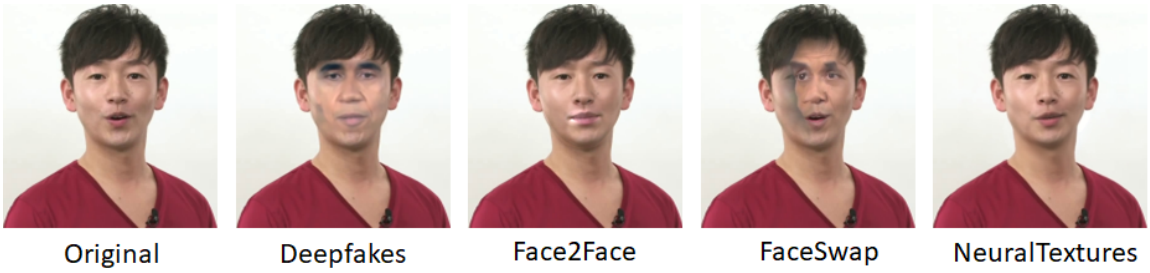


Figure 1.1: A facial image along with the four manipulated versions from the FaceForensics++ (FF++) dataset.

easy to generate highly realistic fake images and videos (see Figure 1.1 for an example of a real image along with four manipulated versions). The advancements in artificial intelligence have made it possible to create deepfakes that are nearly indistinguishable from genuine content, making the detection process even more challenging. This has led to a growing concern about the potential for malicious actors to use these tools for nefarious purposes, such as spreading misinformation, manipulating public opinion, or even causing social unrest. Given the potential risks associated with deepfakes, the importance of developing effective detection methods cannot be overstated [17, 18, 19].

The ease of creating deepfakes necessitates the development of sophisticated methods for the detection of content produced by generative models. Various deepfake detectors can successfully identify forged content in real-world scenarios [20, 21, 22]. To improve detection performance, the biometrics research community has employed machine learning algorithms to discern subtle inconsistencies and artifacts common in synthetic content [23, 24, 25]. Despite their wide application in the community, several studies [26, 27, 28, 9, 29, 30, 3] have shown that these systems are biased toward specific groups with regards to sensitive attributes like gender, racial background, age, and others.

1.2 Problem and Motivation

The field of deepfake detection has seen considerable progress in recent years with a number of sophisticated techniques being proposed in the area [31, 32]. However, in the context of detecting manipulated content in videos, many existing methods primarily focus on spatial features extracted from individual frames [25]. This approach can lead to the overlooking of temporal dynamics that evolve throughout video sequences. This strategy can result in limitations, as temporal artifacts such as flickering and motion discontinuities, are common indicators of deepfake manipulation. Furthermore, sophisticated deepfakes may exhibit subtle spatial inconsistencies that manifest over time, necessitating an integrated analysis of both spatial and temporal information. Moreover, we hypothesize that capturing subtle *spatiotemporal* inconsistencies that are often caused by different deepfake generation methods, could significantly enhance performance by learning representations that generalize to unseen forgery methods, which is often a challenging problem in this area.

Concurrently, the field faces significant open problems concerning fairness in deepfake detection. Although a number of bias-mitigating strategies have been proposed for deepfake detection [3, 28], the embeddings generated by these detectors continue to retain information related to sensitive attributes, which could cause biases in detection of deepfake content. Figure 1.2 shows representations from sample images produced by the encoder of a deepfake detector [2]. We observe that despite the application of a bias-mitigating approach [3], the representations still contain gender-specific information depicted as male/female clusters. As a result, we suggest that there is substantial room for improvement at the core of this problem. Despite efforts to develop less biased deepfake detectors that work fairly across different populations,

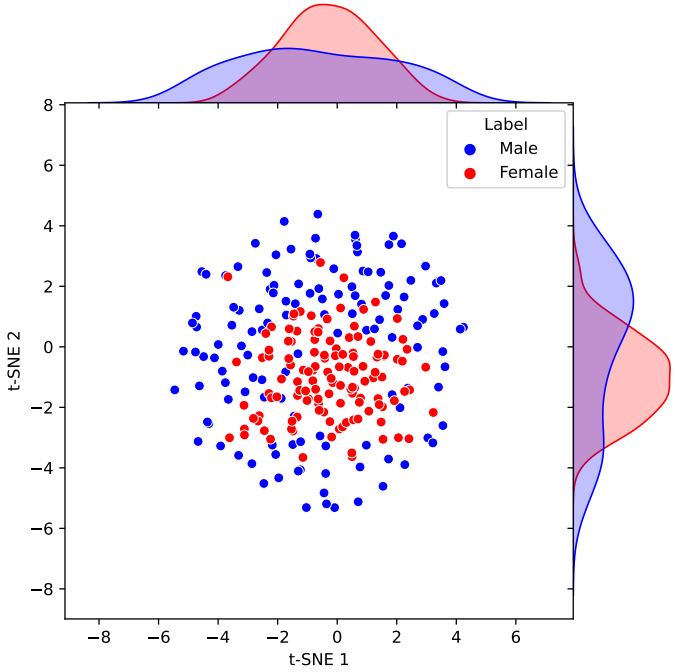


Figure 1.2: Representations for sample images from the CelebDF dataset [1] obtained from the EfficientNet-B4 detector encoder [2] with DAG-FDD [3] bias-mitigation. The male/female clusters indicate the existence of gender-related information in the embeddings.

gender has been the main area of focus. While a few works have focused on 'ethnicity' as a second factor, we argue that 'skin tone' [33] is a more critical and, at the same time, more practical factor to focus on. Moreover, given the prominence of skin tone in facial images, it is more likely for skin tones to be the reason for biased representations being learned by deep learning models as opposed to ethnicity. Lastly, there exists a phenomenon referred to as the 'fairness-accuracy trade-off', which indicates that enhanced fairness may come at the cost of reduced accuracy [34, 33]. While some studies suggest variability in the existence of this trade-off [35, 36, 37], the intertwined relationship of fairness and accuracy has been widely confirmed in prior works [38, 39, 40]. However, to our knowledge, this trade-off has not been studied in

prior works in the context of deepfake detectors.

Addressing these challenges is imperative for the advancement of the field of deepfake detection. Enhancing detection methods to incorporate both spatial and temporal information can lead to more robust and effective identification of deepfakes. Simultaneously, ensuring that these methods are fair and do not perpetuate biases is crucial for maintaining the integrity and societal acceptance of these technologies. Therefore, a balanced approach that addresses both the technical and ethical dimensions is essential for the future development of reliable and equitable deepfake detection systems.

1.3 Solutions Overview

In response to the technical challenges identified above, we propose two novel solutions. To address the first problem of accurate deepfake detection, we propose a dual-component detection system. One component learns spatial information from individual RGB frames of the videos, while the second component leverages optical flow fields to learn temporal consistency across the video. Both components utilize vision transformers [41], which we train in two steps. First, inspired by [42] we pre-train the models in an autoencoding setup using a self-supervised reconstruction scheme. Second, we discard the reconstruction decoder and add a new classification head to each encoder, where they are fine-tuned for deepfake detection, followed by score-level fusion of the results. We pre-train the spatial learning and temporal consistency learning encoders with CelebFaces-Attributes (Celeb-A) [43] and YouTube Faces [44] datasets respectively. For downstream deepfake detection, we evaluate our approach on the FF++ (High Quality) and FF++ (Low Quality) datasets [45] which employ

compression factors of 23% and 40%, respectively, in addition to the CelebDF dataset [1]. Additionally, we synthesize a more challenging variation of the FF++ dataset, which we call FF++ (Very Low Quality) by compressing the data with a rate of 65%. This dataset is then used to further evaluate our approach against existing techniques in the presence of extreme compression artifacts. Experimental results demonstrate that our method achieves state-of-the-art performance on FF++ (LQ and HQ) datasets and highlight the efficacy of our approach in detecting deepfakes across diverse compression levels. Ablation studies demonstrate the importance of different components of our method. Lastly, state-of-the-art results when fine-tuning our model on FF++ and testing it on CelebDF (cross-dataset evaluation) demonstrates the strong generalization of our method.

As the second aspect of our solution, we propose a novel loss term called *FairAlign*, for enhancing fairness via the alignment of conditional distributions of the embeddings in higher-dimensional kernel space. Our method aims to reduce the gap in detection performance across different sensitive attributes such as gender, thus mitigating the risk of biased outcomes. By leveraging the kernel space, our method integrates the cross-covariance and covariance operators of the conditional distributions of the embeddings given sensitive attributes obtained from deepfake detectors into the training process. Our method is a plug-and-play technique that can be integrated with other existing loss functions used in deepfake detection. Using our method, we perform multiple bias-mitigation experiments in deepfake detection on two public datasets (CelebDF [1] and FF++ [46]), wherein we demonstrate the effectiveness of our method in improving the fairness of several state-of-the-art deepfake detectors while retaining strong detection performance. Additionally, we propose a simple yet effective pipeline for detecting skin

tones and using them to mitigate bias for this factor. Our method uses ArcFace [47] to detect and extract the face. Subsequently, we select the facial skin regions using a pre-trained U-Net model [48], based on which the average skin color is measured. Finally, we use the shortest Euclidean distance between the skin tone with respect to Monk Skin Tone (MST) [49] scale to determine the final skin tone. We then apply this pipeline to deepfake detection datasets for the first time, following which we perform skin tone bias mitigation experiments. We find that our method, FairAlign, is effective at reducing skin tone biases in deepfake detection datasets. Lastly, to analyze deepfake detection with a balanced view of both fairness and accuracy, we utilize two metrics, Fairea [5] and Harmonic Mean (HM) [50, 51], and combine fairness and accuracy into unified indices. This is the first time the fairness-accuracy trade-off is being studied in the context of deepfake detection. Our analysis shows that while some existing fairness-promoting techniques do indeed reduce bias to a good degree, this improvement comes at the cost of accuracy, hence not ideal for practical applications. The analysis further demonstrates that our proposed FairAlign maintains the highest performance in terms of both fairness and accuracy as per the unified metrics.

1.4 Contributions

Our contributions in this thesis can be summarized as follows:

- We propose a new approach for effective facial forgery detection. Our method uses a vision autoencoding transformer and is pre-trained in a self-supervised masked reconstruction setup. Our solution consists of two main components which learn spatial (RGB) and temporal consistency information (optical flow fields) separately.

- We leverage relatively small datasets, namely Celeb-A and YouTube Faces, for pretraining our transformers, and achieve state-of-the-art results in the downstream task of deepfake detection on FF++ (LQ, HQ, and VLQ) datasets and competitive results on CelebDF.

- For promoting fairness in deepfake detection we propose a new loss term, FairAlign, that operates in the kernel space, to reduce the distance between distributions of the representations learned by deepfake detectors given different sensitive attribute groups. Our method demonstrates effectiveness in improving the fairness of state-of-the-art deepfake detectors while maintaining strong detection performance on two large-scale datasets, FF++ [46] and CelebDF [1].

- We analyze and improve fairness based on skin tones for deepfake detection tasks. We extract skin tones from existing deepfake datasets using the guidelines given by the MST scale[49], and apply our proposed FairAlign method for enhanced fairness. Our experiments demonstrate that FairAlign improves skin tone fairness across all state-of-the-art deepfake detectors. To our knowledge, this is the first attempt at reducing bias against different skin tones in the context of deepfake detection. Additionally, we enhance fairness based on the *intersection* of gender and skin tone in the context of deepfake detection for the first time.

- To objectively quantify the fairness-accuracy trade-off for bias-mitigating methods, we analyze two unified metrics for the first time in the realm of fair deepfake detection. Results show that our method is highly favorable as a bias-mitigating method that strikes a healthy balance between fairness and accuracy.

1.5 Publications

The following papers have resulted from this research :

- [23]: Sayantan Das, Mojtaba Kolahdouzi, Levent Özparlak, Will Hickie, Ali Etemad, "Unmasking Deepfakes: Masked Autoencoding Spatiotemporal Transformers for Enhanced Video Forgery Detection", *International Joint Conference on Biometrics (IJCB)*, 2023.

- Sayantan Das, Mojtaba Kolahdouzi, Ali Etemad, "FairAlign: Embedding Distribution Alignment for Bias Reduction in Deepfake Detection", *Under Review*, 2024.

1.6 Organization of Thesis

The rest of this thesis is organized as follows. Chapter 2 offers a comprehensive review of deepfake detection methods including classical solutions, Convolutional Neural Network (CNN), and attention mechanisms. Next we explore the role of frequency analysis, temporal approaches, and self-supervised learning in deepfake detection. This is followed by a discussion on multimodal solutions. Next, we review prior work on fairness in deepfake detection, followed by studies on the fairness-accuracy trade-off. Chapter 3 presents our novel method for spatiotemporal deepfake detection using masked reconstruction followed by the experiment setup and results of our approach. Chapter 4 presents our newly proposed loss term for enhancing fairness in deepfake detection, as well as our simple and effective pipeline for skin tone bias mitigation. We then present the experiments and results in comparison to other bias-mitigation techniques for gender, skin tone, and intersection of the two,

across various deepfake detection backbones. We wrap this chapter up by a through analysis on fairness-accuracy trade-off. Finally Chapter 5 concludes the thesis by summarizing our work, followed by a discussion on limitations and potential future research directions.

Chapter 2

Related Work

In this section, we review prior works on deepfake detection using various approaches. We follow this with a review of literature on fairness in deepfake detection. Finally, prior works on skin tone fairness and fairness-accuracy trade-off are reviewed.

2.1 Deepfake Detection

Deepfake detection has traditionally been addressed as a binary classification task [32], where the objective is to discern between authentic and manipulated media. The application of deep learning models, particularly CNN, has been central to achieving this objective [52, 53, 54]. Authors of FF++ dataset [45] used Xception network, which was one of the best-performing architectures at the time, to perform deepfake detection via transfer learning [52].

Researchers proposed a method in [53] that utilizes residual-based descriptors in the form of a constrained CNN for image forgery detection. This approach aims to capture and analyze the residual noise present in manipulated images, which can be a strong indicator of forgery. In contrast, another method introduced a deep learning approach that focuses on the mesoscopic properties of images [54]. In this

context, 'mesoscopic' refers to properties or features that fall between the small scale (microscopic) and the large scale (macroscopic). By concentrating on mesoscopic features, the model can capture subtle artifacts and inconsistencies in manipulated images, potentially making it more effective in detecting forgeries.

2.1.1 Attention Mechanisms

Attention mechanisms [55], combined with CNN [56, 57], have been adopted to enhance interpretability and facilitate the identification of manipulated regions in images. These attention-based models generate attention maps highlighting regions contributing significantly to the detection decision. For instance, [56] used attention maps generated by deep semantic features to outline crucial regions that contributed towards the classification result. The low-level textural feature and high-level semantic features are then aggregated and guided by these attention maps, which helps to capture more subtle artifacts in the image. The work in [57] proposed a new attention mechanism to calculate self-information from the input feature map and output a discriminative attention map that highlights regions contributing significantly to the detection decision.

2.1.2 Frequency Analysis

Various studies have utilized frequency analysis to detect inconsistencies that arise during deepfake creation [58, 59]. In [58], the researchers employed the phase spectrum for forged face image detection, showing that CNN can identify additional implicit phase spectrum features that are advantageous in detecting face forgeries. Concurrently, the study in [59] developed a multi-scale patch similarity module to specifically model

second-order relationships between distinct local regions, forming a similarity pattern through pairwise cosine measurements. These patterns distinguish real from forged regions by recognizing differences such as irregular textures and high-frequency noise.

2.1.3 Temporal Approaches

According to [32], most detection techniques applied to images are not ideally suited for use in videos, as these methods tend to overlook temporal dynamics. In [60], a Recurrent Convolutional Network leveraged spatiotemporal features [61] of videos to identify deepfakes. Similarly, the work in [62] identified inconsistencies within and between frames of deepfake videos. They developed a model combining CNN and Long Short-Term Memory (LSTM) to detect these discrepancies. In their approach, the CNN is tasked with extracting features from individual frames, while the LSTM processes these features to create a descriptor for temporal sequence analysis. A recent work [63] introduces the HCiT framework, blending CNN and Vision Transformers (ViT) models to enhance deepfake detection through detailed spatiotemporal inconsistency analysis, achieving superior performance and generalization across diverse video content. This focus on spatiotemporal analysis shows that our research is in line with current trends in deepfake detection methods.

2.1.4 Self Supervised Learning

Self-supervised learning (SSL) has been explored to address the issue of limited labeled data for deepfake detection [64, 65, 66, 67, 68]. For instance, self-supervised learning was employed in [64] with an auxiliary task specifically designed for deepfake detection, using a synthesizer and adversarial training framework to dynamically

generate forgeries. This approach enriches diversity and strengthens sensitivity to produce strong results. In the method proposed in [65], mouth motion representations were learned by encouraging close-paired video and audio representations while keeping unpaired ones diverse. The study in [66] proposed a decoupling strategy to separate facial authenticity and compression relevance, implemented through a joint self-supervised learning approach using compression ratios as self-supervised signals. Another study utilized a multi-modal backbone trained in a self-supervised manner and adapted it to the video deepfake domain [67]. These self-supervised models leverage unlabeled data to learn useful representations for detection tasks. Contrastive learning is another common pre-text learning approach often considered for deepfake detection [69, 68]. In the study by [69], two different transformed versions of a face image were generated using two distinct transformations. The agreement between these transformed images is maximized after they are passed through an encoder network and a projection head network, effectively training the model without supervision signals. On the other hand, another study employed supervised contrastive learning to learn common features between instances of the same class, while distinguishing between samples from different classes [68].

2.1.5 Multimodal Approaches

In recent studies, researchers have sought to combine multiple modalities, such as visual, audio, and temporal information, to improve detection performance in deepfake detection tasks [58, 70, 71]. These multi-modal approaches provide a comprehensive view of the media, making them more robust against limitations specific to individual modalities [72]. While frequency-based modalities have been employed in multi-modal

deepfake detection solutions [58], we believe that optical flow has the potential to serve as an effective alternative source of information.

Prior works such as [73, 74, 75, 76] have used optical flow features followed by a classifier for deepfake detection. In contrast, for the first time, our method leverages optical flow information alongside RGB features in a masked autoencoding setup to improve cross-dataset generalizability and robustness. To our knowledge, prior deepfake detection methods involving ViT [77, 78, 79, 80, 81] have not used optical flow information to capture the temporal inconsistencies.

2.2 Fairness in Deepfake Detection

Prior works have demonstrated the existence of bias in deepfake detection tasks with respect to gender, age, and ethnicity [33, 26, 82, 27], while a few works have proposed solutions to mitigate this bias [28, 3]. The work in [26] evaluates bias in existing deepfake datasets and detection models for the first time in the literature; however, it doesn't take into account the intersectional bias. The evaluation of bias for a popular detection model (MesoInception-4) on FF++ dataset is done in [82]. This work is limited in terms of the number of detection methods evaluated. A more comprehensive study is proposed in [27], which evaluates fairness over three deepfake detection models. One recent work [28] has attempted to mitigate the aforementioned biases through data-centric approaches, i.e., making the datasets like FF++ balanced with regards to different sensitive attributes like gender. The process of gender-balancing via data annotation is time-consuming and also showed limited improvement in fairness. The work in [3] applies conditional-value-at-risk loss to mitigate bias with regard to both gender and ethnicity in the context of deepfake detection. To our knowledge, this

paper is the first and only one to directly provide a solution for bias in deepfake detection.

2.2.1 Skin Tone Fairness

While ethnicity has been considered for promoting fairness [3, 26, 27] in several deepfake detection literature, skin tone as a factor in reducing bias is more commonly addressed in areas like skin lesion classification [83, 84]. To our knowledge, [33] is the only prior work to study skin tone in the context of deepfake detection. In their study, researchers evaluate biases in deepfake detection by analyzing top models from the DeepFake Detection Challenge [85] on the Casual Conversations dataset, which is rich in diversity across age, gender, and skin tone. Their analysis confirms the importance of skin tone as a crucial sensitive attribute for bias mitigation in deepfake detection.

2.2.2 Fairness-Accuracy Trade-o□

While fairness-accuracy trade-off is a well-known phenomenon [34, 37, 35], only a handful of works have focused on introducing a quantitative measure to assess the trade-off between fairness and accuracy [37, 34, 86], although none are positioned in the context of deepfake detection. One such work defines the Fairness-Area-Under-the-Curve (FAUC) to empirically define the fairness-accuracy Pareto frontier [34]. FAUC provides a model-agnostic metric to measure the Pareto frontier. However, as mentioned in their work, FAUC is ineffective when intersectional fairness is involved or in cases where fairness and accuracy typically do not have an inversely proportional relationship. Another work [86] approaches this trade-off through the lens of multi-task learning by proposing two metrics: Average Relative Fairness Gap and Average Relative

Error. These metrics compare the Fairness-Performance Rate Gap and error rates of multi-task models to those of single-task models with the same architecture, providing a nuanced assessment of the balance between fairness and accuracy in multi-task learning. With a different perspective, [37] approaches the trade-off between fairness and accuracy by quantifying separability with Chernoff information, challenging the use of biased datasets for performance measures, and advocating for ideal, unbiased datasets. Our work utilizes Fairea [5] and HM [50], which were previously not explored in the context of deepfake detection. Fairea quantifies the trade-off by computing the area within an enclosed region formed by the baseline fairness values and the coordinates of any bias-mitigation method in a two-dimensional fairness-accuracy space. HM computes the trade-off using a straightforward formula that balances accuracy and fairness, ensuring neither is overlooked in the evaluation process.

Chapter 3

Masked Autoencoding for Deepfake Detection

3.1 Proposed Method

3.1.1 Overview

Our proposed approach titled Masked Autoencoding Spatiotemporal Deepfake Transformer (MASDT) consists of two components: spatial learning and temporal consistency learning. The spatial learning component has the objective of learning robust spatial features from the RGB images, while temporal consistency learning aims to extract temporal features from optical flow fields derived from the input images. We fuse the classification outputs derived from the spatial and temporal consistency learning components. Both these components follow a self-supervised autoencoding approach in a *two-step* process.

The first step involves a self-supervised pre-training strategy which involves both of the MASDT components in a data reconstruction task. We discuss this strategy in section 3.1.2. The second step is the downstream task of deepfake detection, wherein we re-purpose components trained in the previous step to perform the classification of deepfake data through a model fine-tuning process, followed by fusion of information

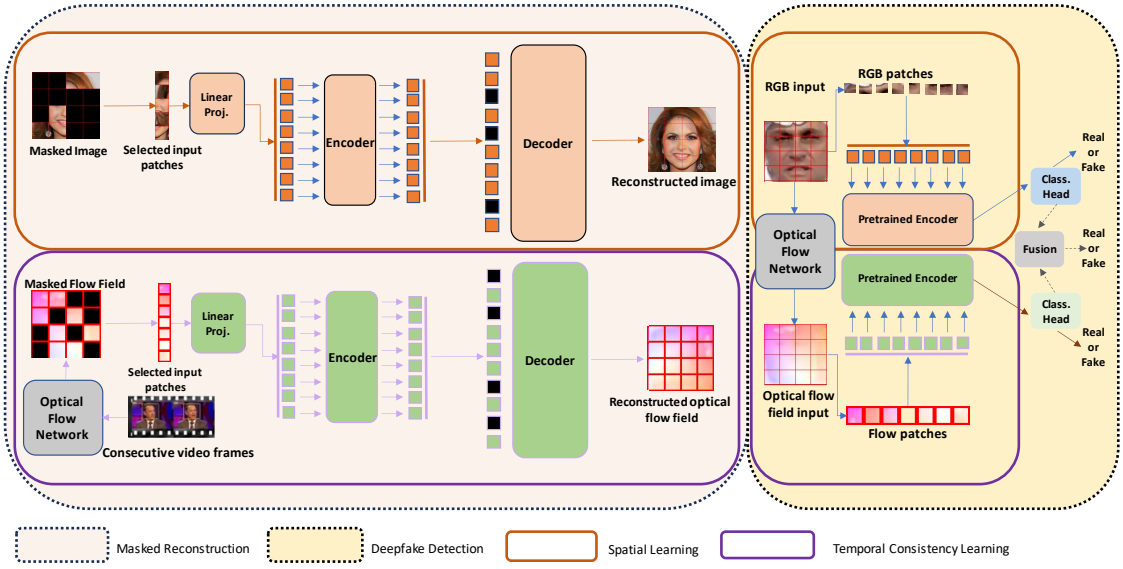


Figure 3.1: An overview of our method (MASDT), which includes the masked facial reconstruction and deepfake detection steps.

from both components (spatial learning and temporal consistency learning). This is discussed in detail in Section 3.1.3. Before we discuss each of the two steps, we discuss the optical flow field generation strategy in Section 3.1.1. A general scheme of the MASDT strategy is presented in Figure 3.1.

Optical flow field estimation

We utilize a CNN model named PWC-Net for generating optical flow fields [87]. Let the model for estimating optical flow be F_θ, and two consecutive frames be f_t and f_{t+1}. Accordingly, the estimated optical flow $\quad_t$ can be denoted by:

$$\quad_t = F_\theta(f_t, f_{t+1}), \tag{3.1}$$

where $_t$ is a 3-channel optical flow matrix of size H W 3 representing the flow field between the consecutive frames.

3.1.2 Masked Facial Reconstruction

The first step of our approach utilizes a masked self-supervised auto-encoder which learns to reconstruct original facial images, given partial observations [42]. This auto-encoder reconstruction pipeline consists of two blocks: a reconstruction encoder, which captures a latent representation from the visible portions of each image, and a reconstruction decoder that aims to reconstruct the masked sections of the image using this latent representation. In this procedure, the encoder is trained to extract robust spatial features from masked facial images, eliminating noise and redundancy while transforming the reconstruction task into a challenging process that requires generalizing features to represent a small subset of available data [88]. Consequently, by masking portions of the facial image using random spatial pixels or patches, we can avoid a potential location bias toward image reconstruction, which can be critical for the detection of deepfake images.

The goal of the decoder is to use the features obtained from the latent space by the encoder to reconstruct the masked information from the original facial image. We train this reconstruction encoder-decoder pair using a simple mean squared error (MSE) reconstruction loss L_r:

$$L_r = \frac{1}{N}\sum_{i=1}^{N}(y_i \quad \hat{y}_i)^2, \tag{3.2}$$

where N represents the number of sampled patches, and $\hat{y}_i$ and y_i are the ith output and expected ith output, respectively.

We perform the above masked reconstruction task for both the components independently where we employ the encoder-decoder pairs for reconstructing RGB images y and optical flow fields for spatial learning and temporal consistency learning respectively, which is also referred to as pre-training in self-supervised learning literature. This prepares the encoders for the fine-tuning step mentioned in the next section.

3.1.3 Deepfake Detection

The second step of MASDT is aimed at the supervised training for the classification of deepfake images. For this purpose, we employ the encoders that learned to extract robust representations in the reconstruction pipeline. Thus, to perform binary classification, a classification head consisting of a simple MLP is attached to each of the pre-trained encoders.

We adopt a dual-encoder setup for the fine-tuning process, utilizing the spatial learning encoder θ_s and the temporal consistency learning encoder θ_t. These encoders were previously trained in the initial step of our proposed solution. In the process of fine-tuning for a binary classification task, we employ a binary cross-entropy loss, denoted as L_b. The formula for this loss is as follows:

$$L_b = \frac{1}{M}\sum_{j=1}^{M}[o_j \log(\hat{o}_j) + (1 - o_j)\log(1 - \hat{o}_j)], \tag{3.3}$$

Here, $\hat{o}_j$ is the predicted output from the network, o_j represents the actual or target class (either 0 or 1), and M denotes the total count of samples in the batch.

Dual Modality Fusion

To further harness the strengths of both θ_s and θ_t, we use a simple fusion mechanism. This method aims to exploit the complementary information that each encoder provides, thereby improving the overall classification performance. The fusion process begins with the individual outputs from θ_s and θ_t, denoted as $\hat{\theta}_s$ and $\hat{\theta}_t$ respectively, which are then combined to create a fused score-level prediction, $\hat{\theta}_f$. Mathematically, this can be expressed as:

$$\hat{\theta}_f = \alpha \quad \hat{\theta}_s + (1 \quad \alpha) \quad \hat{\theta}_t, \tag{3.4}$$

where α is a fusion weight that determines the contribution of each encoder to the final output.

3.2 Experiments

In this section, we present the specifics and details of our method and experiments, describe the datasets used, and discuss the ablation studies conducted to validate the impact of different components of our proposed solution.

3.2.1 Implementation Details

In this section, we outline the implementation details of our deepfake detection method, which incorporates both RGB and optical flow modalities. Our experiments are conducted using the PyTorch framework [89] on 4 Nvidia A100 GPUs, each with 40 GB of vRAM. We generate optical flow fields using the PWC-Net present in the MMFlow toolbox [90].

Our method's performance is evaluated using the top-1 accuracy, which denotes

the percentage of correctly classified deepfake and real videos out of the total number of videos in the test set. This metric is widely used in deepfake detection tasks as it provides a clear indication of a model's ability to distinguish between real and fake videos. Accuracy and Area Under the Curve (AUC) are presented as the metrics for our experiments, following other publications in the area.

For evaluation purposes, we use the FF++ (LQ and HQ) and CelebDF datasets (the details of these datasets are presented in the next Section) and divide them into training, validation, and test sets, ensuring an even distribution of deepfake and real videos across all sets following the instructions provided in the original dataset papers [45, 1]. Data augmentation techniques such as random cropping, horizontal flipping, color jittering, and MixUp augmentation, are employed to improve our model's robustness to input data variations. MixUp augmentation [91] involves generating new training samples by taking linear combinations of input data and their corresponding labels, which encourages the model to learn smooth and robust features. In addition to MixUp, the model employs CutMix [92] data augmentation technique with default settings (alpha set to 0, probability set to 1, and switch probability set to 0.5). Label smoothing is applied with a smoothing factor of 0.1. A drop path rate of 0.1 is used for stochastic depth regularization.

Input images are resized to 224 224, with patches of 16 16. We observe that a masking ratio of 90% is optimal for pre-training. We use the transformer architecture [41] with a default Vit-B configuration as our model. The model is trained using the AdamW optimizer, with a weight decay of 0.05, a base learning rate of 5 10^{-4}, and layer decay of 0.8. The learning rate is scaled according to an effective batch size of 64. We train the model for 300 epochs, using a gradient accumulation of 1 iteration.

We utilize a distributed training approach with distributed evaluation. The CUDA benchmark is enabled, and the model is trained on available CUDA devices. For fine-tuning, the model is initialized with our pre-trained weights from the first step (self-supervised pre-training), and position embeddings are interpolated accordingly.

3.2.2 Datasets

We use the FF++ (LQ), FF++ (HQ) [45], CelebDF [1], Celeb-A [43], and YouTube Faces [44] datasets. The first three datasets are employed for evaluating our proposed method, while the latter two are utilized for pre-training only. Below, we provide a detailed description of each dataset:

FF++ (LQ) [45] simulates various scenarios where manipulated videos appear in compressed formats. With a 40% compression factor using the H.264 video compression standard, the LQ version introduces artifacts that may be present in real-world cases. This dataset challenges researchers to develop techniques capable of detecting manipulations even when the video quality is degraded due to compression.

FF++ (HQ) [45] maintains a higher quality (compression factor of 23%) compared to the LQ version, enabling researchers to study deepfakes and other manipulations with greater detail and less information loss due to compression. Both FF++ versions contain over 1000 original videos, with manipulated videos created using various methods, such as Deepfakes (DF) [93], FaceSwap (FS) [94], Face2Face (FF) [95], and NeuralTextures (NT) [96]. These datasets cover a wide range of manipulation methods, allowing researchers to test their detection algorithms on diverse types of deepfakes.

In order to further push our method to the limit and challenge its detection ability in the presence of significant compression artifacts, we create an even more compressed

version in comparison to FF++ (LQ), which we call **FF++ (VLQ)** where VLQ stands for very low quality. To generate this variant of the dataset, we take the original non-compressed videos of FF++ and compress them by a compression factor of 65%, which we will also use in our experiments besides the datasets with two standard compression ratios. For this purpose, we use the FFMPEG framework [97].

CelebDF [1] includes 590 original videos collected from YouTube, featuring subjects of varying ages, ethnicities, and genders, as well as 5639 corresponding DeepFake videos. The CelebDF dataset's average video length is 13 seconds, and all videos have a standard 30 FPS frame rate.

Celeb-A [43] is a large-scale collection of over 200,000 celebrity images, with 40 attribute labels annotated for each image. The dataset comprises diverse subjects and captures various facial expressions, poses, and lighting conditions.

YouTube Faces [44] is a comprehensive collection of videos from YouTube focusing on individuals' faces. It contains over 3,000 annotated videos of 1,595 people, offering diverse subjects with different ethnicities, ages, and genders. Each video in the dataset is labeled with the corresponding subjects' identities, and is often used for face recognition and verification tasks. It captures various poses, expressions, illuminations, and occlusions.

3.2.3 Pre-training Strategy

For pre-training the RGB modality in our proposed method, we utilize the Celeb-A dataset instead of the typically used ImageNet [98]. Celeb-A is considerably smaller than ImageNet, as Celeb-A contains 200,000 images whereas ImageNet contains over 14 million images. This reduced size allows for faster pre-training and lower

computational requirements, making the process more efficient and accessible to a wider range of researchers and practitioners. Celeb-A is specifically tailored for facial tasks, consisting exclusively of human face images. In contrast, ImageNet covers many object categories and may not be as well-suited and efficient for facial analysis. By pre-training our model on Celeb-A, we ensure that the initial features learned by the model are more relevant to facial structures, expressions, and attributes, which can ultimately contribute to a more effective deepfake detection system.

For pre-training the optical flow modality in our method, we utilize the YouTube Faces dataset. This dataset provides video data, essential for optical flow calculation. Naturally, datasets of images such as ImageNet and Celeb-A cannot be used for optical flow generation. Moreover, the YouTube Faces dataset is specifically designed for facial analysis tasks as it consists exclusively of human face videos. By pre-training our model on this dataset, we ensure that the initial features learned by the temporal consistency encoder can better capture information such as facial structures, expressions, and attributes, ultimately contributing to a more effective deepfake detection system.

3.2.4 Results

In this section, we present the outcome of our experiments, which assess the performance of the proposed method for deepfake detection on the FF++ and CelebDF datasets. Our evaluation concentrates on the effectiveness of integrating both RGB and optical flow modalities, as well as the impact of pre-training on the Celeb-A and YouTube Faces datasets. By contrasting our approach with existing methods and baseline models, we aim to evaluate the benefits of our technique in accurately identifying deepfakes under a range of conditions.

Table 3.1: Quantitative results for ACC and AUC on the FF++ dataset with both quality settings (LQ and HQ). The results are arranged in ascending order on the basis of ACC (LQ).

Methods	ACC (HQ)	AUC (HQ)	ACC (LQ)	AUC (LQ)
Steg. Features [99]	70.97%	-	55.98%	-
LD-CNN [53]	78.45%	-	58.69%	-
CP-CNN [100]	79.08%	-	61.18%	-
Face X-ray [101]	-	87.40%	-	61.60%
C-Conv [102]	82.97%	-	66.84%	-
MesoNet [54]	83.10%	-	70.47%	-
FakeCatcher [103]	94.65%	-	-	-
Two-branch RN [104]	96.43%	88.70%	86.34%	86.59%
Xception [105]	95.73%	-	86.86%	-
LipsDontLie [106]	-	97.10%	-	-
Capsule Net [107]	-	99.50%	-	-
SLADD [64]	-	98.40%	-	-
MADD [56]	97.60%	99.29%	88.69%	90.40%
Self Info. Att. [57]	97.64%	99.35%	90.23%	93.45%
F3-Net [108]	97.52%	98.10%	90.43%	93.30%
E2E Learning [25]	97.06%	99.32%	91.03%	95.02%
Local Relation Learning [59]	97.59%	99.46%	91.47%	95.21%
Ours	98.19%	99.67%	97.79%	98.45%

In Table 3.1 we present the top-1 accuracy and AUC scores of our proposed method compared to the current state-of-the-art approaches. The table presents the quantitative results for various deepfake detection techniques available in the FF++ dataset with both high and low quality settings. It can be observed that our proposed method achieves the highest accuracy and AUC scores in both quality settings, surpassing the prior works and setting a new state-of-the-art.

In our experiments, we assess the performance of different deepfake generation methods in the FF++ (LQ) dataset, comprising four distinct techniques: DeepFakes (DF) [93], FF (FF) [95], FS (FS) [94], and NT (NT) [96], as illustrated in Table 3.2. In this table, we present a breakdown of the performance of our method and others

Table 3.2: Quantitative results (ACC) on the FF++ (LQ) dataset with four manipulation methods: DF, FF, FS, and NT.

Methods	DF [93]	FF [95]	FS [94]	NT [96]
Steg. Features [99]	67.00%	48.00%	49.00%	56.00%
LD-CNN [53]	75.00%	56.00%	51.00%	62.00%
C-Conv [102]	87.00%	82.00%	74.00%	74.00%
CP-CNN [100]	80.00%	62.00%	59.00%	59.00%
MesoNet [54]	90.00%	83.00%	83.00%	75.00%
Xception [105]	96.01%	93.29%	94.71%	79.14%
F3-Net [108]	97.97%	95.32%	96.53%	83.32%
Local Relation Learning [59]	98.84 %	95.53%	97.53%	89.31%
Ours	97.84%	96.27%	97.89%	78.23%

Table 3.3: Quantitative results in terms of ACC and AUC on the CelebDF dataset.

Methods	ACC	AUC
F3-Net [108]	95.95%	98.93%
Xception [105]	97.90%	99.73%
E2E Learning [25]	98.59%	99.94%
Ours	98.00%	98.90%

across these four deepfake generation methods, and compare the accuracy with other state-of-the-art approaches. The results indicate that our method achieves strong results across all four manipulation techniques, particularly in the FF and FS methods, and generates competitive results for the other two. These findings demonstrate the effectiveness of our approach in detecting manipulated face images across different forgery approaches.

Next, we evaluate the performance of our method compared to other recent methods on the CelebDF dataset and present the performance in Table 3.3. It can be observed that our method achieves results competitive to the current state-of-the-art [25].

To further explore the generalization capability of our model, we follow the cross-dataset scheme presented in [25], [105], and [59]. In this experiment, we train the model on the FF++ datasets and test its performance on the CelebDF dataset. We

Table 3.4: Cross-dataset evaluation (AUC) by training on FF++ (LQ) and testing on the CelebDF dataset.

Methods	AUC
Xception [105]	36.19%
E2E Learning [25]	68.71%
Local Relation Learning [59]	78.26%
Ours	80.21%

present the results in Table 3.4, where we observe that our method outperforms prior works in the area, indicating strong generalization ability in detecting deepfakes even when training is done on a different dataset and likely constitutes a different distribution (out-of-distribution).

To further push our approach to the limit, we explore its performance on the VLQ version of the FF++ dataset which we constructed for the first time by applying a 65% compression ratio (see Section 3.2.2). We also use this dataset on two leading methods, namely DCL [109] and E2E Reconstruction Learning [25]. The results are presented in Table 3.5 where we observe that our method outperforms both other solutions, highlighting the efficiency and resilience of our approach in detecting deepfakes, even in the presence of highly compressed data.

To better contextualize our method within prior works, we compare the performance of our method to prior methods that have used optical flow and vision transformers in Tables 3.6 and 3.7 respectively. The results show that we achieve superior results in comparison to other methods that have used optical flow for deepfake detection when evaluating on the FF manipulation method of the FF++ dataset. Similarly, we outperform other methods in the literature that leverage vision transformers.

Lastly, we utilize Grad-CAM [4] visualization on our model and similar performing

Table 3.5: Quantitative results on FF++ (VLQ) dataset which is constructed by applying a 65% compression ratio.

Methods	ACC
DCL [109]	65.20%
E2E Learning [25]	78.20%
Ours	79.70%

methods to demonstrate and investigate the attention patterns of each method. Grad-CAM is capable of pinpointing the areas that the network applies more attention to, and thus deems important. We present a sample image in Figure 3.2, where the red areas highlight parts of the image which are more salient for the models. We observe that our model considers broader areas of the face image as important toward detection of whether the input is a deepfake image or not. This is a noteworthy observation as it indicates that the proposed method is capable of capturing a more comprehensive set of features and artifacts, which might be overlooked by the other models. This ability to focus on multiple areas simultaneously could enable the proposed method to better discern subtle inconsistencies and artifacts that are characteristic of deepfakes or manipulated images. In contrast, the other two models, with their more concentrated attention patterns, may be less effective in capturing the full extent of these subtle cues, which might result in lower overall performance in detecting such forgeries. Another interesting pattern which can be observed is that prior methods seem to focus on select areas, namely the left eye and to some extent the right ear. However, in addition to these regions, our method considers the nose and mouth regions, which are critical areas for authentic face images.

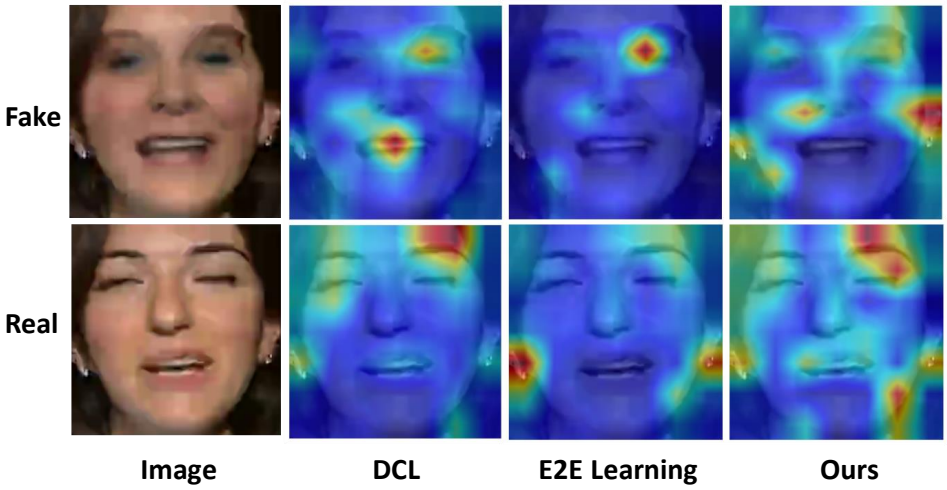


Figure 3.2: Comparison of Grad-CAM visualizations [4] for our method in comparison to two other recent works.

Table 3.6: Comparison to prior deepfake detection methods that use optical flow. The results are reported on the FF manipulation method of FF++.

Methods	Training dataset	AUC
OF + CNN [73]	FF++	-
OF + CNN [74]	FF++	-
OF + CNN-LSTM [75]	-	79.00
Ours	FF++ (LQ)	80.21
Ours	FF++ (HQ)	82.19

3.2.5 Ablation Studies

In this section, we investigate the contributions of different components of our method toward facial forgery detection. As the first step, we remove the temporal consistency encoder and present the results in Tables 3.8, 3.9, and 3.10, for FF++ (LQ), FF++ (HQ), and CelebDF, respectively. When comparing these results to the performance of our original method (also presented in each table), we observe that removing the temporal consistency encoder results in performance drops of 1.2% to 2.9%. This

Table 3.7: Comparison to other ViT-based deepfake detection methods. The results are reported on FF++.

Methods	Training Dataset	AUC
ViXNet [110]	FF++	74.78
Conv ViT [81]	FF++	71.80
UIA-ViT [77]	FF++	99.33
Ours	FF++ (HQ)	99.67

indicates the importance of learning additional temporal information through optical flow which may be difficult for the model to learn without explicit supervision.

Next, we examine the impact using simple score-level fusion in our model. To this end, we adopt two strategies instead. First, we use the joint learning approach proposed in [111], where a single pre-trained encoder accepts patches from both the RGB and optical flow modalities simultaneously. Second, instead of score-level fusion, we use feature-level fusion immediately after the embeddings are obtained from the spatial and temporal consistency encoders. The results for both experiments are presented in Tables 3.8, 3.9, and 3.10, for the three datasets, respectively. We observe that while feature-level fusion achieves results closer to ours in comparison to joint learning, our method still obtains superior results to both these strategies.

Lastly we illustrate the Receiver Operating Characteristic (ROC) curves for our method (depicted in blue) and the three ablated variants discussed above, in Figure 3.3. These results are obtained on the FF++ (LQ), demonstrated in Table 3.8. We observe that the true positive rates are generally higher than the model variants across different false positive rate regions, except for the version where temporal consistency is not used, which shows comparable results in true positive rates for high false positive regions. This indicates that the temporal consistency component is highly effective in reducing the number of false alarms.

Table 3.8: Ablation experiments on FF++ (LQ).

Technique	ACC	AUC
Proposed	97.79%	98.45%
w/o temporal consistency	96.51%	97.03%
w/ joint learning [111]	95.02%	97.05%
w/ feature-level fusion	96.01%	97.10%

Table 3.9: Ablation experiments on FF++ (HQ).

Technique	ACC	AUC
Proposed	98.19%	99.67%
w/o temporal consistency	96.90%	97.35%
w/ joint learning [111]	95.81%	97.58%
w/ feature-level fusion	98.01%	99.09%

Table 3.10: Ablation experiments on CelebDF.

Technique	ACC	AUC
Proposed	98.00%	98.90%
w/o temporal consistency	95.08%	97.17%
w/ joint learning [111]	95.06%	96.55%
w/ feature-level fusion	96.81%	98.10%

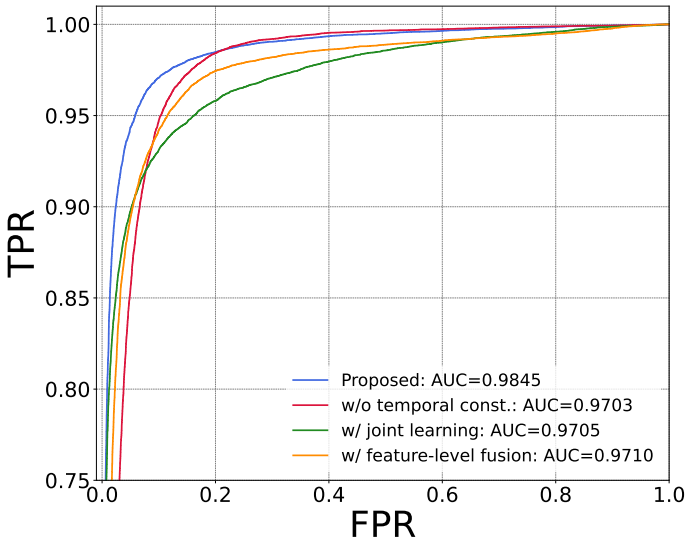


Figure 3.3: ROC curves for our proposed method (blue) and three ablations on the FF++ (LQ) dataset.

Chapter 4

Toward Fair Deepfake Detection via Embedding Distribution Alignment

4.1 Proposed Method

In this work, we introduce a novel loss term called FairAlign for enhancing fairness in deepfake detection. Training a deepfake detector using FairAlign causes the alignment of conditional distributions of embeddings given different sensitive attributes. Our fairness-enhancing approach is inspired by domain adaptation tasks [112, 113, 114] and is considered an in-processing fairness technique [115], as it intervenes directly within the learning algorithm to promote fairness.

4.1.1 Problem Setup

Let us represent a triplet $\{(x_i, y_i, z_i)\}_{i=1}^{N}$, where x_i denotes the embedding generated by deepfake detection model for the ith input image, y_i is its corresponding forgery label (real, fake), and z_i is the associated sensitive attribute (e.g., female, male, etc.). For simplicity, we can assume that the embeddings x_i are independent and identically distributed. Additionally, let X, Y, and Z be the random variables associated with the

embeddings, their labels, and the sensitive attributes, respectively. Let the distribution of X be defined over the set $\mathcal{X}$ and the distribution of Z be defined over the set $\mathcal{Z}$. Also, let the cardinality of the set $\mathcal{Z}$ be equal to ζ; for example, in the FF++ dataset, $\zeta = 2$ for gender, which corresponds to male and female. We alternatively denote embeddings x_i as x_i^a and x_i^b where a and b denote elements in the set $\mathcal{Z}$. Let us denote $P_{(.)}$ as the probability distribution of an arbitrary random variable. Accordingly, the optimum condition for a fair classifier is denoted by

$$P_{\hat{Y}}(\hat{y}|Z = z_i) = P_{\hat{Y}}(\hat{y}|Z = z_j) \quad \forall z_i, z_j \in \mathcal{Z}, \tag{4.1}$$

where $\hat{Y}$ represents the random variable associated with the classifier's output, i.e., the predicted forgery category. This optimum condition is also known as demographic parity, which is the objective of many bias-mitigating methods [116]. In this thesis, we aim to reduce the information related to the sensitive attribute z_i from the embeddings using our proposed loss.

4.1.2 FairAlign

Our overall goal is to accurately capture and minimize the distance between the distributions of embeddings given different sensitive attributes, through a novel loss term. The discrepancy between two distributions can be captured with regard to different statistical measures like expected value, covariance, etc. However, it has been shown that simply considering the arithmetic difference of such statistical measures, particularly in lower dimensions, cannot effectively capture the discrepancy [117]. As an example, Maximum Mean Discrepancy (MMD) [118] uses the difference in expected values of two distributions in high dimensions (specifically, Reproducing Kernel Hilbert Space (RKHS) [119]) to render the discrepancy between two distributions. In contrast,

the Bures metric [120], which defines the discrepancy between two distributions as the difference between their covariance matrices, cannot effectively capture the discrepancy unless the data are mapped onto a high dimensional space. To avoid explicit data projection onto higher dimensions and the additional computational load required to measure the difference in high-dimensional space, kernel functions [121] can be used to operate directly in low-dimensional space. Usage of kernel functions allows establishing the Bures metric in RKHS [119] where it is viable to be used as a distance metric [121]. This is also termed the Kernel Bures metric. Building on this concept, we utilize the Conditional Kernel Bures (CKB) metric, which is particularly designed for conditional distributions [122, 117].

As defined earlier, $\{x_i^a, z_i^a\}_{i=1}^n$ and $\{x_j^b, z_j^b\}_{j=1}^m$ are sets of embeddings corresponding to different sensitive attributes, drawn from the conditional distributions $P_{X|Z=z_a}$ and $P_{X|Z=z_b}$, respectively. Let's define kernel functions k_X and k_Z on the space of embeddings X and Z, respectively. Further, we define $\phi(x) = k_X(x, .)$ and $\psi(z) = k_Z(z, .)$ as feature mappings from X to RKHS H_X and Z to RKHS H_Z respectively. Now, let us denote either a or b by the notation a/b. Accordingly $K_{XX}^{a/b}$, $K_{ZZ}^{a/b}$, and K_{XX}^{ba} are the kernel matrices, where $(K_{XX}^{a/b})_{ij} = k_X(x_i^{a/b}, x_j^{a/b})$, $(K_{ZZ}^{a/b})_{ij} = k_Z(z_i^{a/b}, z_j^{a/b})$, and $(K_{XX}^{ba})_{ij} = k_X(x_i^b, x_j^a)$.

The feature mappings can therefore be represented by $_{a/b} = \left[\phi(x_1^{a/b}), \ldots, \phi(x_{n/m}^{a/b})\right]$ and $_{a/b} = \left[\psi(z_1^{a/b}), \ldots, \psi(z_{n/m}^{a/b})\right]$. Consequently, as illustrated in [117], the empirical cross-covariance matrices are denoted by $\hat{A}_{XZ}^a = \frac{1}{n}\ {}_aJ_n\ {}_a^{>}$ and $\hat{A}_{XZ}^b = \frac{1}{m}\ {}_bJ_m\ {}_b^{>}$, with $J_n = (I_n\ \ \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^{>})$ as the centering matrix of size $n\ \ n$, I_n as the identity matrix, and $\mathbf{1}_n$ as a vector of ones of dimension n (similarly for J_m, I_m, and $\mathbf{1}_m$). The covariance matrices $\hat{A}_{XX}^{a/b}$ and $\hat{A}_{ZZ}^{a/b}$ are defined

in a similar fashion. Moreover, the empirical conditional covariance is defined as

$$\hat{A}^{a/b}_{XXjZ} = \hat{A}^{a/b}_{XX} \quad \hat{A}^{a/b}_{XZ} \left(\hat{A}^{a/b}_{ZZ} + \epsilon I\right)^{-1} \hat{A}^{a/b}_{ZX}, \tag{4.2}$$

where ϵI acts as a regularizer to the $\hat{A}_{ZZ}$ matrix and $\epsilon > 0$ is the regularization factor. The regularization is done due to the rank deficiency of the matrix $\hat{A}_{ZZ}$. We denote the matrices

$$Q_a \triangleq I_n \quad \frac{1}{n\epsilon}\left[M^a_Z \quad M^a_Z \left(M^a_Z + \epsilon n I_n\right)^{-1} M^a_Z\right] \tag{4.3}$$

$$Q_b \triangleq I_m \quad \frac{1}{m\epsilon}\left[M^b_Z \quad M^b_Z \left(M^b_Z + \epsilon m I_m\right)^{-1} M^b_Z\right] \tag{4.4}$$

where the centralized kernel matrices are defined as $M^a_Z = J_n K^a_{ZZ} J_n$ and $M^b_Z = J_m K^b_{ZZ} J_m$. Using the Cholesky decomposition [123] $Q_{a/b} = S_{a/b} S^T_{a/b}$, where Q is a positive-definite matrix [117] and S is the lower-triangular matrix obtained from the decomposition. Accordingly, we can reformulate the conditional covariance operator $\hat{A}^a_{XXjZ}$ as

$$\hat{A}^a_{XXjZ} = \frac{1}{n} \; {}_a J_n S_a \left({}_a J_n S_a\right)^T, \tag{4.5}$$

and respectively for $\hat{A}^b_{XXjZ}$. The empirical CKB metric is accordingly defined as

$$\begin{aligned} \hat{d}^2_{CKB}(P_{XjZ=z_a}, P_{XjZ=z_b}) &= \hat{d}^2_{CKB}(\hat{A}^a_{XXjZ}, \hat{A}^b_{XXjZ}) \\ &= \epsilon \, \mathrm{tr}\left[M^a_X \left(\epsilon n \mathbf{I}_n + M^a_Z\right)^{-1}\right] + \epsilon \, \mathrm{tr}\left[M^b_X \left(\epsilon m \mathbf{I}_m + M^b_Z\right)^{-1}\right] \\ &\quad - \frac{2}{m \quad n} \left\| (J_m S_b)^T K^{ba}_{XX} (J_n S_a) \right\|, \end{aligned} \tag{4.6}$$

where $\| \cdot \|$ is the nuclear norm. The empirical CKB metric is differentiable and highly suitable for usage as a loss function.

Total Loss for Deepfake Detection.

Based on Equation 4.6, we define $L_{FairAlign}$ as

$$L_{FairAlign} = \sum_{8(z_i, z_j) 2 Z} \hat{d}^2_{CKB}(P_{XjZ=z_i}, P_{XjZ=z_j}). \tag{4.7}$$

Additionally, we use a binary cross-entropy loss, L_{ce} for supervising the deepfake detector to discriminate between real and fake samples, defined as

$$L_{ce} = \quad \frac{1}{} \sum_{j=1} [y_j \log(\hat{y}_j) + (1 \quad y_j) \log(1 \quad \hat{y}_j)] . \tag{4.8}$$

Here, denotes the total count of samples in the batch. Finally, we define the total loss as

$$L = L_{ce} + \lambda \quad L_{FairAlign} \tag{4.9}$$

where λ is a hyperparameter to control the contribution of the CKB term.

4.1.3 Skin Tone Fairness Enhancement

As indicated earlier, in addition to a fair deepfake detection solution, we aim to detect skin tone. To this end, we first perform face detection using MobileFaceNet backbone [124] along with the ArcFace loss [47]. We then employ the U-Net model presented in [48] to segment skin regions from the extracted facial image. Next, we compute the average color of the facial skin pixels to obtain the overall skin tone. Finally, we intend to use a standard definition for characterizing the estimated skin tone. To do so, we use the Monk Skin Tone (MST) scale [49]. Therefore, to identify the corresponding tone from the MST scale, we compute the closest neighbour based on the Euclidean norm between our measured average tone and the tones in the MST scale.

4.1.4 Fairness-Accuracy Trade-o Assessment

Finally, the third goal of our work is to conduct a comprehensive analysis of fairness-accuracy trade-off in the context of deepfake detection. Multiple studies have previously indicated the presence of an intrinsic trade-off between fairness and accuracy [34, 33, 37, 36], although not in the area of deepfake detection. To this end, we employ two

metrics to characterize this trade-off in this context for the first time: (1) Fairea: The Fairea approach [5] first assesses how a model's predictions would change if it were less biased. This is done by manipulating the model's predictions to reflect a range of hypothetical scenarios from slightly to fully unbiased, which is referred to as 'mutation' in [5]. The range of mutation can be from 10% to 100%, with 10% increments at each step. These adjusted predictions create a spectrum of potential fairness values within the model, referred to as the 'baseline'. This baseline, along with the coordinates of an arbitrary bias-mitigation method, form an enclosed region whose area quantifies the trade-off. When evaluating two bias mitigation methods, the one with the larger area is considered to have achieved a better fairness-accuracy trade-off. We illustrate this approach in Figure 4.1. (2) Harmonic Mean: HM takes into account both accuracy and fairness in the form of $HM = \frac{2A \ F}{A+F}$, where A represents accuracy and F stands for fairness. We apply the same rationale as various works that use F1 score [50], wherein a harmonic mean formulation has been used to balance two diverging objectives [51].

4.2 Experiments

4.2.1 Experiment Setup

Datasets. We conduct all the experiments based on two popular datasets, **FF++** [46], **CelebDF** [1]. FF++ comprises 1000 Baseline and 4000 forged videos with several visual quality levels, raw (no compression), high quality, and low quality. CelebDF contains 590 real and 5639 fake videos. Since both the FF++ and CelebDF provide only the video-level labels, we sample frames out of these videos using FFMPEG [97] and perform facial cropping on these frames using the ArcFace detection model [47].

Evaluation metrics. To comprehensively assess fairness, we employ five bias metrics.

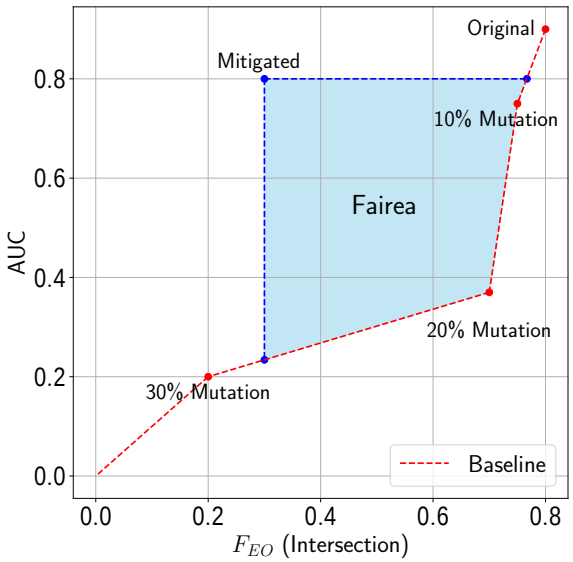


Figure 4.1: Schematic of the Fairea [5] trade-off evaluation metric.

First, following [125, 3], we use the maximum difference in False Positive Rate (FPR) gap, equal FPR, and equal odds, denoted by G_{FPR}, F_{FPR}, and F_{EO} respectively. These metrics are defined as

$$G_{\mathsf{FPR}} := \max_{8z_i,z_j 2 Z} \left|\mathsf{FPR}_{z_i} \quad \mathsf{FPR}_{z_j}\right|, \tag{4.10}$$

$$F_{\mathsf{FPR}} := \sum_{z_i 2 Z} \left| \frac{\sum_{i=1}^{n} \mathbb{1}_{[\hat{Y}_i=1, Z=z_i, Y_i=0]}}{\sum_{i=1}^{n} \mathbb{1}_{[Z=z_i, Y_i=0]}} \quad \frac{\sum_{i=1}^{n} \mathbb{1}_{[\hat{Y}_i=1, Y_i=0]}}{\sum_{i=1}^{n} \mathbb{1}_{[Y_i=0]}} \right|, \tag{4.11}$$

$$F_{\mathsf{EO}} := \sum_{z_i 2 Z} \sum_{q=0}^{1} \left| \frac{\sum_{i=1}^{n} \mathbb{1}_{[\hat{Y}_i=1, Z=z_i, Y_i=q]}}{\sum_{i=1}^{n} \mathbb{1}_{[Z=z_i, Y_i=q]}} \quad \frac{\sum_{i=1}^{n} \mathbb{1}_{[\hat{Y}_i=1, Y_i=q]}}{\sum_{i=1}^{n} \mathbb{1}_{[Y_i=q]}} \right|, \tag{4.12}$$

where FPR_{z_i} represents the FPR scores of group z_i, $\mathbb{1}_{[]}$ denotes the indicator function, and q represents the forgery label (1 is real and 0 is fake). Also note that in the special case when Z corresponds to gender, which means $\zeta = 2$, the metrics G_{FPR} and F_{FPR} will return the same value. Additionally, following [126], we use Demographic Parity Difference (DPD) and Demographic Parity Ratio (DPR), which are formulated as

$$\mathsf{DPD} = \max_{z_i 2 Z} P(\hat{Y} = 1 j Z = z_i) \quad \min_{z_i 2 Z} P(\hat{Y} = 1 j Z = z_i), \tag{4.13}$$

$$\text{DPR} = \frac{\min_{z_i \in Z} P(\hat{Y} = 1 | Z = z_i)}{\max_{z_i \in Z} P(\hat{Y} = 1 | Z = z_i)}. \tag{4.14}$$

DPD measures the disparity in positive outcomes across different groups, with an ideal value of 0 indicating no disparity. Conversely, DPR assesses the relative disparity, with an ideal value of 1 suggesting equal positive outcome rates across all groups.

Finally, to assess the performance of different deepfake detectors, we utilize four widely-used metrics [3]: AUC, FPR, True Positive Rate (TPR), and top-1 accuracy (ACC).

Baseline methods. To validate the efficacy of our proposed loss term, $L_{FairAlign}$, we integrate it into the training process of 5 state-of-the-art deepfake detector backbones: RECCE [25], MASDT [23], AltFreezing [24], EfficientNet-B3 [127], and EfficientNet-B4 [2]. The objective is to assess the impact of $L_{FairAlign}$ on the fairness of these models. For a thorough analysis, we benchmark our approach against 4 state-of-the-art bias-mitigating methods: DAG-FDD [3], DAW-FDD [3], DRO-χ^2 [128], and MMD loss [129]. Furthermore, to establish baseline performances, each model is also trained without any fairness-enhancing module.

Implementation details. All experiments are conducted using the PyTorch framework [130] on up to 8 NVIDIA A100 PCIE GPUs. We train all methods using the AdamW optimizer [131] with a batch size of 32, a maximum of 100 epochs, and a learning rate of 0.001. The optimizer employs first and second momentum decays of 0.9 and 0.999, respectively. Additionally, we use a weight decay of 0.01 to refine the training process. The learning rate is adjusted using a step scheduler, which decreases the learning rate by a factor of 0.5 every 5 epochs. The video frame input size is set to 380 pixels, with training augmentations including resizing, normalization and horizontal flipping. For the face detection process mentioned in Section 4.1.3, we use

the InsightFace toolkit [132]. For the ϵ used in Equation 4.2, we adhere to 0.01 as per the design choice outlined in [117].

4.2.2 Results

Performance. We present a thorough analysis of the performance of our approach on the FF++ and CelebDF datasets in Tables 4.1 and 4.2 respectively. First, evaluating the average results of FairAlign against the baselines demonstrates that our approach substantially promotes fairness across all three metrics for both gender and skin tone, as well as the intersection of the two. Comparing the performance of our method averaged across backbones, against other fairness-promoting solutions, we observe that our method generally achieves the best fairness scores, with exceptions in the intersection group, where MMD achieves marginally higher scores for F_{FPR} and F_{EO} on the FF++ dataset. Similarly, MMD obtains slightly better results in the intersection group with G_{FPR} on CelebDF.

Delving deeper into the results, we observe that on FF++, our method FairAlign applied to EfficientNet-B3 achieves the lowest G_{FPR} (and similarly F_{FPR}) of 0.16% in gender, in a tie with MMD when applied to the RECCE network. Similarly in CelebDF, FairAlign with MASDT achieves the lowest G_{FPR} (and F_{FPR}) in gender with a score of 0.19%. For F_{EO} our method obtains the best performance on gender when coupled with EfficientNet-B4 on FF++, while on CelebDF, MMD results in the best outcome with the same backbone. For skin tone, on FF++, MMD achieves the best performance for G_{FPR} along with the AltFreezing deepfake detection method. However, for F_{FPR} and F_{EO}, our method obtains the best results with EfficientNet-B3 and MASDT backbones respectively. On CelebDF, our method consistently achieves

Table 4.1: Results on FF++. Best results in each column are in **bold** and second-best results are underlined.

Methods	Backbones	Bias Metrics (%)↓									Detection Metrics (%)			
		Gender			Skin Tone			Intersection			Overall			
		G_{FPR}	F_{FPR}	F_{EO}	G_{FPR}	F_{FPR}	F_{EO}	G_{FPR}	F_{FPR}	F_{EO}	AUC↑	FPR↓	TPR↑	ACC↑
Baseline	EfficientNet-B3 [2]	1.97	1.97	8.15	11.05	10.86	32.19	14.38	22.65	44.13	94.72	20.25	97.21	94.09
	RECCE [25]	1.27	1.27	9.14	18.81	29.65	25.07	30.26	69.38	82.34	98.05	21.20	98.21	94.74
	EfficientNet-B4 [2]	1.97	1.97	7.85	11.56	10.86	34.12	23.89	20.56	42.13	95.91	20.25	97.21	94.09
	MASDT [23]	1.38	1.38	19.71	14.64	11.89	11.39	18.07	14.32	41.46	96.21	3.65	97.13	97.60
	AltFreezing [24]	2.82	2.82	10.54	18.37	9.85	18.09	12.02	33.74	40.74	97.84	8.42	96.27	98.10
	Average	1.88	1.88	11.08	14.89	14.62	24.17	19.72	32.13	50.16	96.55	14.75	97.21	95.72
DAG-FDD [3]	EfficientNet-B3 [2]	0.67	0.67	5.36	11.48	9.58	13.50	12.87	19.34	46.08	97.01	8.40	92.87	92.65
	RECCE [25]	0.75	0.75	5.71	14.68	19.41	19.33	25.40	38.17	76.24	98.33	12.01	96.80	95.23
	EfficientNet-B4 [2]	0.61	0.61	4.86	13.81	16.85	20.55	17.50	30.88	52.63	94.15	21.58	95.60	92.92
	MASDT [23]	0.58	0.58	18.70	10.60	9.84	7.36	16.04	14.27	30.39	96.95	5.67	97.63	98.29
	AltFreezing [24]	2.64	2.64	10.01	11.18	9.27	17.42	11.30	32.62	37.69	97.10	7.87	95.59	97.44
	Average	1.05	1.05	8.93	12.35	12.99	15.63	16.62	27.06	48.61	96.71	11.11	95.70	95.31
DAW-FDD [3]	EfficientNet-B3 [2]	0.34	0.34	6.53	6.79	11.67	12.63	8.43	12.57	43.72	95.96	8.22	91.43	91.49
	RECCE [25]	0.45	0.45	7.95	6.99	9.96	13.95	23.54	25.44	54.95	98.35	8.15	94.59	94.10
	EfficientNet-B4 [2]	0.55	0.55	3.71	13.65	17.35	20.30	15.34	36.00	56.19	90.44	2.00	96.91	95.40
	MASDT [23]	0.45	0.45	17.71	8.62	8.86	9.87	12.05	13.29	33.42	96.86	5.41	97.81	98.13
	AltFreezing [24]	2.72	2.72	10.03	6.20	9.39	17.71	11.79	32.24	37.54	97.64	8.19	95.91	97.81
	Average	0.90	0.90	9.19	8.45	11.45	14.89	14.23	23.90	45.16	95.85	6.39	95.33	95.39
DRO_{χ^2} [133]	EfficientNet-B3 [2]	0.23	0.23	4.42	4.71	6.58	12.38	6.30	12.32	42.85	94.37	8.06	89.60	89.66
	RECCE [25]	0.33	0.33	5.46	6.15	9.08	11.71	20.27	24.97	64.89	98.32	7.99	96.48	95.98
	EfficientNet-B4 [2]	0.54	0.54	3.64	11.40	17.00	19.89	15.03	35.28	51.11	93.90	1.96	98.10	96.77
	MASDT [23]	0.64	0.64	15.68	6.51	9.04	7.24	12.93	13.04	41.15	98.29	3.22	**98.96**	97.37
	AltFreezing [24]	2.67	2.67	9.83	5.10	9.20	17.36	11.55	31.60	26.19	**98.86**	8.03	97.22	**98.75**
	Average	0.88	0.88	7.81	6.77	10.18	13.72	13.22	23.44	45.24	96.75	5.85	96.07	95.71
MMD [129]	EfficientNet-B3 [2]	0.35	0.35	6.65	5.88	4.78	12.91	5.62	12.84	44.95	93.57	8.49	94.11	94.01
	RECCE [25]	**0.16**	**0.16**	6.19	7.14	10.18	9.19	18.98	20.08	66.46	96.96	8.42	93.65	93.53
	EfficientNet-B4 [2]	0.39	0.39	2.59	8.72	13.03	15.05	10.34	25.47	39.50	92.64	1.58	97.08	95.60
	MASDT [23]	0.32	0.32	14.20	5.89	6.94	5.87	10.20	**11.43**	29.18	97.04	**1.23**	98.02	98.32
	AltFreezing [24]	2.02	2.02	7.06	**3.74**	7.18	12.61	7.78	22.30	**21.01**	97.82	5.78	96.14	98.05
	Average	0.65	0.65	7.34	6.27	8.42	11.13	10.58	18.42	40.22	95.61	5.10	95.80	95.90
FairAlign (Ours)	EfficientNet-B3 [2]	**0.16**	**0.16**	5.78	3.97	**3.59**	10.15	**4.74**	13.09	45.74	92.87	8.59	95.19	93.66
	RECCE [25]	0.19	0.19	4.98	6.02	10.03	10.50	14.21	21.58	57.54	96.74	8.53	93.60	93.03
	EfficientNet-B4 [2]	0.39	0.39	**2.07**	6.54	12.29	11.79	8.10	22.95	45.08	91.78	1.47	97.24	95.75
	MASDT [23]	0.29	0.29	12.00	4.01	5.28	**5.38**	8.26	11.72	23.07	97.23	1.67	98.11	98.46
	AltFreezing [24]	1.74	1.74	6.03	4.10	7.05	10.17	5.90	23.78	31.23	97.97	5.36	96.26	98.13
	Average	0.55	0.55	6.17	4.93	7.65	9.60	8.24	18.62	40.53	95.32	5.12	96.08	95.81

the lowest bias when coupled with MASDT, EfficientNet-B3, and MASDT, for the three metrics respectively. For the intersection of the two (gender and skin tone), on FF++, our approach outperforms the others based on G_{FPR} using EfficientNet-B3, while MMD shows better performance on F_{FPR} and F_{EO} using MASDT and AltFreezing respectively. On CelebDF, MMD achieves better intersection results based

Table 4.2: Results on CelebDF. Best results in each column are in **bold** and second-best results are underlined.

Methods	Backbones	Bias Metrics (%)↓									Detection Metrics (%)			
		Gender			Skin Tone			Intersection			Overall			
		G_{FPR}	F_{FPR}	F_{EO}	G_{FPR}	F_{FPR}	F_{EO}	G_{FPR}	F_{FPR}	F_{EO}	AUC↑	FPR↓	TPR↑	ACC↑
Baseline	EfficientNet-B3 [2]	3.82	3.82	7.54	11.37	9.85	14.09	21.25	45.47	70.74	94.13	9.25	92.27	95.04
	RECCE [25]	2.71	2.71	3.14	18.81	27.65	30.07	30.26	67.38	80.34	94.05	12.20	95.21	95.74
	EfficientNet-B4 [2]	1.21	1.21	5.15	10.05	20.86	34.12	13.38	22.65	40.13	93.91	10.25	97.21	93.09
	MASDT [23]	0.48	0.48	3.71	8.64	10.89	9.39	12.07	15.32	16.46	96.61	5.05	94.13	95.60
	AltFreezing [24]	1.71	1.71	8.52	7.25	12.63	22.65	23.83	42.65	**40.13**	95.91	8.25	94.11	96.09
	Average	1.99	1.99	5.61	11.22	16.38	22.06	20.16	38.69	49.56	94.92	9.00	94.59	95.11
DAG-FDD [3]	EfficientNet-B3 [2]	3.65	3.65	8.01	9.18	9.27	11.42	12.30	42.62	67.96	93.02	10.72	92.59	95.20
	RECCE [25]	1.45	1.45	3.71	12.68	17.41	19.33	15.40	36.17	64.24	94.33	10.01	95.80	93.23
	EfficientNet-B4 [2]	0.61	0.61	4.86	12.81	16.85	20.55	17.50	30.88	52.63	92.15	11.58	95.60	94.92
	MASDT [23]	0.58	0.58	3.70	8.60	10.84	9.36	12.04	15.27	16.39	95.95	5.55	93.63	96.29
	AltFreezing [24]	0.97	0.97	6.63	7.81	11.83	18.24	20.81	39.34	46.08	97.20	7.02	94.75	96.65
	Average	1.45	1.45	5.38	10.22	13.24	15.78	15.61	32.86	49.46	94.53	8.98	94.47	95.26
DAW-FDD [3]	EfficientNet-B3 [2]	3.56	3.56	7.03	7.20	6.39	9.18	11.79	42.24	67.45	93.41	9.98	92.91	95.17
	RECCE [25]	0.95	0.95	4.75	6.99	7.96	11.95	13.54	23.44	62.95	92.35	12.15	94.59	92.10
	EfficientNet-B4 [2]	0.45	0.45	3.71	12.65	17.35	20.30	15.34	36.00	54.19	91.44	12.00	96.91	93.40
	MASDT [23]	0.38	0.38	3.71	8.62	10.86	9.37	12.05	15.29	16.02	95.56	4.55	94.81	96.13
	AltFreezing [24]	0.43	0.43	5.34	5.96	10.62	17.82	11.53	37.57	43.72	96.30	7.26	93.32	97.49
	Average	1.15	1.15	4.91	8.28	10.64	13.72	12.85	30.91	48.87	93.81	9.19	94.51	94.86
DRO_{χ^2} [133]	EfficientNet-B3 [2]	3.40	3.40	6.83	8.10	9.20	12.36	11.55	39.60	66.95	93.63	6.32	93.22	96.62
	RECCE [25]	0.84	0.84	4.66	6.85	7.80	10.71	13.27	22.97	71.89	91.32	10.99	93.48	90.98
	EfficientNet-B4 [2]	0.44	0.44	3.64	12.40	17.00	19.89	15.03	35.28	53.11	90.90	11.96	**98.10**	94.77
	MASDT [23]	0.22	0.22	3.68	8.51	10.70	9.24	9.93	15.04	16.15	95.29	5.54	93.96	96.37
	AltFreezing [24]	0.33	0.33	5.23	6.16	10.83	17.91	11.14	32.32	42.85	94.37	6.62	90.01	95.66
	Average	1.05	1.05	4.81	8.40	11.11	14.02	12.18	29.04	50.19	93.10	8.29	93.75	94.88
MMD [129]	EfficientNet-B3 [2]	3.02	3.02	5.06	6.74	11.18	10.61	**7.78**	28.30	51.17	94.20	6.88	92.14	95.25
	RECCE [25]	0.76	0.76	4.89	7.14	8.18	12.19	13.98	24.08	64.46	92.96	10.42	92.65	91.53
	EfficientNet-B4 [2]	0.29	0.29	**2.59**	8.72	13.03	15.05	10.34	25.47	37.50	92.64	12.58	97.08	94.60
	MASDT [23]	0.31	0.31	3.20	6.89	7.94	7.87	9.20	**12.43**	14.18	96.04	**4.40**	92.02	97.32
	AltFreezing [24]	0.52	0.52	4.54	6.85	10.97	18.83	11.32	28.84	44.95	94.57	6.91	96.15	97.01
	Average	0.98	0.98	4.06	7.27	10.26	12.91	10.52	23.82	42.45	94.08	8.24	94.01	95.14
FairAlign (Ours)	EfficientNet-B3 [2]	2.67	2.67	4.39	5.19	**5.54**	7.77	9.08	23.78	41.39	94.71	5.61	93.26	96.33
	RECCE [25]	0.86	0.86	4.98	7.32	8.30	10.50	14.21	24.58	55.54	93.74	7.53	91.60	94.03
	EfficientNet-B4 [2]	0.29	0.29	2.77	6.54	12.99	11.79	12.10	22.95	33.08	93.78	11.47	97.24	94.75
	MASDT [23]	**0.19**	**0.19**	3.00	**5.01**	7.88	**7.38**	9.16	12.72	**12.07**	96.23	5.40	92.11	97.46
	AltFreezing [24]	0.29	0.29	4.86	5.78	10.17	16.95	11.45	31.09	45.74	**97.87**	6.28	96.94	**97.66**
	Average	0.86	0.86	4.00	5.97	8.98	10.88	11.20	23.02	37.56	95.27	7.26	94.23	96.05

on G_{FPR} and F_{FPR} using EfficientNet-B3 and MASDT, while ours outperforms other methods based on F_{EO} using the MASDT method.

A similar trend is observed for the additional two metrics, DPD and DPR, presented in Tables 4.3 and 4.4. Our method's performance, averaged across backbones, achieves the lowest DPD and the highest DPR across all groups based on gender, skin tone, and

Table 4.3: DPR and DPD results on FF++. Best results in each column are in **bold** and second-best results are underlined.

Methods	Backbones	Bias Metrics					
		Gender		Skin Tone		Intersection	
		DPD ↓	DPR ↑	DPD ↓	DPR ↑	DPD ↓	DPR ↑
Baseline	EfficientNet-B3 [2]	0.29	0.79	0.38	0.64	0.55	0.56
	RECCE [25]	0.39	0.77	0.39	0.67	0.49	0.72
	EfficientNet-B4 [2]	0.49	0.69	0.46	0.73	0.53	0.66
	MASDT [23]	0.47	0.77	0.54	0.68	0.42	0.52
	AltFreezing [24]	0.38	0.67	0.47	0.73	0.54	0.65
	Average	0.40	0.74	0.45	0.69	0.51	0.62
DAG-FDD [3]	EfficientNet-B3 [2]	0.20	0.82	0.22	0.85	0.32	0.88
	RECCE [25]	0.22	0.80	0.41	0.89	0.36	0.85
	EfficientNet-B4 [2]	0.30	0.72	0.27	0.85	0.24	0.88
	MASDT [23]	0.32	0.89	0.35	0.79	0.33	0.74
	AltFreezing [24]	0.29	0.89	0.48	0.84	0.43	0.56
	Average	0.27	0.82	0.35	0.84	0.34	0.78
DAW-FDD [3]	EfficientNet-B3 [2]	0.21	0.88	0.30	0.89	0.39	0.91
	RECCE [25]	0.24	0.84	0.41	0.92	0.20	0.87
	EfficientNet-B4 [2]	0.21	0.84	0.28	0.87	0.54	0.90
	MASDT [23]	0.28	0.92	0.46	0.83	0.17	0.86
	AltFreezing [24]	0.21	0.92	0.39	0.87	0.37	0.79
	Average	0.23	0.88	0.37	0.88	0.33	0.87
DRO_{χ^2} [133]	EfficientNet-B3 [2]	0.29	0.82	0.29	0.85	0.40	0.87
	RECCE [25]	0.13	0.89	0.20	0.88	0.28	0.83
	EfficientNet-B4 [2]	0.39	0.70	0.36	0.84	0.37	0.87
	MASDT [23]	0.17	0.88	0.34	0.79	0.35	0.63
	AltFreezing [24]	0.18	0.88	0.27	0.84	0.30	0.76
	Average	0.23	0.83	0.29	0.84	0.34	0.79
MMD [129]	EfficientNet-B3 [2]	0.20	0.85	0.19	0.87	0.35	0.89
	RECCE [25]	0.16	0.90	0.31	0.89	0.26	0.86
	EfficientNet-B4 [2]	0.40	0.61	0.47	0.85	0.44	0.89
	MASDT [23]	0.28	0.90	0.48	0.81	0.27	0.74
	AltFreezing [24]	0.14	0.90	0.28	0.85	0.24	0.83
	Average	0.24	0.83	0.35	0.85	0.31	0.84
FairAlign (Ours)	EfficientNet-B3 [2]	0.17	0.90	**0.11**	**0.97**	0.23	**0.93**
	RECCE [25]	0.12	0.94	0.32	0.93	0.20	0.90
	EfficientNet-B4 [2]	0.12	0.76	0.29	0.87	0.39	0.92
	MASDT [23]	0.18	0.93	0.37	0.83	0.18	0.85
	AltFreezing [24]	**0.11**	**0.94**	0.29	0.89	**0.15**	0.89
	Average	0.14	0.89	0.28	0.90	0.23	0.90

intersectional categories. Among the backbones, the AltFreezing detector consistently achieves either the best or the second-best position for gender metrics in both the FF++ and CelebDF benchmarks. Similarly, EfficientNet-B3 tends to achieve the best

Table 4.4: DPR and DPD results on CelebDF. Best results in each column are in **bold** and second-best results are underlined.

Methods	Backbones	Bias Metrics					
		Gender		Skin Tone		Intersection	
		DPD ↓	DPR ↑	DPD ↓	DPR ↑	DPD ↓	DPR ↑
Baseline	EfficientNet-B3 [2]	0.39	0.89	0.28	0.84	0.44	0.86
	RECCE [25]	0.36	0.77	0.39	0.67	0.45	0.62
	EfficientNet-B4 [2]	0.38	0.72	0.36	0.83	0.43	0.62
	MASDT [23]	0.37	0.63	0.44	0.78	0.42	0.62
	AltFreezing [24]	0.38	0.87	0.31	0.83	0.34	0.85
	Average	0.38	0.78	0.36	0.79	0.42	0.71
DAG-FDD [3]	EfficientNet-B3 [2]	0.40	0.92	0.28	0.85	0.45	0.88
	RECCE [25]	0.22	0.79	0.41	0.69	0.56	0.65
	EfficientNet-B4 [2]	0.28	0.73	0.37	0.85	0.64	0.58
	MASDT [23]	0.27	0.79	0.32	0.79	0.33	0.64
	AltFreezing [24]	0.29	0.89	0.38	0.84	0.44	0.86
	Average	0.29	0.82	0.35	0.80	0.48	0.72
DAW-FDD [3]	EfficientNet-B3 [2]	0.41	0.94	0.30	0.89	0.66	0.91
	RECCE [25]	0.24	0.73	0.41	0.62	0.36	0.67
	EfficientNet-B4 [2]	0.29	0.77	0.38	0.87	0.34	0.60
	MASDT [23]	0.28	0.62	0.27	0.83	0.53	0.66
	AltFreezing [24]	0.25	0.92	0.39	0.87	0.25	0.89
	Average	0.29	0.80	0.35	0.82	0.43	0.75
DRO_{χ^2} [133]	EfficientNet-B3 [2]	0.39	0.90	0.28	0.85	0.54	0.87
	RECCE [25]	0.31	0.78	0.40	0.68	0.45	0.63
	EfficientNet-B4 [2]	0.35	0.73	0.36	0.84	0.43	0.57
	MASDT [23]	0.17	0.88	0.34	0.79	0.32	0.63
	AltFreezing [24]	0.18	0.88	0.35	0.84	0.34	0.86
	Average	0.28	0.83	0.35	0.80	0.42	0.71
MMD [129]	EfficientNet-B3 [2]	0.40	0.91	0.29	0.87	0.25	0.89
	RECCE [25]	0.19	0.87	0.41	0.69	0.56	0.66
	EfficientNet-B4 [2]	0.29	0.75	0.37	0.85	0.24	0.79
	MASDT [23]	0.10	0.90	0.30	0.81	0.22	0.64
	AltFreezing [24]	0.15	0.90	0.38	0.85	0.30	0.88
	Average	0.23	0.87	0.35	0.81	0.31	0.77
FairAlign (Ours)	EfficientNet-B3 [2]	0.27	**0.96**	0.29	**0.97**	0.35	**0.93**
	RECCE [25]	0.20	0.88	0.42	0.63	0.37	0.67
	EfficientNet-B4 [2]	0.24	0.77	0.29	0.87	**0.20**	0.72
	MASDT [23]	**0.08**	0.93	**0.22**	0.89	0.19	0.65
	AltFreezing [24]	0.10	0.94	0.39	0.89	0.25	0.89
	Average	0.18	0.90	0.32	0.85	0.27	0.77

scores for skin tone and intersectional group metrics in both datasets.

Comparing the bias metrics for gender with those of skin tone, we notice considerably higher, i.e., more biased, values for skin tone. We believe this due to two main

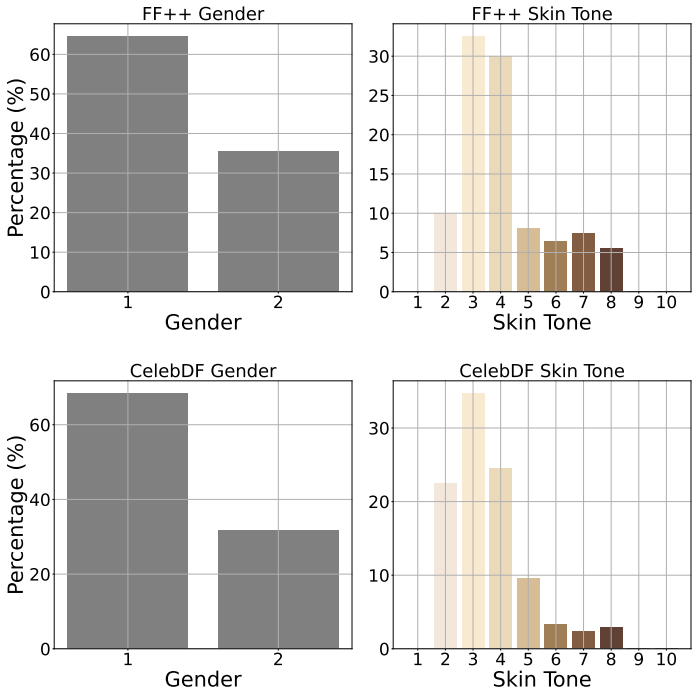
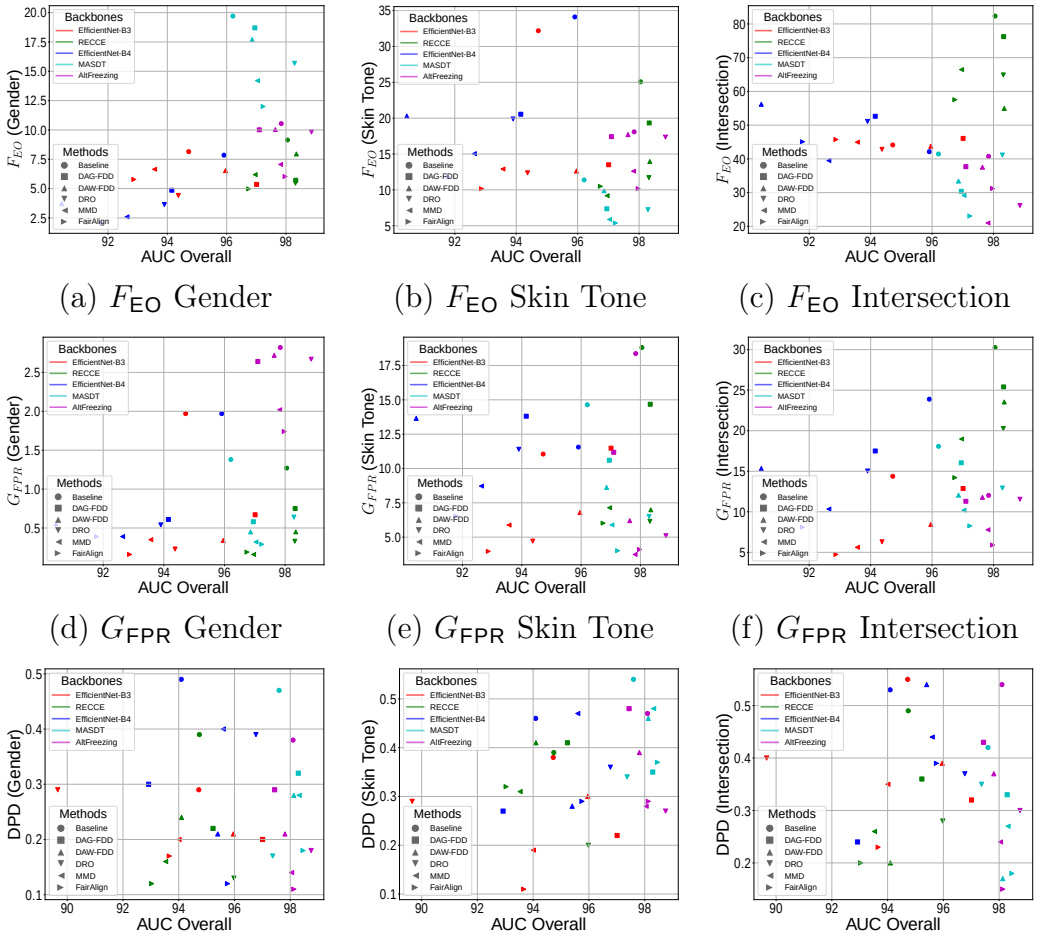


Figure 4.2: Distribution of genders and skin tones in the FF++ and CelebDF datasets.

reasons. First, gender is currently defined as a binary class in the dataset, whereas our definition of skin tones consist of 10 unique classes. This difference between the number of classes is an important reason behind skin tone showing more bias as measured by the metrics. The second reason could be that skin tone is inherently more challenging in terms of bias mitigation, for instance due to the heavily imbalanced nature of the datasets in this regard. We present the distributions for gender and measured skin tones in Figure 4.2, where we observe a less balanced, i.e., long-tailed distribution for skin tones.

Finally, our method demonstrates strong deepfake detection performances across all four metrics for both datasets. On the FF++ dataset, FairAlign obtains very competitive results, while on CelebDF, it generally achieves better performances with



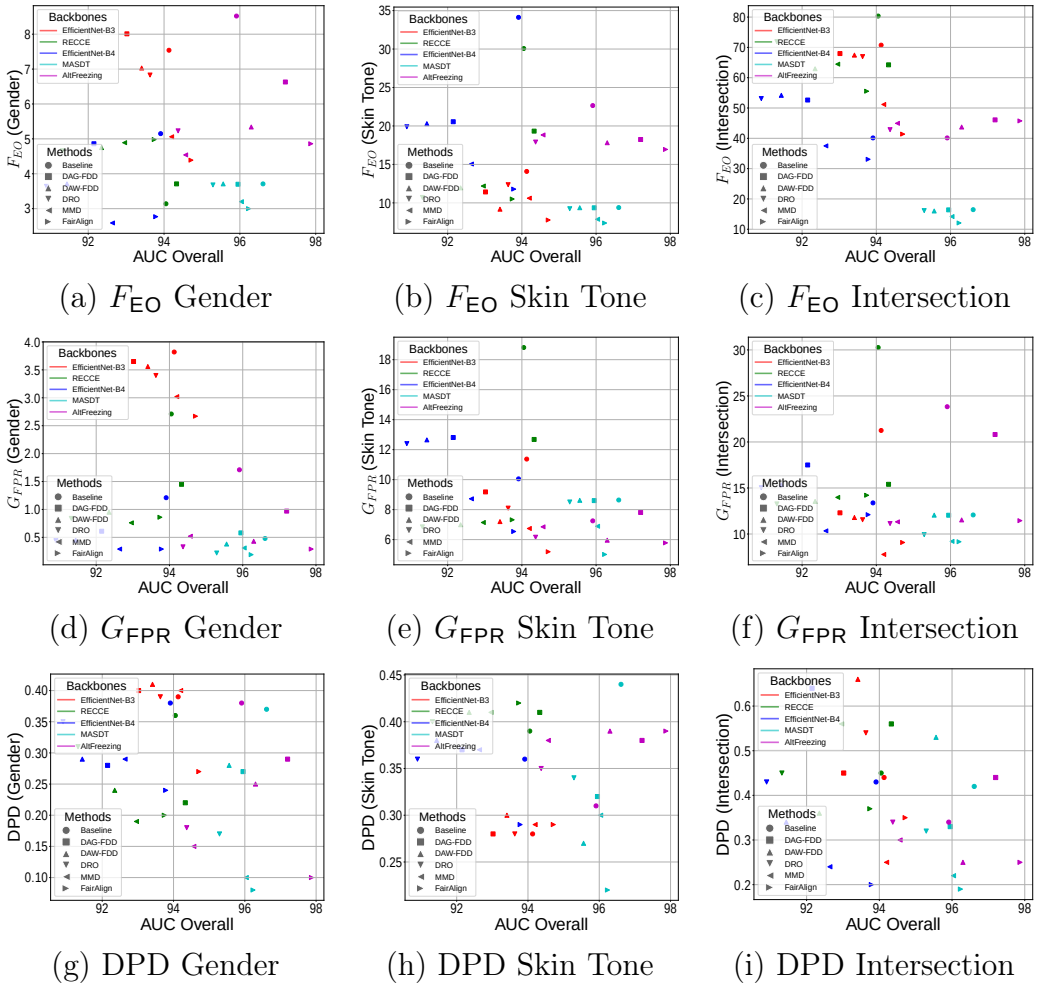
(a) F_{EO} Gender (b) F_{EO} Skin Tone (c) F_{EO} Intersection

(d) G_{FPR} Gender (e) G_{FPR} Skin Tone (f) G_{FPR} Intersection

(g) DPD Gender (h) DPD Skin Tone (i) DPD Intersection

Figure 4.3: Fairness vs AUC plots for all detectors and loss techniques on the FF++ dataset.

respect to others. We visualize all the results in Figure 4.3 and Figure 4.4. A detailed comparison between both performance aspects (fairness alongside deepfake detection) from Tables 4.1 through 4.4 and Figures 4.3 and 4.4 highlights that drawing a high-level conclusion about the best fairness promoting approach remains complicated and nuanced when considering the deepfake detection results. This is especially the case on FF++ where the trade-off between fairness and performance seems more complex,



(a) F_{EO} Gender (b) F_{EO} Skin Tone (c) F_{EO} Intersection

(d) G_{FPR} Gender (e) G_{FPR} Skin Tone (f) G_{FPR} Intersection

(g) DPD Gender (h) DPD Skin Tone (i) DPD Intersection

Figure 4.4: Fairness vs AUC plots for all detectors and loss techniques on the CelebDF dataset.

further demonstrating the need for an in-depth trade-off analysis.

Trade-o Analysis. To perform an analysis on fairness-accuracy trade-off, we use $1/F_{\mathsf{EO}}$ following [3] to represent fairness performance for the intersection of gender and skin tone to capture a holistic view of both sensitive attributes. Moreover, following [24] we select AUC to represent the the deepfake detection performance of different methods. Using these metrics, we present Fairea and HM as discussed in

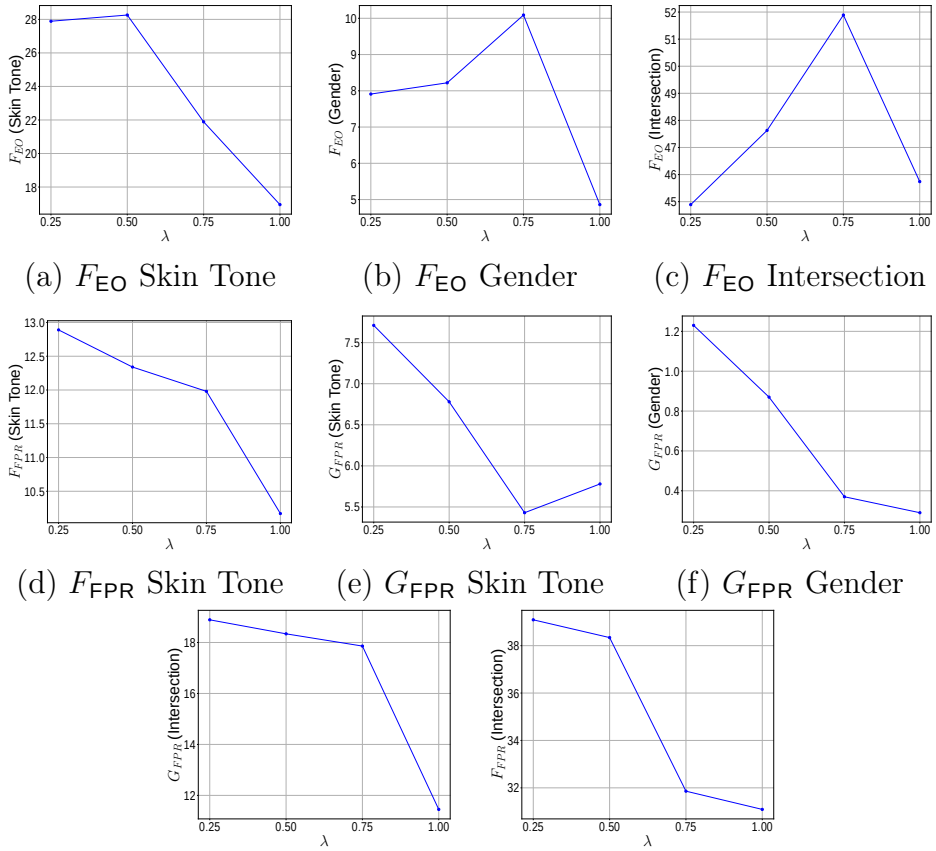
Table 4.5: Fairness-accuracy trade-off on the FF++ dataset.

Methods	Backbones									
	E□.-B3 [2]		RECCE [25]		E□.-B4[2]		MASDT[23]		AltFreezing [24]	
	Fairea↑	HM↑	Fairea↑	HM↑	Fairea↑	HM↑	Fairea↑	HM↑	Fairea↑	HM↑
DAG-FDD [3]	0.04	1.64	0.04	1.35	0.05	1.46	0.04	1.68	0.04	1.82
DAW-FDD [3]	0.05	1.55	0.06	1.27	0.05	1.50	0.05	1.55	0.03	1.68
DRO_{χ^2} [133]	0.06	1.49	0.06	1.46	0.04	1.48	0.04	1.49	0.05	1.77
MMD [129]	0.05	1.49	0.05	1.50	0.06	1.47	0.04	1.61	0.03	1.82
FairAlign (Ours)	0.07	1.80	0.05	1.64	0.04	1.50	0.05	1.73	0.06	1.91

Table 4.6: Fairness-accuracy trade-off on the CelebDF Dataset.

Methods	Backbones									
	E□.-B3 [2]		RECCE [25]		E□.-B4[2]		MASDT[23]		AltFreezing [24]	
	Fairea↑	HM↑	Fairea↑	HM↑	Fairea↑	HM↑	Fairea↑	HM↑	Fairea↑	HM↑
DAG-FDD [3]	0.03	1.34	0.02	1.12	0.02	1.26	0.03	1.49	0.03	1.42
DAW-FDD [3]	0.02	1.35	0.03	1.27	0.04	1.20	0.04	1.46	0.03	1.43
DRO_{χ^2} [133]	0.02	1.34	0.02	1.20	0.02	1.27	0.03	1.39	0.03	1.57
MMD [129]	0.03	1.31	0.03	1.17	0.03	1.36	0.03	1.51	0.03	1.62
FairAlign (Ours)	0.04	1.30	0.04	1.24	0.03	1.40	0.04	1.58	0.04	1.70

Section 4.1.4, and present the results in Tables 4.5 and 4.6 for FF++ and CelebDF respectively. We observe that FairAlign generally outperforms other bias-mitigation methods across different backbones when considering both fairness and accuracy. For instance, FairAlign achieves the highest scores on FF++ for EfficientNet-B3 and AltFreezing according to Fairea and HM respectively. On the CelebDF dataset, the best results are obtained by our method using all four backbones as per Fairea, while HM indicates the highest score using AltFreezing. An important observation from this analysis is that while theoretically both metrics (Fairea and HM) are capable of quantifying the fairness-accuracy trade-off, Fairea seems to produce less discriminatory outcomes. In contrast, HM generates a wider range of values, offering a more effective and discriminative means of capturing the trade-off.



(a) F_{EO} Skin Tone (b) F_{EO} Gender (c) F_{EO} Intersection

(d) F_{FPR} Skin Tone (e) G_{FPR} Skin Tone (f) G_{FPR} Gender

(g) G_{FPR} Intersection (h) F_{FPR} Intersection

Figure 4.5: Effect of tuning the λ hyperparameter on bias metrics for AltFreezing backbone trained on CelebDF dataset.

Impact of λ on Fairness. We further investigate the impact of the λ hyperparameter in Equation 4.9 on the bias metrics. Figure 4.5 illustrates the relationship between the bias metrics and various values of λ using the AltFreezing backbone on the CelebDF dataset. From this figure, we observe that for six out of the eight plots, $\lambda = 1$ results in the least amount of bias, while for F_{EO} for Intersection and G_{FPR} for Skin Tone, $\lambda = 0.25$ and $\lambda = 0.75$ are marginally better than 1. For consistency, we set $\lambda = 1$ throughout all the experiments in this work.

Computational Cost. Additionally, we assess the computational efficiency of our

Table 4.7: Average time per epoch for bias mitigating methods on CelebDF across multiple backbones on a single NVIDIA A100 GPU.

Methods	Time per epoch (s)
DAG-FDD [3]	109.71
DAW-FDD [3]	220.84
DRO$_{\chi^2}$[133]	187.32
MMD [129]	141.60
FairAlign (Ours)	167.65

method by comparing the average time per epoch for all involved bias-mitigating methods on the CelebDF dataset. We present the results in Table 4.7 where we observe that although FairAlign does not demonstrate the lowest time, it is able to achieve substantial performance gains with a reasonable computational overhead.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

In this thesis, we addressed two key problems in the realm of deepfake detection. The first problem centers on the considering temporal information in for detecting deepfake videos. The second problem is concerning bias in deepfake detection to improve fairness for gender and skin tone, especially without compromising deepfake detection accuracy. To address the first problem, we proposed an effective solution based on a masked autoencoding transformer architecture. To tackle the second problem, we proposed a plug-and-play loss term to promote fairness in deepfake detection across gender and skin tone groups, while considering fairness-accuracy trade-off. Following is a summary of our work in this thesis.

In Chapter 3, we introduced MASDT, a learning framework for enhanced deepfake detection. Our method consists of two components, spatial and temporal consistency learning. The model follows a sequential two-step process. Initially, it employs self-supervised pre-training where both spatial learning and temporal consistency learning components engage in data reconstruction. Spatial learning makes use of

a masked self-supervised auto-encoder to derive robust spatial features from partial facial images, while temporal consistency learning employs a similar auto-encoder to extract temporal features from partial optical flow fields. Subsequently, deepfake detection is executed through fine-tuning of the encoders of both learning components followed by simple score-level fusion. Various experiments on FF++ (LQ and HQ) and CelebDF datasets demonstrate that our approach outperforms state-of-the-art methods by effectively learning spatial and temporal information, resulting in enhanced classification performance.

In Chapter 4, we proposed FairAlign, a novel loss term designed to promote fairness in deepfake detection. Our method operates by aligning conditional embedding distributions within the higher-dimensional kernel space, thus reducing information related to sensitive attributes that could potentially bias the detection process. In addition to standard practice of bias reduction for gender, we implemented a simple yet effective pipeline to annotate and reduce bias in deepfake detection for the sensitive attribute of skin tone. Lastly, we performed a study on the fairness-accuracy trade-off in deepfake detection for the first time. Through various experiments on two commonly used public datasets, we demonstrated that FairAlign outperforms other bias-mitigation techniques while integrating smoothly with various deepfake detectors to improve fairness. The experiments demonstrated that not only FairAlign is highly effective in reducing gender and skin tone bias, but it does so while retaining strong deepfake detection performance.

5.2 Future Work

We identify several areas for potential future work. In the context of our work in Chapter 3, while the integration of temporal information through optical flow improves the detection performance of our method, it also increases the computational complexity of the system, potentially limiting its real-time applicability. Second, the proposed approach may not be robust to novel deepfake techniques or attacks targeting the identified limitations. Therefore, the effectiveness and generalizability of our proposed method will need to be validated further on new datasets and deepfake scenarios as they become available in the future. Third, we observe that the temporal consistency contributed mostly to the reduction of false positive detection. While this can indeed be valuable for practical applications, designing additional components to further enhance the true positive detection is also of critical importance. Additional future research directions in this area may include a lightweight version of MASDT for real-time or edge deployment, which could be achieved through distillation. Moreover, by integrating various modalities such as visual, audio, and text data and leveraging the strengths and complementary aspects of each modality, a unified framework could significantly enhance detection capabilities and overall performance through a holistic understanding of manipulated content. Lastly, extending our framework to accurately detect deepfakes beyond only faces, for example to full body video clips, natural sceneries, and others, can be an interesting future research direction.

Regarding our work in Chapter 4, as seen from the results, FairAlign has the potential for speed improvements. We believe the calculation of the matrix inversions in Equations 4.3 and 4.4 can be sped up using estimation strategies, resulting in an overall faster FairAlign. Moreover, while FairAlign is designed for deepfake detection,

its applicability and efficacy in other domains that involve sensitive attributes, such as facial expression recognition or affect analysis, remain uncertain. Each domain could introduce distinct challenges that may impact FairAlign's performance. Lastly, our study on skin tone bias reduction relies on the detection and quantification of skin tones. While we resorted to standard and widely accepted off-the-shelf modules for our skin tone detection pipeline, custom-designing this component with fairness-promoting procedures embedded into it could be a promising and interesting future research avenue. Additional future directions may involve exploring the robustness of FairAlign against varied and potentially adversarial inputs warrants further investigation. Moreover, the notion of robustness, especially its trade-off with respect to fairness [134] can be studied. Investigating the intertwined relationship of fairness and robustness in the context of deepfake detection remains an open question in the field.

Bibliography

[1] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-df: A large-scale challenging dataset for deepfake forensics," *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3204–3213, 2020. ix, 4, 6, 8, 23, 24, 25, 39

[2] D. A. Coccomini, N. Messina, C. Gennaro, and F. Falchi, "Combining efficientnet and vision transformers for video deepfake detection," *International Conference on Image Analysis and Processing*, 2022. ix, 3, 4, 41, 43, 44, 45, 46, 50

[3] Y. Ju, S. Hu, S. Jia, G. H. Chen, and S. Lyu, "Improving fairness in deepfake detection," *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024. ix, 2, 3, 4, 15, 16, 40, 41, 43, 44, 45, 46, 49, 50, 52

[4] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," *IEEE International Conference on Computer Vision*, pp. 618–626, 2017. ix, 29, 31

[5] M. Hort, J. M. Zhang, F. Sarro, and M. Harman, "Fairea: A model behaviour mutation approach to benchmarking bias mitigation methods," *ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 2021. ix, 7, 17, 39, 40

[6] N. Yu, V. Skripniuk, S. Abdelnabi, and M. Fritz, "Artificial fingerprinting for generative models: Rooting deepfake attribution in training data," *IEEE/CVF International Conference on Computer Vision*, 2021. 1

[7] A. Borji, "Generated faces in the wild: Quantitative comparison of stable diffusion, midjourney and dall-e 2," *arXiv:2210.00586*, 2022. 1

[8] D. K. Kanbach, L. Heiduk, G. Blueher, M. Schreiter, and A. Lahmann, "The genai is out of the bottle: generative artificial intelligence from a business model innovation perspective," *Review of Managerial Science*, pp. 1–32, 2023. 1

[9] M. Masood, M. Nawaz, K. M. Malik, A. Javed, A. Irtaza, and H. Malik, "Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward," *Applied Intelligence*, pp. 1–53, 2022. 1, 2

[10] M. Westerlund, "The emergence of deepfake technology: A review," *Technology Innovation Management Review*, vol. 9, no. 11, 2019. 1

[11] S. Ahmed, "Navigating the maze: Deepfakes, cognitive ability, and social media news skepticism," *New Media & Society*, vol. 25, no. 5, pp. 1108–1129, 2023. 1

[12] T. Weikmann and S. Lecheler, "Cutting through the hype: Understanding the implications of deepfakes for the fact-checking actor-network," *Digital Journalism*, pp. 1–18, 2023. 1

[13] A. Tiwari, R. Dave, and M. Vanamala, "Leveraging deep learning approaches for deepfake detection: A review," *International Conference on Intelligent Systems, Metaheuristics & Swarm Intelligence*, p. 12–19, 2023. 1

[14] D. P. Kingma, S. Mohamed, D. Jimenez Rezende, and M. Welling, "Semi-supervised learning with deep generative models," *Advances in neural information processing systems*, vol. 27, 2014. 1

[15] J. Xu, H. Li, and S. Zhou, "An overview of deep generative models," *IETE Technical Review*, vol. 32, no. 2, pp. 131–139, 2015. 1

[16] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," *International Conference on Machine Learning*, vol. 32, pp. 1278–1286, 2014. 1

[17] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, vol. 65, no. 1, p. 99–106, 2021. 2

[18] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," *International Conference on Machine Learning*, pp. 8748–8763, 2021. 2

[19] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," *IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 2

[20] M. Masood, M. Nawaz, K. M. Malik, A. Javed, A. Irtaza, and H. Malik, "Deepfakes generation and detection: state-of-the-art, open challenges, countermeasures, and way forward," *Applied Intelligence*, vol. 53, no. 4, pp. 3974–4026, 2023. 2

[21] L. A. Passos, D. Jodas, K. A. P. da Costa, L. A. S. Júnior, D. Rodrigues, J. D. Ser, D. Camacho, and J. P. Papa, "A review of deep learning-based approaches for deepfake content detection," *arXiv: 2202.06095*, 2022. 2

[22] C. R. Leibowicz, S. McGregor, and A. Ovadya, "The deepfake detection dilemma: A multistakeholder exploration of adversarial dynamics in synthetic media," *AAAI/ACM Conference on AI, Ethics, and Society*, 2021. 2

[23] S. Das, M. Kolahdouzi, L. Özparlak, W. Hickie, and A. Etemad, "Unmasking deepfakes: Masked autoencoding spatiotemporal transformers for enhanced video forgery detection," *IEEE International Joint Conference on Biometrics*, 2023. 2, 9, 41, 43, 44, 45, 46, 50

[24] Z. Wang, J. Bao, W. Zhou, W. Wang, and H. Li, "AltFreezing for more general video face forgery detection," *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2, 41, 43, 44, 45, 46, 49, 50

[25] J. Cao, C. Ma, T. Yao, S. Chen, S. Ding, and X. Yang, "End-to-end reconstruction-classification learning for face forgery detection," *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2, 3, 27, 28, 29, 30, 41, 43, 44, 45, 46, 50

[26] L. Trinh and Y. Liu, "An examination of fairness of ai models for deepfake detection," *International Joint Conference on Artificial Intelligence*, 2021. 2, 15, 16

[27] Y. Xu, P. Terhörst, K. Raja, and M. Pedersen, "A comprehensive analysis of AI biases in deepfake detection with massively annotated databases," *arXiv:2208.05845*, 2022. 2, 15, 16

[28] A. V. Nadimpalli and A. Rattani, "GBDF: Gender balanced deepfake dataset towards fair deepfake detection," *arXiv:2207.10246*, 2022. 2, 3, 15

[29] L. Zhang, H. Chen, S. Hu, B. Zhu, X. Wu, J. Hu, and X. Wang, "X-transfer: A transfer learning-based framework for robust gan-generated fake image detection," *Computing Research Repository (CoRR)*, 2023. 2

[30] S. Yang, S. Hu, B. Zhu, Y. Fu, S. Lyu, X. Wu, and X. Wang, "Improving cross-dataset deepfake detection with deep information decomposition," *arXiv:2310.00359*, 2023. 2

[31] P. Yu, Z. Xia, J. Fei, and Y. Lu, "A survey on deepfake video detection," *IET Biometrics*, vol. 10, no. 6, pp. 607–624, 2021. 3

[32] M. S. Rana, M. N. Nobi, B. Murali, and A. H. Sung, "Deepfake detection: A systematic literature review," *IEEE Access*, 2022. 3, 11, 13

[33] C. Hazirbas, J. Bitton, B. Dolhansky, J. Pan, A. Gordo, and C. C. Ferrer, "Towards measuring fairness in ai: The casual conversations dataset," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 4, no. 3, pp. 324–332, 2021. 4, 15, 16, 38

[34] C. O. Little, M. Weylandt, and G. I. Allen, "To the fairness frontier and beyond: Identifying, quantifying, and optimizing the fairness-accuracy pareto frontier," *arXiv:2206.00074*, 2022. 4, 16, 38

[35] S. Maity, D. Mukherjee, M. Yurochkin, and Y. Sun, "Does enforcing fairness mitigate biases caused by subpopulation shift?" *Advances in Neural Information Processing Systems*, 2021. 4, 16

[36] M. L. Wick, S. Panda, and J.-B. Tristan, "Unlocking fairness: a trade-off revisited," *Neural Information Processing Systems*, 2019. 4, 38

[37] S. Dutta, D. Wei, H. Yueksel, P.-Y. Chen, S. Liu, and K. Varshney, "Is there a trade-off between fairness and accuracy? A perspective using mismatched hypothesis testing," *International Conference on Machine Learning*, 2020. 4, 16, 17, 38

[38] Q. Wang, Z. Xu, Z. Chen, Y. Wang, S. Liu, and H. Qu, "Visual analysis of discrimination in machine learning," *IEEE Transactions on Visualization and Computer Graphics*, vol. 27, no. 2, pp. 1470–1480, 2021. 4

[39] J. Li, Y. Moskovitch, and H. V. Jagadish, "Denouncer: detection of unfairness in classifiers," *VLDB Endowment*, vol. 14, no. 12, 2021. 4

[40] H. Zhang, N. Shahbazi, X. Chu, and A. Asudeh, "Fairrover: explorative model building for fair and responsible machine learning," *Workshop on Data Management for End-To-End Machine Learning*, 2021. 4

[41] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *International Conference on Learning Representations*, 2021. 5, 23

[42] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15 979–15 988, 2022. 5, 20

[43] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," *International Conference on Computer Vision*, 2015. 5, 24, 25

[44] L. Wolf, T. Hassner, and I. Maoz, "Face recognition in unconstrained videos with matched background similarity," *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 529–534, 2011. 5, 24, 25

[45] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to detect manipulated facial images," *International Conference on Computer Vision*, 2019. 5, 11, 23, 24

[46] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to detect manipulated facial images," *IEEE/CVF International Conference Computer Vision*, 2019. 6, 8, 39

[47] J. Deng, J. Guo, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," *Computer Vision and Pattern Recognition*, 2018. 7, 38, 39

[48] H. Xu, A. Sarkar, and A. L. Abbott, "Color invariant skin segmentation," *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2022. 7, 38

[49] C. M. Heldreth, E. P. Monk, A. T. Clark, C. Schumann, X. Eyee, and S. Ricco, "Which skin tone measures are the most inclusive? an investigation of skin tone

measures for artificial intelligence." *ACM Journal on Responsible Computing*, 2023. 7, 8, 38

[50] O. Lesota, S. Brandl, M. Wenzel, A. B. Melchiorre, E. Lex, N. Rekabsaz, and M. Schedl, "Exploring cross-group discrepancies in calibrated popularity for accuracy/fairness trade-off optimization." *RecSys*, 2022. 7, 17, 39

[51] X. Li, P. Wu, and J. Su, "Accurate fairness: Improving individual fairness without trading accuracy," *AAAI Conference on Artificial Intelligence*, 2023. 7, 39

[52] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1800–1807, 2017. 11

[53] D. Cozzolino, G. Poggi, and L. Verdoliva, "Recasting residual-based local descriptors as convolutional neural networks: an application to image forgery detection," *ACM Workshop on Information Hiding and Multimedia Security*, 2017. 11, 27, 28

[54] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "Mesonet: a compact facial video forgery detection network," *IEEE International Workshop on Information Forensics and Security*, pp. 1–7, 2018. 11, 27, 28

[55] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017. 12

[56] H. Zhao, W. Zhou, D. Chen, T. Wei, W. Zhang, and N. Yu, "Multi-attentional deepfake detection," *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 12, 27

[57] K. Sun, H. Liu, T. Yao, X. Sun, S. Chen, S. Ding, and R. Ji, "An information theoretic approach for attention-driven face forgery detection," *European Conference on Computer Vision*, pp. 111–127, 2022. 12, 27

[58] H. Liu, X. Li, W. Zhou, Y. Chen, Y. He, H. Xue, W. Zhang, and N. Yu, "Spatial-phase shallow learning: rethinking face forgery detection in frequency domain," *IEEE/CVF conference on computer vision and pattern recognition*, pp. 772–781, 2021. 12, 14, 15

[59] S. Chen, T. Yao, Y. Chen, S. Ding, J. Li, and R. Ji, "Local relation learning for face forgery detection," *AAAI Conference on Artificial Intelligence*, pp. 1081–1088, 2021. 12, 27, 28, 29

[60] E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, and P. Natarajan, "Recurrent convolutional strategies for face manipulation detection in videos," *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 80–87, 2019. 13

[61] I. Ganiyusufoglu, L. M. Ngô, N. Savov, S. Karaoglu, and T. Gevers, "Spatio-temporal features for generalized detection of deepfake videos," *arXiv preprint arXiv:2010.11844*, 2020. 13

[62] D. Güera and E. J. Delp, "Deepfake video detection using recurrent neural networks," *IEEE International Conference on Advanced Video and Signal Based Surveillance*, 2018. 13

[63] B. Kaddar, S. A. Fezza, Z. Akhtar, W. Hamidouche, A. Hadid, and J. Serra-Sagristà, "Deepfake detection using spatiotemporal transformer," *ACM Transactions on Multimedia Computing, Communications and Applications*, 2024. 13

[64] L. Chen, Y. Zhang, Y. Song, L. Liu, and J. Wang, "Self-supervised learning of adversarial examples: Towards good generalizations for deepfake detections," *IEEE/CVF conference on computer vision and pattern recognition*, 2022. 13, 27

[65] H. Zhao, W. Zhou, D. Chen, W. Zhang, and N. Yu, "Self-supervised transformer for deepfake detection," *arXiv preprint arXiv:2203.01265*, 2022. 13, 14

[66] J. Zhang, J. Ni, and H. Xie, "Deepfake videos detection using self-supervised decoupling network," *IEEE International Conference on Multimedia and Expo*, pp. 1–6, 2021. 13, 14

[67] G. Knafo and O. Fried, "Fakeout: Leveraging out-of-domain self-supervision for multi-modal video deepfake detection," *arXiv preprint arXiv:2212.00773*, 2022. 13, 14

[68] Y. Xu, K. Raja, and M. Pedersen, "Supervised contrastive learning for generalizable and explainable deepfakes detection," *IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 379–389, 2022. 13, 14

[69] S. Fung, X. Lu, C. Zhang, and C.-T. Li, "Deepfakeucl: Deepfake detection via unsupervised contrastive learning," *International Joint Conference on Neural Networks*, pp. 1–8, 2021. 14

[70] Z. Cai, K. Stefanov, A. Dhall, and M. Hayat, "Do you really mean that? content driven audio-visual deepfake dataset and multimodal method for temporal forgery localization," *International Conference on Digital Image Computing: Techniques and Applications*, pp. 1–10, 2022. 14

[71] H. Ilyas, A. Javed, and K. M. Malik, "Avfakenet: A unified end-to-end dense swin transformer deep learning model for audio-visual deepfakes detection," *Applied Soft Computing*, vol. 136, p. 110124, 2023. 14

[72] H. Khalid, S. Tariq, M. Kim, and S. S. Woo, "Fakeavceleb: A novel audio-video multimodal deepfake dataset," *Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2021. 14

[73] I. Amerini, L. Galteri, R. Caldelli, and A. D. Bimbo, "Deepfake video detection through optical flow based cnn," *IEEE/CVF International Conference on Computer Vision Workshop*, pp. 1205–1207, 2019. 15, 31

[74] R. Caldelli, L. Galteri, I. Amerini, and A. Del Bimbo, "Optical flow based cnn for detection of unlearnt deepfake manipulations," *Pattern Recognition Letters*, vol. 146, pp. 31–37, 2021. 15, 31

[75] P. Saikia, D. Dholaria, P. Yadav, V. Patel, and M. Roy, "A hybrid cnn-lstm model for video deepfake detection by leveraging optical flow features," *2022 International Joint Conference on Neural Networks*, pp. 1–7, 2022. 15, 31

[76] Z. Jiang, P. Zhao, and Z. Zheng, "Optical flow-attention fusion model for deepfake detection," *International Conference on Algorithms, Computing and Artificial Intelligence*, 2023. 15

[77] W. Zhuang, Q. Chu, Z. Tan, Q. Liu, H. Yuan, C. Miao, Z. Luo, and N. Yu, "Uia-vit: Unsupervised inconsistency-aware method based on vision transformer for face forgery detection," *European Conference on Computer Vision*, 2022. 15, 32

[78] J. Hu, X. Liao, D. Gao, S. Tsutsui, Q. Wang, Z. Qin, and M. Z. Shou, "Mover: Mask and recovery based facial part consistency aware method for deepfake video detection," *arXiv preprint arXiv: 2303.01740*, 2023. 15

[79] L. Shi, J. Zhang, and S. Shan, "Real face foundation representation learning for generalized deepfake detection," *arXiv preprint arXiv: 2303.08439*, 2023. 15

[80] Y.-J. Heo, W.-H. Yeo, and B.-G. Kim, "Deepfake detection algorithm based on improved vision transformer," *Applied Intelligence*, vol. 53, no. 7, pp. 7512–7527, 2023. 15

[81] D. Wodajo and S. Atnafu, "Deepfake video detection using convolutional vision transformer," *arXiv:2102.11126*, 2021. 15, 32

[82] M. Pu, M. Y. Kuan, N. T. Lim, C. Y. Chong, and M. K. Lim, "Fairness evaluation in deepfake detection models using metamorphic testing," *International Workshop on Metamorphic Testing*, 2022. 15

[83] N. M. Kinyanjui, T. Odonga, C. Cintas, N. C. F. Codella, R. Panda, P. Sattigeri, and K. R. Varshney, "Estimating skin tone and effects on classification performance in dermatology datasets," *Advances in Neural Information Processing Systems Fair Machine Learning for Health Workshop*, 2019. 16

[84] C. M. Heldreth, E. P. Monk, A. T. Clark, C. Schumann, X. Eyee, and S. Ricco, "Which skin tone measures are the most inclusive? an investigation of skin tone measures for artificial intelligence." *ACM Journal on Responsible Computing*, 2023. 16

[85] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. C. Ferrer, "The deepfake detection challenge (DFDC) dataset," *arXiv:2006.07397*, 2020. 16

[86] Y. Wang, X. Wang, A. Beutel, F. Prost, J. Chen, and E. H. Chi, "Understanding and improving fairness-accuracy trade-offs in multi-task learning," *Conference on Knowledge Discovery and Data Mining*, 2021. 16

[87] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz, "PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume," *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018. 19

[88] D. Ibanez, R. Fernandez-Beltran, F. Pla, and N. Yokoya, "Masked auto-encoding spectral–spatial transformer for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–14, 2022. 20

[89] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito,

M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," *Advances in Neural Information Processing Systems*, pp. 8024–8035, 2019. 22

[90] M. Contributors, "MMFlow: Openmmlab optical flow toolbox and benchmark," https://github.com/open-mmlab/mmflow, 2021. 22

[91] H. Zhang, M. Cissé, Y. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *International Conference On Learning Representations*, 2017. 23

[92] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "Cutmix: Regularization strategy to train strong classifiers with localizable features," *IEEE/CVF International Conference on Computer Vision*, pp. 6023–6032, 2019. 23

[93] Deepfakes, "Faceswap: Deepfakes Software," https://github.com/deepfakes/faceswap/, accessed 2023. 24, 27, 28

[94] M. Kowalski, "FaceSwap: Deep Learning for Face Swapping," https://github.com/MarekKowalski/FaceSwap, accessed 2023. 24, 27, 28

[95] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Nießner, "Face2face: Real-time face capture and reenactment of rgb videos," *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016. 24, 27, 28

[96] J. Thies, M. Zollhöfer, and M. Nießner, "Deferred neural rendering: Image synthesis using neural textures," *ACM Transactions on Graphics (ToG)*, 2019. 24, 27, 28

[97] S. Tomar, "Converting video formats with ffmpeg," *Linux Journal*, vol. 2006, no. 146, p. 10, 2006. 25, 39

[98] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. 25

[99] J. Fridrich and J. Kodovský, "Rich models for steganalysis of digital images," *IEEE Transactions on Information Forensics and Security*, vol. 7, pp. 868–882, 2012. 27, 28

[100] N. Rahmouni, V. Nozick, J. Yamagishi, and I. Echizen, "Distinguishing computer graphics from natural images using convolution neural networks," *IEEE Workshop on Information Forensics and Security (WIFS)*, pp. 1–6, 2017. 27, 28

[101] L. Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wen, and B. Guo, "Face x-ray for more general face forgery detection," *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5000–5009, 2020. 27

[102] B. Bayar and M. C. Stamm, "A deep learning approach to universal image manipulation detection using a new convolutional layer," *ACM Workshop on Information Hiding and Multimedia Security*, 2016. 27, 28

[103] U. A. Ciftci, I. Demir, and L. Yin, "Fakecatcher: Detection of synthetic portrait videos using biological signals," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2020. 27

[104] I. Masi, A. Killekar, R. M. Mascarenhas, S. Pratik Gurudatt, and W. AbdAlmageed, "Two-branch Recurrent Network for Isolating Deepfakes in Videos," *European Conference on Computer Vision*, 2020. 27

[105] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "Faceforensics++: Learning to detect manipulated facial images," *IEEE/CVF International Conference on Computer Vision*, pp. 1–11, 2019. 27, 28, 29

[106] A. Haliassos, K. Vougioukas, S. Petridis, and M. Pantic, "Lips don't lie: A generalisable and robust approach to face forgery detection," *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5039–5049, 2021. 27

[107] R. Tolosana, S. Romero-Tapiador, J. Fierrez, and R. Vera-Rodriguez, "Deepfakes evolution: Analysis of facial regions and fake detection performance," *International Conference on Pattern Recognition*, 2021. 27

[108] Y. Qian, G. Yin, L. Sheng, Z. Chen, and J. Shao, "Thinking in frequency: Face forgery detection by mining frequency-aware clues," *European Conference on Computer Vision*, 2020. 27, 28

[109] K. Sun, T. Yao, S. Chen, S. Ding, J. Li, and R. Ji, "Dual contrastive learning for general face forgery detection," *AAAI Conference on Artificial Intelligence*, vol. 36, pp. 2316–2324, 2022. 29, 30

[110] S. Ganguly, A. Ganguly, S. Mohiuddin, S. Malakar, and R. Sarkar, "Vixnet: Vision transformer with xception network for deepfakes based video and image forgery detection," *Expert Systems with Applications*, 2022. 32

[111] R. Bachmann, D. Mizrahi, A. Atanov, and A. Zamir, "MultiMAE: Multi-modal multi-task masked autoencoders," *European Conference on Computer Vision*, 2022. 32, 33

[112] D. Mukherjee, F. Petersen, M. Yurochkin, and Y. Sun, "Domain adaptation meets individual fairness. and they get along," *Advances in Neural Information Processing Systems*, 2022. 34

[113] T.-D. Truong, N. Le, B. Raj, J. Cothren, and K. Luu, "Fredom: Fairness domain adaptation approach to semantic scene understanding," *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 34

[114] N. Joshi and P. Burlina, "Ai fairness via domain adaptation," *arXiv: 2104.01109*, 2021. 34

[115] X. Han, J. Chi, Y. Chen, Q. Wang, H. Zhao, N. Zou, and X. Hu, "Ffb: A fair fairness benchmark for in-processing group fairness methods," *International Conference on Learning Representations*, 2024. 34

[116] M. Srivastava, H. Heidari, and A. Krause, "Mathematical notions vs. human perception of fairness: A descriptive approach to fairness for machine learning," *Conference on Knowledge Discovery and Data Mining*, 2019. 35

[117] Y.-W. Luo and C.-X. Ren, "Conditional bures metric for domain adaptation," *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 35, 36, 37, 42

[118] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *Journal of Machine Learning Research*, vol. 13, p. 723–773, 2012. 35

[119] F. Steinke and B. Schölkopf, "Kernels, regularization and differential equations," *Pattern Recognition*, vol. 41, no. 11, pp. 3271–3286, 2008. 35, 36

[120] R. Bhatia, T. Jain, and Y. Lim, "On the bures-wasserstein distance between positive definite matrices," *Expositiones mathematicae*, 2017. 36

[121] Z. Zhang, M. Wang, and A. Nehorai, "Optimal transport in reproducing kernel hilbert spaces: Theory and applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 7, 2020. 36

[122] K. Fukumizu, F. R. Bach, and M. I. Jordan, "Kernel dimension reduction in regression," *The Annals of Statistics*, vol. 37, no. 4, pp. 1871–1905, 2009. 36

[123] R. S. Schulman and E. M. Cramer, "Comparison of the accuracies of some regression algorithms," *Journal of Statistical Computation and Simulation*, vol. 4, no. 2, pp. 85–93, 1975. 37

[124] S. Chen, Y. Liu, X. Gao, and Z. Han, "Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices," *Chinese Conference on Biometric Recognition*, 2018. 38

[125] J. Wang, X. E. Wang, and Y. Liu, "Understanding instance-level impact of fairness constraints," *International Conference on Machine Learning*, 2022. 40

[126] H. Weerts, M. Dudík, R. Edgar, A. Jalali, R. Lutz, and M. Madaio, "Fairlearn: Assessing and improving fairness of ai systems," *arXiv:2303.16626*, 2023. 40

[127] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," *International Conference on Machine Learning*, 2019. 41

[128] J. Chai, T. Jang, and X. Wang, "Fairness without demographics through knowledge distillation," *Advances in Neural Information Processing Systems*, 2022. 41

[129] N. Deka and D. J. Sutherland, "Mmd-b-fair: Learning fair representations with statistical testing," *International Conference on Artificial Intelligence and Statistics*, 2022. 41, 43, 44, 45, 46, 50, 52

[130] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," *Advances in Neural Information Processing Systems*, 2019. 41

[131] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *International Conference on Learning Representations*, 2017. 41

[132] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Insightface: 2d and 3d face analysis project," https://github.com/deepinsight/insightface, 2022. 42

[133] T. Hashimoto, M. Srivastava, H. Namkoong, and P. Liang, "Fairness without demographics in repeated loss minimization," *International Conference on Machine Learning*, 2018. 43, 44, 45, 46, 50, 52

[134] H. Xu, X. Liu, Y. Li, and J. Tang, "To be robust or to be fair: Towards fairness in adversarial training," *International Conference on Machine Learning*, 2020. 56