Multi-target Deep Domain Adaptation for Deepfake Detection

Md Shamim Seraj
Department of Computer Science
Florida State University

Shayok Chakraborty
Department of Computer Science
Florida State University

***Abstract*—While the recent progress of generative AI technology (primarily through diffusion models) has revolutionized AI research, they have also posed significant challenges for real-world deepfake detection. Existing detection techniques demonstrate promise in detecting deepfakes on which the model has been trained; however, their performance drops significantly when applied to detect forgeries created using other manipulation techniques, on which the model has not been sufficiently trained. Thus, there is a pressing need for a technology that can detect deepfakes created using unknown faking techniques, without losing prior knowledge about already learned faking techniques. In this paper, we propose a novel multi-target deep domain adaptation framework to address this practically relevant and timely problem. Our framework can leverage a large amount of annotated data (fake/genuine) generated using a particular faking technique (source domain) and a small amount of labeled data generated using different unknown faking techniques (target domains) to induce a deep neural network with good generalization capability on the source domain, as well as all the target domains of interest. Further, our framework can efficiently utilize unlabeled data in the target domains, which are more readily available than labeled data. We design a novel loss function specific to the multi-target domain adaptation task and use the SGD method to optimize the loss and train the deep neural network. Our extensive empirical studies on benchmark datasets, using multiple types of deepfakes, corroborate the promise and potential of our framework for real-world applications. To the best of our knowledge, this is the first research effort to develop a multi-target deep domain adaptation framework for deepfake detection. Our code is open-sourced at https://github.com/shuvornb/MultiTargetDomainAdaptation.**

***Index Terms*—Deepfake Detection, Domain Adaptation, Deep Learning**

I. Introduction

Generative AI techniques (such as GANs, and particularly, diffusion models) have recently demonstrated their potential for highly realistic visual content generation, which makes them an excellent tool for creating deepfakes [1][2][3]. Of particular concern, are human facial manipulations that can generate entirely synthetic faces even at very high resolutions [4], animate a subject's face to make it express the desired emotions [5], or modify facial expressions [6]. The misuse of deepfake technology poses increasing risks, including creating child sexual abuse materials, celebrity pornographic videos and fake propaganda videos for gaining unlawful political influence [7][8][9]. Thus, reliably detecting deepfakes has become a problem of immense practical importance [1].

Convolutional Neural Networks (CNNs) have depicted promising performance for deepfake detection [10][11][12][13][14][15][16]. However, they have been shown to suffer from the transferability problem; that is, while they depict impressive performance on manipulation techniques on which they have been trained, they tend to overfit to the manipulation specific artifacts, and their performance drops drastically when applied on new types of forgeries, even though they are semantically similar [17][18][19][20]. Accurately detecting fakes generated using the new manipulations using supervised learning methods, necessitates a large amount of annotated data (fake / genuine) from the new manipulation techniques. However, with the progress of generative AI at an unprecedented pace, newer and more sophisticated faking technologies are emerging on a regular basis; thus, collecting a large amount of labeled data for every single faking technique is not feasible. However, it is comparatively easier to obtain unlabeled data; for instance, we may have access to several images which may or may not be forged using unknown faking techniques, and that information is not available to us, that is, the labels of these images are unknown. The problem at hand is therefore is to detect deepfakes, given only a few (or none) labeled samples and a moderate amount of unlabeled samples from unknown faking techniques. Further, while accurately detecting deepfakes generated using unknown faking techniques is important, it is also equally important to retain prior knowledge about the known faking techniques to avoid catastrophic forgetting [21][22][23].

The research task can thus be posed as follows: *we are given abundant data generated using a particular type of deepfake (for instance, FaceSwap); this data is completely labeled as genuine / fake. We are also given a small amount (or none) of labeled data and a moderate amount of unlabeled data generated using multiple unknown manipulation technologies. Our objective is to train a deep CNN to accurately identify genuine and fake images generated using the known, as well as the unknown manipulation techniques.*

In this paper, we present a novel multi-target deep domain

This research was supported in part by the National Science Foundation under Grant Number: 2143424 (NSF CAREER Award)

[1]we use *deepfake* as a generic terminology to denote any kind of fake content

adaptation framework to address this timely and relevant problem in deepfake detection research. *Domain Adaptation (DA)* or *Transfer Learning (TL)* algorithms utilize abundant labeled data in one domain (source domain) to develop a model for related domain(s) of interest (target domain), where labeled data is scarce [24]. The data in the source and target domains are derived from different probability distributions; thus, a model trained on the source may not generalize to the target. Our problem can be formulated as a multi-target DA task, where the data from the given manipulation technique forms the source domain, and the data from the unknown techniques form multiple target domains. We propose a novel loss function specific to the multi-target DA task and utilize the SGD algorithm to optimize the loss and train a deep CNN. Our framework can also utilize unlabeled data in the target domains, which is more readily available than labeled data. We extensively validate our framework on benchmark datasets with multiple faking techniques, on challenging low resolution data, and also conduct a cross-dataset experiment with different manipulation techniques. Our framework depicts impressive performance even when no labeled data is available in any of the target domains. To our knowledge, multi-target domain adaptation has not been studied in deepfake detection research.

II. Related Work

Deepfake Detection: Fueled by generative AI, the rapid progress in deepfake creation technologies has posed a significant security threat, with the potential for misuse in spreading misinformation and engaging in malicious activities. This has spurred active research in the development of reliable deepfake detection algorithms. Please refer to [25], [26] for a comprehensive survey. Most of the current methods use some kind of a deep neural network architecture [27][18][28]. As mentioned before, they suffer from poor generalization, when tested on samples generated using unknown manipulation techniques.

Domain Adaptation: *Domain Adaptation (DA)* algorithms attempt to induce a model in a target domain of interest (with scarce labeled samples) by leveraging abundant labeled data from a related source domain, under the constraint of a probability distribution difference between the domains [24], [29]. A body of research in DA has focused on the development of metrics to quantify the disparity between the source and target domains (such as Maximum Mean Discrepancy (MMD) [30][31][32], Kullback Leibler Divergence [33], Jensen Shannon Divergence [34] among others) and then devising techniques to minimize the values of these metrics to address the domain disparity. With the advent and popularity of generative adversarial networks (GANs), different variants of GAN have been exploited for DA, including Domain Adversarial Neural Network (DANN) [35], Coupled Generative Adversarial Network (CoGAN) and its combination with Variational Autoencoder (VAE) [36], and Wasserstein GAN [37] among others. *Multi-target domain adaptation (MTDA)* is an extension of DA where multiple target domains are available. Isobe *et al.* [38] proposed a collaborative consistency learning framework for MTDA. Other strategies include a curriculum graph co-teaching framework [39], a multi-target knowledge distillation method [40], a post-training Hessian and distant regularizing quantization strategy to merge models for MTDA [41] and a dynamic generator with attention (DGWA) method for multi-source, multi-target DA [42].

Domain Adaptation for Deepfake Detection: DA for deepfake detection is a comparatively less explored research topic. The concept that the deeper layers of a DNN contain domain specific information was exploited by Tariq *et al.* [18], where the DNN was first trained on the source domain data and then the deeper layers were fine-tuned using the target domain data for detecting deepfakes. The *ForensicTransfer* framework, proposed by Cozzolino *et al.* [19], also adopted a similar strategy, where an autoencoder was first trained on the source domain data, which was then fine-tuned using target domain data. Kim *et al.* introduced *FReTAL* [20], a transfer learning-based method for deepfake detection, that leveraged representation learning and knowledge distillation; the same authors also introduced *CoReD* [43], another DA method for deepfake detection, that combined continual learning, representation learning and knowledge distillation. All these methods, however, require all the samples in the target domain to be annotated (fake / genuine), and cannot leverage the presence of unlabeled data in the target domain. Very recently, Chen and Tan [44] and Seraj *et al.* [45][46] used the concept of adversarial DA for deepfake detection. These methods only require access to domain labels (whether a sample is derived from the source or target domain) rather than fake / genuine labels, and can thus utilize unlabeled data in the target domain. Zhang *et al.* [17] proposed to minimize the MMD between the source and target domains and trained a CNN to learn domain invariant features, in the absence of labeled target domain samples. All these methods, however, assume a single target domain and do not consider the challenging scenario where the target domain itself can contain deepfakes generated by multiple unknown image manipulation techniques.

Unlike existing MTDA frameworks, which typically train a separate model for each source-target pair, and then collate their knowledge into a single model through a knowledge distillation / consistency learning framework [38], we design one loss function, and train a single DNN architecture end-to-end. We use the concept of moment matching to address the disparity between multiple target domains and the source domain simultaneously. We also include a class alignment loss term (to learn from the unlabeled target domain samples) which requires each sample to learn feature representations that strongly align with one of the possible classes (genuine / fake) and be distinct from the other class. We conduct extensive experimental studies on benchmark datasets under challenging real-world conditions (low-resolution data, limited labeled data in the target domain) to validate the performance of our framework.

III. Proposed Framework

A. Problem Setup

We are given data from a source domain S and N_T target domains $T_1, T_2, \ldots T_{N_T}$. Each domain represents data generated using a particular forgery technique (such as *Neural Textures*, *Face2Face* etc.). Each faking technology operates in a different manner and thus, the data in each domain is derived from a different probability distribution. The data in the source domain is labeled: $S = \{x_i, y_i\}_{i=1}^{|S|}$. Here $\{x_i\}$ denotes an image sample and $\{y_i\}$ denotes the binary label (fake / genuine). Let D_T denote the combined data from all the target domains: $D_T = T_1 \cup T_2 \cup \ldots \cup T_{N_T}$. The data in each target domain j consists of a labeled subset $D_{T_j}^L = \{x_i, y_i\}_{i=1}^{N_{T_j}^L}$, as well as an unlabeled subset: $D_{T_j}^U = \{x_i\}_{i=1}^{N_{T_j}^U}$, $\forall j = 1, 2, \ldots N_T$. As explained in Section I, obtaining annotated samples in the target domains is difficult, while obtaining unlabeled samples is relatively easier, that is, $|D_{T_j}^L| \ll |D_{T_j}^U|$, $\forall j = 1, 2, \ldots N_T$. Our objective is to train a deep convolutional neural network (CNN) which will furnish good generalization performance on the source domain, as well as all the target domains; that is, we would like our trained CNN to reliably detect deepfakes generated using the known faking technique, as well as all the unknown faking techniques. We propose to formulate a novel loss function, specific to the MTDA task, and train the deep network to optimize that loss. Our loss function consists of three components: (i) *a supervised loss* on the labeled data, which encourages the network to predict the labeled source and target samples accurately; (ii) *a moment matching loss* to align the source and all the target data distributions, and learn domain invariant feature representations; and (iii) *an unsupervised loss* on unlabeled target data, which encourages the network to predict the unlabeled samples from all the target domains with high confidence. These are detailed in the following sections.

B. Supervised Loss on Labeled Source and Target Data

Since we are solving a binary classification problem (fake / genuine), we use the binary cross entropy (BCE) loss [47] as the supervised loss term in our framework. Let the labeled source and target domain samples be denoted by $D_L = D_S \cup D_T^L = \{x_1, x_2, \ldots, x_{n_L}\}$, with corresponding labels $\{y_1, y_2, \ldots, y_{n_L}\}$. The BCE loss to train the deep CNN is given by:

$$\mathcal{L}_{sup} = -\frac{1}{n_L} \sum_{i=1}^{n_L} \Big(y_i . \log(p(y_i)) + (1 - y_i) . \log(1 - p(y_i)) \Big) \tag{1}$$

where $p(y_i)$ denotes the probability obtained from the softmax activation layer of the CNN. This term enforces the CNN to learn feature representations that are consistent with the labeled samples.

C. Moment Matching Loss on Source and Target Data

Matching moments of distributions has been explored as a strategy to address the disparity between the source and target domains in domain adaptation research [48][49][50][51]. The commonly used maximum mean discrepancy (MMD) measure, for instance, attempts to align the means of the source and target distributions in a Reproducing Kernel Hilbert Space (RKHS) [31][32] by minimizing the distance between them. Given the multiple target domains in our framework, we need to align each of them with the source domain, as well as align all the target domains among themselves. We leverage the technique proposed by Peng *et al.* [52], which attempts to align domains by directly aligning the moments of their deep feature distributions:

$$\mathcal{L}_{moment} = \sum_{m=1}^{M} \Bigg(\underbrace{\frac{1}{N_T} \sum_{i=1}^{N_T} \|\mathbb{E}(\mathcal{F}_i^m) - \mathbb{E}(\mathcal{F}_S^m)\|_2}_{\text{source-target alignment}} + \underbrace{\frac{1}{\binom{N_T}{2}} \sum_{i,j=1, i \neq j}^{N_T} \|\mathbb{E}(\mathcal{F}_i^m) - \mathbb{E}(\mathcal{F}_j^m)\|_2}_{\text{target-target alignment}} \Bigg) \tag{2}$$

Here, $\mathbb{E}(.)$ denotes the expectation operator, $\|.\|_2$ denotes the 2-norm of a vector and M denotes the maximum order of the moments considered. The first term aligns the source domain separately with each target domain, while the second term aligns the target domains among themselves (considering a single pair at a time). Minimizing this term ensures that the source and all the target domains are aligned and the underlying CNN learns domain invariant feature representations.

D. Unsupervised Loss on Unlabeled Target Data

One of the useful features of our framework is that it can utilize the presence of unlabeled data in the target domains (generated using the unknown faking techniques), which are easier to obtain than labeled data. In the absence of target domain labels, our strategy is to enforce the CNN to predict the label of each unlabeled target domain sample with maximum confidence (least uncertainty). Since each target domain sample can belong to exactly one class (fake / genuine), our rationale is to ensure that the CNN learns feature representations for each target sample that strongly align with exactly one of the two classes, and does not align with the other class. Such learned representations will enable the model to appropriately distinguish the two classes, and assign a label confidently to each sample.

We consider K labeled samples from each class from the source domain. Let $\mathcal{F}_S^{ck}$ denote the learned feature representation for the k^{th} source sample from class c, where $c \in \{\text{fake, genuine}\}$, and $\mathcal{F}_T^i$ denote the feature representation of an unlabeled target domain sample x_i. Since the feature distributions of the source and all the target domain samples have already been aligned through the moment matching loss (as described in Section III-C), we consider all the target domain samples as being derived from a single domain and do not make the domain-wise distinction in formulating this loss term. Our objective is to ensure that $\mathcal{F}_T^i$ is similar to all

the K learned source representations from one of the classes c, and dissimilar to the other class. We consider alignment with K randomly selected source samples, instead of a single sample, to ensure a more robust alignment. We formulate the following probability term to quantify the extent of alignment of the target sample x_i with respect to class c:

$$p_{ic} = \frac{\sum_{k=1}^{K} exp\langle \mathcal{F}_T^i, \mathcal{F}_S^{ck} \rangle}{\sum_{c=1}^{2} \sum_{k=1}^{K} exp\langle \mathcal{F}_T^i, \mathcal{F}_S^{ck} \rangle} \tag{3}$$

We compute the feature similarity using their dot product (denoted by $\langle \cdot, \cdot \rangle$) and use the exponential function for ease of differentiability. The denominator sums over both the classes to ensure that the probability vector is normalized and sums to 1. Since our goal is to align x_i to exactly one of the two classes, we want one of the probability values to be high, and the other one to be low. We formulate the following class alignment (CA) term to capture this notion, which is given by the entropy of the probability distribution p_{ic}:

$$\mathcal{L}_{CA} = -\frac{1}{N_T^U} \sum_{i=1}^{N_T^U} \sum_{c=1}^{2} p_{ic} \log p_{ic} \tag{4}$$

where N_T^U denotes the total number of unlabeled samples across all the target domains. Minimizing this loss produces probability vectors that are one-hot vectors, that is, the deep network furnishes confident predictions on the unlabeled target data. We combine the three terms to design the overall loss term to train the deep CNN:

$$\mathcal{L} = \mathcal{L}_{sup} + \lambda_1 \mathcal{L}_{moment} + \lambda_2 \mathcal{L}_{CA} \tag{5}$$

where λ_1, λ_2 are weight parameters governing the relative importance of each term. We use the stochastic gradient descent (SGD) method to optimize the loss and train the deep network.

IV. Experiments and Results

Datasets. We used the *FaceForensic++ (FF++)* benchmark dataset [28] in our experimental studies. It contains $1,000$ genuine videos, each of which was converted to a fake video using the following four four faking techniques: *Face2Face (F2F), FaceSwap (FS), DeepFakes (DF)* and *NeuralTextures (NT)*. Each video was split to generate 50 images (128×128) and we used an off-the-shelf face recognition software [2] to detect and crop the facial regions from these images. We also conducted a cross-dataset study, where we used the *Celeb-DF* dataset [53], which contains high-quality deepfake videos of celebrities.

Comparison Baselines. We used DA algorithms that have been explicitly studied for the deepfake detection problem as comparison baselines in our work: (i) *Fine-tuning (FT)* [18]; (ii) *Transferable GAN-images Detection (TGD)* [54]; (iii) *Feature Representation Transfer Adaptation Learning (FReTAL)* [20]; and (iv) *Knowledge Distillation (KD)* [20]. All these baselines are supervised, that is, they require all the samples in the target domain to be labeled. Further, all these baselines are only designed for a single source and a single target domain; to extend them to the multi-target setting, we combined all the target domain data and treated that as a single target domain. This is the most common strategy to extend single target DA to the multi-target setting and has been used in previous research on multi-target DA [38].

Experimental Setup. In each experiment, we were given images from one source domains and multiple target domain, where each domain represents a particular faking technique (*DF, F2F, FS* etc.). We experimented with two and three target domains in this research. The data in the source domain was completely labeled. The target domain contained a small amount of labeled data and a large amount of unlabeled data (to appropriately capture a real-world scenario). We used $40,000$ images as the source domain data, $2,000$ images as the labeled target domain data and $18,000$ images as the unlabeled target domain data (taken equally from all the target domains). The test set contained $10,000$ images from the source domain, and $10,000$ images taken equally from all the target domains, to validate the performance of the model on all the domains. Each set contained an equal number of images from the two classes: *fake* and *genuine*. Each experiment was conducted 3 times, and the results were averaged to rule out the effects of randomness. The parameters λ_1 and λ_2 in Equation (5) were taken as 0.9 and 0.4 respectively, based on preliminary experiments. The number of moments M in Equation (2) was taken as 2.

We used the *Xception* network as the base model due to its promising performance on the FaceForensics++ dataset [3]. The F1 score was used as the evaluation metric, similar to [20]. We report the results separately on the source samples and the target samples (from all the target domains) in the test set.

Implementation Details. We implement our framework using the PyTorch deep learning library. The backbone Xception architecture was initialized with ImageNet-pretrained weights. Model optimization is performed using the *AdamW* optimizer with a learning rate of 0.0003 and a weight decay of 0.0001. Training was carried out for a fixed number of epochs, and the number of steps per epoch was set to 500. At each training step, mini-batches were independently sampled from the labeled source set, labeled target sets, and unlabeled target sets. All experiments were conducted on a DELL Alienware Aurora R15 workstation equipped with an NVIDIA GeForce RTX 3090 GPU with 24 GB of memory. The system runs Ubuntu 22.04.3 LTS, and all experiments were executed using Python 3.11.5 with PyTorch 2.5.1 and CUDA 12.3 support.

A. Main Results

Table I reports the mean ($\pm$ std) F1 scores for multiple multi-target domain adaptation tasks, where one domain is treated as the source and multiple domains are treated as targets. The results are reported separately for the source and target domains. Across all evaluated tasks, the proposed

[2]https://pypi.org/project/face-recognition/

TABLE I

MEAN ($\pm$ STD) F1 SCORES (%) FOR MULTI-TARGET DOMAIN ADAPTATION TASKS. BEST RESULTS ARE HIGHLIGHTED IN BOLD. THE NOTATION $x \rightarrow y, z$ DENOTES x AS SOURCE DOMAIN AND y, z AS THE TARGET DOMAINS.

DA Task	Domain	Proposed	FReTAL	FT	KD	TGD
DF $\rightarrow$ FS, F2F	Source	$\mathbf{97.96 \pm 0.03}$	89.74 ± 0.74	86.14 ± 1.41	95.70 ± 2.97	88.58 ± 5.92
	Target	$\mathbf{92.27 \pm 0.41}$	78.44 ± 0.79	75.12 ± 0.39	63.09 ± 4.85	70.48 ± 0.99
FS $\rightarrow$ F2F, DF	Source	$\mathbf{97.94 \pm 0.02}$	90.66 ± 0.41	86.06 ± 0.40	95.12 ± 0.72	86.92 ± 6.56
	Target	$\mathbf{94.51 \pm 0.32}$	80.11 ± 0.17	78.55 ± 0.55	69.28 ± 3.18	74.09 ± 1.22
FS $\rightarrow$ F2F, NT	Source	$\mathbf{94.31 \pm 1.16}$	87.18 ± 1.00	80.77 ± 1.82	84.84 ± 8.24	79.61 ± 10.44
	Target	$\mathbf{85.94 \pm 0.79}$	68.19 ± 1.91	67.28 ± 0.06	53.80 ± 2.42	61.03 ± 2.83
DF $\rightarrow$ FS, F2F, NT	Source	$\mathbf{93.32 \pm 0.36}$	90.62 ± 2.72	85.90 ± 1.28	88.27 ± 7.24	83.42 ± 3.35
	Target	$\mathbf{87.36 \pm 0.39}$	68.90 ± 1.72	69.92 ± 1.14	63.72 ± 4.58	68.71 ± 1.24
FS $\rightarrow$ DF, F2F, NT	Source	$\mathbf{95.74 \pm 0.45}$	93.17 ± 3.90	82.45 ± 1.08	90.13 ± 5.55	82.52 ± 1.72
	Target	$\mathbf{88.72 \pm 0.95}$	65.03 ± 7.59	70.84 ± 0.80	61.98 ± 1.69	68.23 ± 0.57

TABLE II

PERFORMANCE COMPARISON ON LOW-RESOLUTION (64×64) MULTI-TARGET DEEPFAKE DETECTION TASKS. BEST RESULTS ARE HIGHLIGHTED IN BOLD. THE NOTATION $x \rightarrow y, z$ DENOTES x AS SOURCE DOMAIN AND y, z AS THE TARGET DOMAINS.

DA Task	Domain	Proposed	FReTAL	KD	FT	TGD
DF $\rightarrow$ FS, F2F	Source	$\mathbf{95.20 \pm 2.31}$	90.30 ± 1.19	93.12 ± 3.93	86.73 ± 1.66	87.88 ± 4.21
	Target	$\mathbf{80.09 \pm 2.18}$	76.41 ± 0.21	62.85 ± 4.40	74.73 ± 1.21	69.94 ± 1.36
FS $\rightarrow$ F2F, DF	Source	$\mathbf{95.83 \pm 0.59}$	90.65 ± 1.65	92.27 ± 3.74	86.19 ± 0.90	76.98 ± 14.48
	Target	77.75 ± 6.05	$\mathbf{78.52 \pm 0.52}$	69.18 ± 6.17	77.33 ± 0.81	64.90 ± 5.82
DF $\rightarrow$ NT, FS	Source	$\mathbf{93.94 \pm 1.79}$	89.72 ± 0.34	87.79 ± 1.21	85.98 ± 0.87	93.40 ± 0.45
	Target	$\mathbf{70.64 \pm 0.74}$	69.50 ± 0.21	53.39 ± 2.12	68.68 ± 1.34	57.57 ± 1.04
DF $\rightarrow$ FS, F2F, NT	Source	92.54 ± 1.11	92.31 ± 0.12	$\mathbf{97.33 \pm 1.04}$	80.40 ± 1.14	73.65 ± 0.52
	Target	$\mathbf{73.27 \pm 2.64}$	66.51 ± 1.20	59.31 ± 0.54	67.56 ± 2.10	61.90 ± 1.36
F2F $\rightarrow$ DF, FS, NT	Source	$\mathbf{93.34 \pm 0.12}$	89.07 ± 4.88	89.64 ± 0.84	76.60 ± 1.02	79.96 ± 1.36
	Target	$\mathbf{76.40 \pm 0.29}$	73.11 ± 6.21	61.52 ± 0.39	70.66 ± 0.74	68.42 ± 0.93

method consistently achieves the highest F1 scores on both the source and target domains. In particular, for challenging multi-target settings such as $\{DF \rightarrow FS, F2F\}$ and $\{FS \rightarrow DF, F2F, NT\}$, the proposed approach demonstrates a substantial performance margin over all competing methods on the target domains. This indicates that our model not only adapts effectively to multiple unseen target distributions (faking techniques), but also preserves strong performance on the source domain, thereby mitigating catastrophic forgetting. In contrast, supervised adaptation baselines such as *FT*, *KD*, and *TGD* exhibit noticeable degradation in target-domain performance, especially as the number of target domains increase. Although *FReTAL* performs better than other baselines in some cases, its performance remains consistently below that of the proposed method.

B. Performance on Low-Resolution Deepfakes

To evaluate robustness of our framework under the challenging setup of degraded visual quality, we perform experiments on low-resolution deepfake images (64×64). Table II reports the performance comparison across multiple multi-target domain adaptation tasks under this setting. Despite the significant loss of fine-grained visual details at lower resolutions, the proposed method achieves strong performance across the majority of tasks and outperforms the baselines on both source and target domains in 8 out of the 10 experiments. While it is marginally outperformed in 2 experiments, the proposed approach exhibits markedly more consistent behavior overall, particularly on the target domains where supervised baselines experience substantial degradation. These results suggest that the proposed framework learns more resolution-invariant and transferable feature representations, leading to improved robustness under low-resolution conditions.

C. Cross-Dataset Study

We further evaluate the generalizability of the proposed method through a cross-dataset study, where the source and target domains originate from different datasets. Table III presents results for several cross-dataset multi-target adaptation tasks involving the FaceForensics++, Celeb-DF [53], and DFDC [55] datasets. The proposed method achieves the highest F1 score consistently in all the source and target domains, compared to all the baselines, demonstrating strong generalization between datasets. Importantly, our framework successfully transfers the knowledge learned from one dataset to detect deepfakes generated by different manipulation pipelines and data distributions. Although baseline methods such as *FT* and *KD* show competitive performance in some cases, their results are less consistent and often degrade significantly on target datasets. In contrast, the proposed approach maintains stable and high performance, indicating its ability to learn dataset-invariant features while retaining discriminative power.

TABLE III

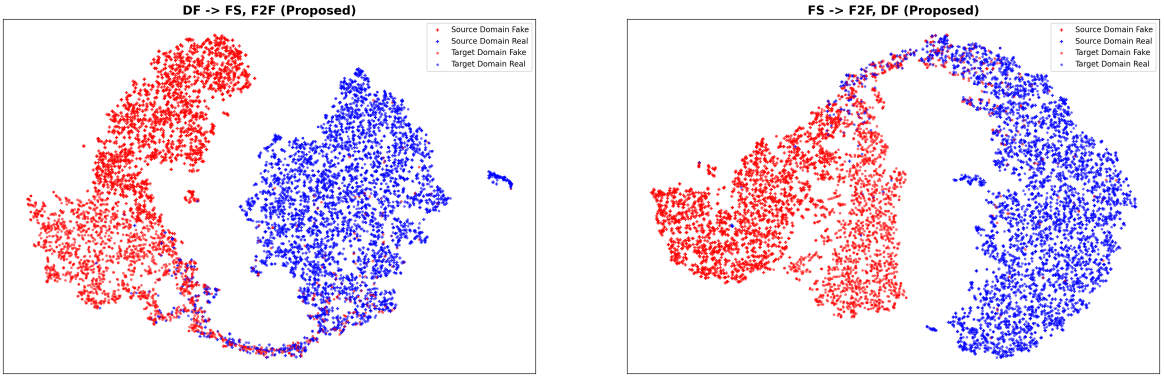
CROSS-DATASET PERFORMANCE COMPARISON FOR MULTI-TARGET DEEPFAKE DETECTION TASKS. BEST RESULTS ARE HIGHLIGHTED IN BOLD. THE NOTATION $x \rightarrow y, z$ DENOTES x AS SOURCE DATASET/DOMAIN AND y, z AS THE TARGET DATASETS/DOMAINS.

DA Task	Domain	Proposed	FReTAL	KD	FT	TGD
DF $\rightarrow$ Celeb-DF, DFDC	Source	$\mathbf{99.35 \pm 0.07}$	96.04 ± 1.82	96.30 ± 1.07	77.82 ± 1.48	90.49 ± 0.79
	Target	$\mathbf{98.25 \pm 0.59}$	97.44 ± 1.94	95.27 ± 0.64	97.39 ± 0.42	97.05 ± 1.74
FS $\rightarrow$ Celeb-DF, DFDC	Source	$\mathbf{98.31 \pm 0.89}$	56.06 ± 0.47	76.16 ± 0.41	65.09 ± 0.76	77.04 ± 1.21
	Target	$\mathbf{98.86 \pm 0.33}$	96.87 ± 0.52	91.23 ± 0.88	97.18 ± 0.61	95.78 ± 1.36
Celeb-DF $\rightarrow$ DF, DFDC	Source	$\mathbf{99.10 \pm 0.04}$	93.67 ± 1.03	78.35 ± 0.82	74.46 ± 0.91	74.82 ± 0.96
	Target	$\mathbf{96.84 \pm 0.64}$	91.67 ± 0.58	83.93 ± 0.73	88.53 ± 0.58	91.10 ± 0.64
DFDC $\rightarrow$ FS, Celeb-DF	Source	$\mathbf{99.43 \pm 0.05}$	84.43 ± 1.11	76.97 ± 0.91	74.44 ± 0.83	85.43 ± 1.02
	Target	$\mathbf{90.75 \pm 4.61}$	89.84 ± 0.66	83.66 ± 0.69	88.89 ± 0.55	91.01 ± 0.71

TABLE IV

ABLATION STUDY ANALYZING THE CONTRIBUTION OF INDIVIDUAL LOSS COMPONENTS. $\mathcal{L}_{\text{sup}}$ DENOTES THE SUPERVISED BCE LOSS, $\mathcal{L}_{\text{moment}}$ THE MOMENT MATCHING LOSS, AND $\mathcal{L}_{\text{CA}}$ THE CLASS ALIGNMENT LOSS. BEST RESULTS ARE HIGHLIGHTED IN BOLD.

DA Task	Loss Function			
	Proposed	Proposed w/o $\mathcal{L}_{\text{moment}}$	Proposed w/o $\mathcal{L}_{\text{CA}}$	Proposed w/o $\mathcal{L}_{\text{moment}}$ and $\mathcal{L}_{\text{CA}}$
DF $\rightarrow$ FS, F2F	$\mathbf{92.27 \pm 0.41}$	90.67 ± 0.91	91.26 ± 1.14	82.28 ± 1.87
FS $\rightarrow$ F2F, DF	$\mathbf{94.51 \pm 0.32}$	92.41 ± 1.34	91.60 ± 2.24	88.56 ± 1.45



(a) $DF \rightarrow FS, F2F$ (b) $FS \rightarrow F2F, DF$

Fig. 1. *t-SNE* visualization results. The data from the source domain are represented with a plus sign. The data from all the target domains are represented with a circle. Blue color denotes the *Genuine* class; red color denotes the *Fake* class. Best viewed in color.

D. Ablation Study

We conducted an ablation study to assess the effects of the individual components in our loss function in Equation (5). The results for two multi-target experiments are reported in Table IV. We note that removing the moment matching loss term decreases the F1 score, showing its importance in aligning the source and target domains. Removing the class alignment loss term also results in a drop in the F1 score, demonstrating its importance in learning from the unlabeled samples in the target domains. Removing both the moment matching and class alignments loss terms results in the maximum reduction in the performance. The best results are obtained when all the three terms are used in the loss function, as in our method.

E. Feature Visualization

In this experiment, we utilized *t-SNE* embeddings to visualize the features learned by our framework. The results for two different DA tasks are depicted in Figure 1. The data from the source domain are represented with a plus sign; the data from all the target domains are represented with a circle. Blue color denotes the *Genuine* class; red color denotes the *Fake* class. As evident from the figure, the proposed method aggregates the fake samples from the source and target domains into one cluster (red) and the genuine samples from the source and target domains into a separate cluster (blue). Thus, the moment matching and the class alignment loss terms in our framework enable the deep CNN to learn discriminating feature representations that minimize the disparity between the source and all the target domains, and also separate the genuine and fake images from all the domains.

F. Effect of Varying Size of the Target Dataset

We study the effect of varying the size of the target dataset on detection performance. Table V presents results for different sizes of labeled target data. As shown in the table, increasing the number of labeled target samples con-

TABLE V
EFFECT OF VARYING SIZE OF THE LABELED DATA IN THE TARGET DOMAIN (DF → FS, F2F). BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Labeled Size	Domain	Method				
		Proposed	**FReTAL**	**KD**	**TGD**	**FT**
0	Source	$\mathbf{96.88 \pm 0.25}$	N/A	N/A	N/A	N/A
	Target	$\mathbf{70.13 \pm 0.24}$	N/A	N/A	N/A	N/A
800	Source	$\mathbf{96.64 \pm 0.25}$	90.10 ± 0.95	72.80 ± 1.05	67.40 ± 1.20	64.10 ± 1.25
	Target	$\mathbf{90.86 \pm 0.06}$	78.20 ± 1.90	82.60 ± 1.05	73.30 ± 1.70	80.10 ± 1.35
1600	Source	$\mathbf{97.73 \pm 0.29}$	90.80 ± 0.80	73.70 ± 0.90	68.10 ± 1.05	65.30 ± 1.10
	Target	$\mathbf{92.13 \pm 1.20}$	84.10 ± 1.35	80.40 ± 1.20	78.90 ± 1.25	85.20 ± 1.05
2400	Source	$\mathbf{98.26 \pm 0.72}$	91.20 ± 0.72	74.10 ± 0.82	68.60 ± 0.96	66.00 ± 1.00
	Target	$\mathbf{93.94 \pm 1.74}$	87.30 ± 1.10	76.10 ± 1.60	81.50 ± 1.10	88.10 ± 0.92

TABLE VI
EFFECT OF VARYING SIZE OF THE UNLABELED DATA IN THE TARGET DOMAIN (DF → FS, F2F).

Method	**Domain**	**4,000**	**8,000**	**12,000**
Proposed	Source	96.96 ± 0.21	97.48 ± 1.16	97.17 ± 0.21
	Target	91.72 ± 0.10	92.14 ± 0.97	93.42 ± 0.10

sistently improves performance across all methods. However, the proposed method achieves strong performance even with limited labeled data. It also significantly outperforms all the competing baselines across all sizes of the labeled target set. More importantly, it depicts impressive performance even when the target domains are completely unsupervised, and do not contain any labeled samples. The baseline methods, on the other hand, are unable to operate in the absence of labeled samples in the target domains.

Table VI evaluates the impact of varying the number of unlabeled target domain samples. We report the results of our method only, since the baselines cannot operate in the absence of labeled target domain samples. The results show that increasing the amount of unlabeled target data leads to steady performance improvements for the proposed method. This highlights the effectiveness of leveraging unlabeled samples through the proposed alignment strategy and underscores the framework's suitability for scenarios where labeled data is scarce but unlabeled data is abundant.

V. CONCLUSION AND FUTURE WORK

While generative AI technology has pushed the boundaries of AI research, it has also made deepfakes commonplace and easy to generate. Reliably detecting deepfakes has thus attracted significant attention in the research community. While CNNs have depicted promise, their performance drops drastically when applied to deepfakes created using unknown techniques (out-of-domain data). In this paper, we proposed a novel multi-target domain adaptation framework using deep learning for robust detection of deepfakes. Our framework can detect deepfakes generated using multiple unknown faking techniques; more importantly, it can operate in scenarios where only unlabeled samples are available from the unknown faking techniques (target domains). Our extensive empirical studies corroborated the promise and potential of our framework against competing baselines, in several challenging real-world scenarios. To our knowledge, this is the first research effort to study the performance of multi-target deep domain adaptation techniques for deepfake detection. As part of future research, we plan to study the performance of our framework using transformer based architectures [56] as the underlying base model; we also plan to extend our framework to multi-source multi-target scenarios, where both the source and the target domains can contain deepfakes generated using multiple image manipulation techniques.

REFERENCES

[1] C. Bhattacharyya, H. Wang, F. Zhang, S. Kim, and X. Zhu, "Diffusion deepfake," in *arXiv:2404.01579v1*, 2024.

[2] M. Kowalski, "Faceswap - github repository," https://github.com/MarekKowalski/FaceSwap, 2016, accessed: March 14, 2024.

[3] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner, "Faceforensics++: Learning to detect manipulated facial images," in *IEEE/CVF International conference on computer vision (ICCV)*, 2019.

[4] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[5] A. Pumarola, A. Agudo, A. Martinez, A. Sanfeliu, and F. Moreno-Noguer, "Ganimation: Anatomically-aware facial animation from a single image," in *European Conference on Computer Vision (ECCV)*, 2018.

[6] S. Suwajanakorn, S. Seitz, and I. Kemelmacher-Shlizerman, "Synthesizing obama: learning lip sync from audio," *ACM Transactions on Graphics*, vol. 36, no. 4, 2017.

[7] J. Vincent, "Watch jordan peele use ai to make barack obama deliver a psa about fake news," *The Verge*, vol. 17, 2018.

[8] D. Patterson, "President's words used to create "deepfakes" at davos," *Video*, 2020.

[9] J. DelViscio, "A nixon deepfake, a "moon disaster" speech and an information ecosystem at risk," *Scientific American*, vol. 20, 2020.

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

[11] S. Agarwal, H. Farid, Y. Gu, M. He, K. Nagano, and H. Li, "Protecting world leaders against deep fakes," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019.

[12] L. Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wen, and B. Guo, "Face X-Ray for more general face forgery detection," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[13] R. Salloum, Y. Ren, and C. J. Kuo, "Image splicing localization using a multi-task fully convolutional network (MFCN)," *Journal of Visual Communication and Image Representation*, vol. 51, pp. 201–209, 2018.

[14] X. Yang, Y. Li, and S. Lyu, "Exposing deep fakes using inconsistent head poses," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 8261–8265.

[15] P. Zhou, X. Han, V. I. Morariu, and L. S. Davis, "Two-stream neural networks for tampered face detection," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 2017, pp. 1831–1839.

[16] ——, "Learning rich features for image manipulation detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 1053–1061.

[17] M. Zhang, H. Wang, P. He, A. Malik, and H. Liu, "Improving GAN-generated image detection generalization using unsupervised domain adaptation," in *IEEE International Conference on Multimedia and Expo (ICME)*, 2022.

[18] S. Tariq, S. Lee, and S. S. Woo, "One detector to rule them all: Towards a general deepfake attack detection framework," in *Proceedings of the web conference (WWW)*, 2021, pp. 3625–3637.

[19] D. Cozzolino, J. Thies, A. Rossler, C. Riess, M. Niebner, and L. Verdoliva, "Forensictransfer: Weakly-supervised domain adaptation for forgery detection," in *arXiv:1812.02510v2*, 2019.

[20] M. Kim, S. Tariq, and S. S. Woo, "Fretal: Generalizing deepfake detection using knowledge distillation and representation learning," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2021, pp. 1001–1012.

[21] R. Kemker, M. McClure, A. Abitino, T. Hayes, and C. Kanan, "Measuring catastrophic forgetting in neural networks," in *AAAI Conference on Artificial Intelligence*, 2018.

[22] B. Thompson, J. Gwinnup, H. Khayrallah, K. Duh, and P. Koehn, "Overcoming catastrophic forgetting during domain adaptation of neural machine translation," in *North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 2062–2068.

[23] Y. Xu, X. Zhong, A. Yepes, and J. Lau, "Forget me not: Reducing catastrophic forgetting for domain adaptation in reading comprehension," in *IEEE International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2020, pp. 1–8.

[24] S. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, vol. 22, no. 10, 2010.

[25] T. Wang, X. Liao, K. P. Chow, X. Lin, and Y. Wang, "Deepfake detection: A comprehensive survey from the reliability perspective," *ACM Computing Surveys*, vol. 57, no. 3, 2024.

[26] G. Pei, J. Zhang, M. Hu, Z. Zhang, C. Wang, Y. Wu, G. Zhai, J. Yang, C. Shen, and D. Tao, "Deepfake generation and detection: A benchmark and survey," in *arXiv:2403.17881v4*, 2024.

[27] A. Malik, M. Kuribayashi, S. Abdullahi, and A. Khan, "Deepfake detection for human face images and videos: A survey," *IEEE Access*, vol. 10, 2022.

[28] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner, "Faceforensics: A large-scale video dataset for forgery detection in human faces," in *arXiv:1803.09179*, 2018.

[29] P. Singhal, R. Walambe, S. Ramanna, and K. Kotecha, "Domain adaptation: Challenges, methods, datasets, and applications," *IEEE Access*, vol. 11, 2023.

[30] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[31] M. Long, Y. Cao, J. Wang, and M. Jordan, "Learning transferable features with deep adaptation networks," in *International Conference on Machine Learning (ICML)*, 2015.

[32] H. Venkateswara, J. Eusebio, S. Chakraborty, and S. Panchanathan, "Deep hashing network for unsupervised domain adaptation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[33] A. Nguyen, T. Tran, Y. Gal, P. Torr, and A. Baydin, "Kl guided domain adaptation," in *International Conference on Learning Representations (ICLR)*, 2022.

[34] C. Shui, Q. Chen, J. Wen, F. Zhou, C. Gagne, and B. Wang, "A novel domain adaptation theory with jensen-shannon divergence," *Knowledge-Based Systems*, vol. 257, 2022.

[35] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *Journal of Machine Learning Research (JMLR)*, vol. 17, 2016.

[36] D. Kingma and M. Welling, "Auto-encoding variational bayes," in *arXiv preprint arXiv:1312.6114*, 2013.

[37] J. Shen, Y. Qu, W. Zhang, and Y. Yu, "Wasserstein distance guided representation learning for domain adaptation," in *Association for the Advancement of Artificial Intelligence (AAAI)*, 2018.

[38] T. Isobe, X. Jia, S. Chen, J. He, Y. Shi, J. Liu, H. Lu, and S. Wang, "Multi-target domain adaptation with collaborative consistency learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 8187–8196.

[39] S. Roy, E. Krivosheev, Z. Zhong, N. Sebe, and E. Ricci, "Curriculum graph co-teaching for multi-target domain adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 5351–5360.

[40] Y. Xu, P. Ghamisi, and Y. Avrithis, "Multi-target unsupervised domain adaptation for semantic segmentation without external data," in *arXiv:2405.06502v1*, 2024.

[41] J. Shin, M. Seok, S. Kim, and E. Park, "Merge-friendly post-training quantization for multi-target domain adaptation," in *International Conference on Machine Learning (ICML)*, 2025.

[42] Y. Lu, H. Huang, B. Zeng, Z. Lai, and X. Li, "Multi-source and multi-target domain adaptation based on dynamic generator with attention," *IEEE Transactions on Multimedia*, vol. 26, 2024.

[43] M. Kim, S. Tariq, and S. Woo, "Cored: Generalizing fake media detection with continual representation using distillation," in *ACM International Conference on Multimedia*, 2021, pp. 337–346.

[44] B. Chen and S. Tan, "Featuretransfer: Unsupervised domain adaptation for cross-domain deepfake detection," *Security and Communication Networks*, 2021.

[45] S. Seraj, A. Singh, and S. Chakraborty, "Semi-supervised deep domain adaptation for deepfake detection," in *Workshop on Manipulation, Adversarial, and Presentation Attacks in Biometrics at IEEE Winter Conference on Applications of Computer Vision (WACV-W)*, 2024.

[46] M. S. Seraj and S. Chakraborty, "Multi-source deep domain adaptation for deepfake detection," in *International Conference on Pattern Recognition (ICPR)*, 2024.

[47] J. Liu, W. Chang, Y. Wu, and Y. Yang, "Deep learning for extreme multi-label text classification," in *ACM SIGIR Conference on Information Retrieval*, 2017.

[48] Y. Li, K. Swersky, and R. Zemel, "Generative moment matching networks," in *International Conference on Machine Learning (ICML)*, 2015.

[49] Y. Mroueh, T. Sercu, and V. Goel, "McGAN: Mean and covariance feature matching GAN," in *International Conference on Machine Learning (ICML)*, 2017.

[50] B. Sun, J. Feng, and K. Saenko, "Return of frustratingly easy domain adaptation," in *AAAI Conference on Artificial Intelligence*, 2016.

[51] W. Zellinger, T. Grubinger, E. Lughofer, T. Natschlager, and S. Saminger-Platz, "Central moment discrepancy (cmd) for domain-invariant representation learning," in *International Conference on Learning Representations (ICLR)*, 2017.

[52] X. Peng, Q. Bai, X. Xia, Z. Huang, K. Saenko, and B. Wang, "Moment matching for multi-source domain adaptation," in *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[53] Y. Li, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A large-scale challenging dataset for deepfake forensics," in *IEEE Conference on Computer Vision and Patten Recognition (CVPR)*, 2020.

[54] H. Jeon, Y. Bang, J. Kim, and S. S. Woo, "TGD: Transferable GAN-generated images detection framework," in *arXiv:2008.04115*, 2020.

[55] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. C. Ferrer, "The deepfake detection challenge (dfdc) dataset," in *arXiv:2006.07397v4*, 2020.

[56] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Neural Information Processing Systems (NeurIPS)*, 2017.