Quality-Agnostic Deepfake Detection with Intra-model Collaborative Learning

Binh M. Le
Dept. of Computer Science & Engineering
Sungkyunkwan University
Suwon, South Korea
bmle@g.skku.edu

Simon S. Woo*
Dept. of Computer Science & Engineering
Sungkyunkwan University
Suwon, South Korea
swoo@g.skku.edu

Abstract

Deepfake has recently raised a plethora of societal concerns over its possible security threats and dissemination of fake information. Much research on deepfake detection has been undertaken. However, detecting low quality as well as simultaneously detecting different qualities of deepfakes still remains a grave challenge. Most SOTA approaches are limited by using a single specific model for detecting certain deepfake video quality type. When constructing multiple models with prior information about video quality, this kind of strategy incurs significant computational cost, as well as model and training data overhead. Further, it cannot be scalable and practical to deploy in real-world settings. In this work, we propose a universal intra-model collaborative learning framework to enable the effective and simultaneous detection of different quality of deepfakes. That is, our approach is the quality-agnostic deepfake detection method, dubbed QAD. In particular, by observing the upper bound of general error expectation, we maximize the dependency between intermediate representations of images from different quality levels via Hilbert-Schmidt Independence Criterion. In addition, an Adversarial Weight Perturbation module is carefully devised to enable the model to be more robust against image corruption while boosting the overall model's performance. Extensive experiments over seven popular deepfake datasets demonstrate the superiority of our QAD model over prior SOTA benchmarks.

1. Introduction

Deep learning approaches for facial manipulation, such as deepfakes, have recently received considerable attention [54, 31, 61, 25, 20, 23], because they can be abused for the malicious purposes such as fake news, pornography, etc. Due to the advancements made in Generative Adversarial Networks and other deep learning-based computer vision algorithms, deepfakes have also become more realistic and natural, making it harder not only for humans, but also for classifiers to tell them apart. Moreover, it has been simpler than ever before to create convincing deepfakes using simple programs and apps without requiring advanced machine learning knowledge. Such easy-to-create and realistic fake images and videos can be maliciously exploited, raising significant security, privacy, and societal concerns such as fake news propagation [39], and stealing personal information via phishing and scams [10].

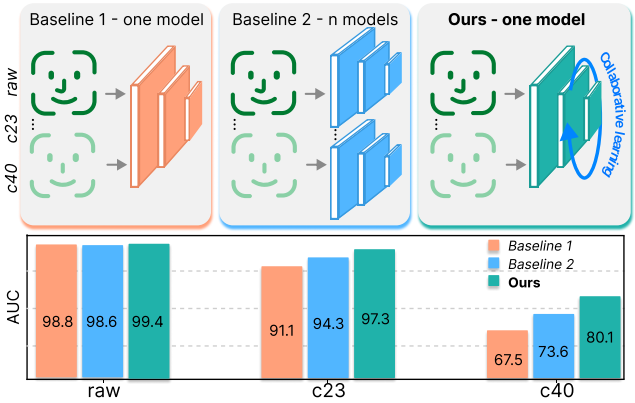

Figure 1. **A summary of our goal.** Our approach stands out from previous works that detect deepfakes using separate models for different qualities (*e.g.* Baseline 2 [28]) or a single model without considering the interaction between qualities (*e.g.* Baseline 1 [43]). Instead, our method employs all quality levels and improves the performance of the model on each quality level, leading to overall enhanced performance.

To mitigate such problems caused by deepfakes, there has been a tremendous research effort put into constructing reliable detectors [32, 25, 8, 38, 61, 44]. Although they have achieved outstanding performance with high-quality deepfakes, most of them have failed to detect low-quality deepfakes effectively [8, 43]. While video compression steps do not significantly impact on visualization, it drastically drop deepfake detectors' performance on low-quality deepfakes (c40). A handful of research has been focused on detecting low-quality deepfakes such as ADD [28] and BZNet [29]. However, their methods can only detect low-quality com-

*Corresponding author.

pressed deepfakes. And, those prior approaches expose a critical problem, when deployed in practice since the prior video quality information of the input is unknown. Moreover, developing different models for each input quality induces significant computational overhead. Other works, such as LipForensics [16], also attempt to make their detectors robust against various corruptions and compression. Nevertheless, it is unable to detect image-based deepfakes with random lossy compression like JPEG.

In this research, we propose the novel deepfake detection method, QAD, which can simultaneously detect both high and low-quality (quality-agnostic) deepfakes in a single model, as illustrated in Fig. 1. Especially, we propose a universal intra-model collaborative learning framework to provide the effective detection of different quality deepfakes. We modulate the conventional model-based collaborative learning [47] to an instance-based intra-model collaborative learning framework in our training. During the training phase, our single model simultaneously learns the representations of one image, but with different qualities. By utilizing the collaborative learning framework, our QAD can align the distributions of high and low-quality image representations to be geometrically similar. Hence, it can avoid the overfitting caused by compressed images and the overconfidence caused by raw images, while boosting its overall performance.

In particular, we perform a rigorous theoretical analysis, and show that the low-quality deepfake classification error can be bounded by two terms: classification loss and the distance between the representations of high and low-quality images. Instead of using a direct pairwise regularization to minimize the gaps between the high and low-quality image representations, we propose to apply *Hilbert-Schmidt Independence Criterion* (HSIC) to maximize the dependence between a mini-batch of high and low-quality images, thus maximizing the mutual information between them, and supporting the high-level representations and effective output predictions. Meanwhile, to enhance the model's robustness under heavy input compression, we propose *Adversarial Weight Perturbation* (AWP) [56, 3], which can further flatten the weight loss landscape of the model, bridging the gap in multiple quality learning for deepfake detection.

Finally, we conduct extensive experiments to show the effectiveness of our QAD with seven different popular benchmark datasets. We first show that our method can outperform previous baselines when training with data from various video and image compression qualities. Furthermore, we show that our QAD exceeds the performance of the SOTA quality-aware models such as BZNet [29] by a significant margin, while requiring remarkably fewer computational parameters and no prior knowledge of the inputs. Our contributions are summarized as follows:

1) We theoretically analyze and prove that the classification error of low-quality deepfakes can be bounded by its classification loss and the representation distance with its corresponding high-quality images.

2) We propose a unified quality-agnostic deepfake detection framework (QAD), utilizing instance-based intra-model collaborative learning. We use the *Hilbert-Schmidt Independence Criterion* (HSIC) to maximize the geometrical similarity between intermediate representations of high and low-quality deepfakes, and *Adversarial Weight Perturbation* (AWP) to make our model robust under varying input compression.

3) We demonstrate that our approach outperforms well-known baselines, including the total of *eight* quality-agnostic and quality-aware SOTA methods with *seven* popular benchmark datasets.

2. Related works

2.1. Deepfake detection

Recently, deepfakes have been of the utmost crucial because they can cause serious security and privacy threats. Threrefore, a large number of detection methods have been proposed to effectively identify such deepfakes [43, 31, 30, 22, 41, 54, 32, 25, 8, 61]. However, the majority of the aforementioned works focus on mining visual artifacts of deepfakes, such as the blending boundaries of generated faces [31], the irregularity of pupil shapes [14], the spatiotemporal inconsistency [6, 44], or exploring deep learning-based attention methods [61] to identify such artifacts. Meanwhile, several approaches also showed that exposing deepfakes in the frequency domain is effective. Such methods include analyzing the discrepancies of frequency spectrum [8, 27, 25], employing the checkerboard artifacts caused by the transposed convolutional operator [60, 11], or mining the statistical frequency features with dual deep learning models [38]. Nevertheless, such models' performance substantially decreases when encountering low-quality compressed images. To remedy the above shortcoming, recent studies proposed methods to detect the deepfake in highly compressed low-quality versions such as [28], which utilized a knowledge distillation. Also, [29] presented a supper-resolution-based network for enhancing the performance of low-quality deepfake detection. However, all of the aforementioned approaches are limited in developing a single model for each quality of deepfakes, which is impractical to deploy in real-world scenarios due to the requirement of prior knowledge of the input quality.

2.2. Collaborative learning

Collaborative learning proposed by [47] is designed to achieve a global minimum of a deep neural network, while maintaining the same computational complexity at inference time as at training time. Collaborative learning inherits

the advantages of auxiliary training [48], multi-task learning [58], and knowledge distillation [18]. Its applications cover supporting weakly-supervised learning [24], or integrating with online knowledge distillation [57, 15]. And, its training graph is divided into two or more sub-networks to ensure global minimum achievement [47]. Besides, [9] proposed an intra-model collaborative learning framework that shares a similar characteristic with self-knowledge distillation. However, all of the approaches are model-based collaborative learning, in which a single input generates multiple outputs (or *views*) through multiple classifier heads of one target network in both training and inference phase.

In this work, we distinguish ourselves by deploying the collaborative learning framework for simultaneously training deepfakes of various qualities with an undeviated single model, namely *instance-based collaborative learning*. Different from conventional mini-batch stochastic optimization, which independently samples random images from different qualities and optimizes the detector, our collaborative learning approach allows us to utilize the common features in the same image but from different qualities simultaneously. Thus, our deepfake detector circumvent the overfitting caused by compressed images or the overconfidence from raw images, enhancing its overall performance.

2.3. Hilbert-Schmidt Independence Criterion

The Hilbert-Schmidt Independence Criterion (HSIC) [13] measures the statistical dependency between probability distributions. In fact, HSIC differs from the covariance, where $Cov(X, Y) = 0$ does not imply that *two* random variables X and Y are independent [42], while HSIC shows its tractable computation and equivalency in terms of the independence property [13]. Moreover, HISC is easy to be estimated statistically and algorithmically. In practice, applications based on HSIC are found in a variety of practical domains, including maximizing the dependencies for self-supervised learning [33] and classification learning [34, 12], or defense against model inversion attacks [37]. In this paper, we utilize the HSIC to maximize the dependency between distributions of deepfake images of different qualities at intermediate layers. Therefore, we aim to constrain low-level representations of images not to be exactly the same, but to share a geometrical similarity of learning features that can support high-level output predictions.

3. Methods

In this section, we first theoretically examine the upper bound for our optimization problem by considering a DNN classifier of K classes, and two modalities of input quality: raw (high-quality) and compressed (low-quality). Then, we discuss how to more efficiently collaborate on the representations of deepfake images of differing quality.

3.1. Preliminary & our inspiration

Given a sample x_r from a space $\mathcal{X}$ and its compressed version at quantile c, x_c can be expressed as $x_c = x_r - \delta_c$, and we define the corresponding label $y \in \{0, 1\}$ (real and fake). Next, a family $\mathcal{F}$ of learning functions $f : \mathcal{X} \to \mathbb{R}^2$ returns a 2-tuple $f(x) = [f(x, j)]_{j=1}^2$, whose $f(x, j)$ is proportional to the probability to assign x to the j-th class, and f is defined by learning parameters θ. Given a training data $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n \subset \mathcal{X} \times \mathcal{Y}$, our goal is to minimize the expectation of a loss function $\mathcal{L} : \mathbb{R}^2 \times \mathcal{Y} \to \mathbb{R}$. Here, we consider $\mathcal{L}(f(x), y) = 1 - \sigma_T(f(x, y))$, where σ_T is the softmax function with temperature $T > 0$:

$$\sigma_T(f(x), y) = \frac{\exp(f(x, y)/T)}{\sum_{k=1}^{2} \exp(f(x, k)/T)}. \tag{1}$$

Theorem 1. *(Proof can be done similarly as [36] and [19] in Supp. Material) For any $f \in \mathcal{F}$, and with probability $1 - \delta$ over the draw of $\mathcal{D}$,*

$$\begin{aligned}\mathbb{E}[\mathbb{I}\{\hat{y}(x_c) \neq y\}] \leq\ & 2\mathbb{E}_{\mathcal{D}}\mathcal{L}(f(x_c), y) \\ & + \frac{8}{T}\mathbb{E}_{\mathcal{D}}\mathcal{L}_{i\text{-}col}(f(x_r), f(x_c)) + 4\mathfrak{R}_{\mathcal{D}}(\Phi_{\mathcal{W}}) \\ & + \frac{16}{n} + \mathcal{O}\left(\sqrt{\frac{\log(2/\delta)}{2n}}\right),\end{aligned} \tag{2}$$

where $\mathfrak{R}_{\mathcal{D}}$ is the Rademacher complexity, $\Phi_{\mathcal{W}} = \{\mathcal{L}(f(x_r), y), f \in \mathcal{F}\}$, and

$$\mathcal{L}_{i\text{-}col}(f(x_r), f(x_c)) = \| f(x_r) - f(x_c) \|. \tag{3}$$

Insight of Theorem 1. On the right-hand side of Eq. 2, our classifier f depends on two terms, where the first term is the classification loss $\mathcal{L}(f(x_c), y)$ applied to the prediction of the compressed image x_c. And, the second term is the instance-based collaborative loss $\mathcal{L}_{i\text{-}col}(f(x_r), f(x_c))$ that measures the pairwise difference between predictions of the raw image and its compressed version. Therefore, minimizing the expectation over training data $\mathcal{D}$ of $2\mathcal{L}(f(x_c), y) + 8\mathcal{L}_{i\text{-}col}(f(x_r), f(x_c))/T$, can decrease the true error. Note that Eq. 2 is also general so that it can be applicable for raw images. Hence, in practice, the first term can be generalized to both raw and compressed images.

In order to minimize the expectation of errors in both raw and compressed deepfake image predictions, our theoretical analysis shows that we can minimize both classification loss and collaborative loss at the output. However, as observed by [9] and in our experiments (see Tab. 4), this instance-based collaborative learning loss fails to achieve the best performance. Additionally, training solely with highly-compressed images makes the detector prone to overfitting, yielding a considerable gap between training and test performance [28]. As a result, the major research

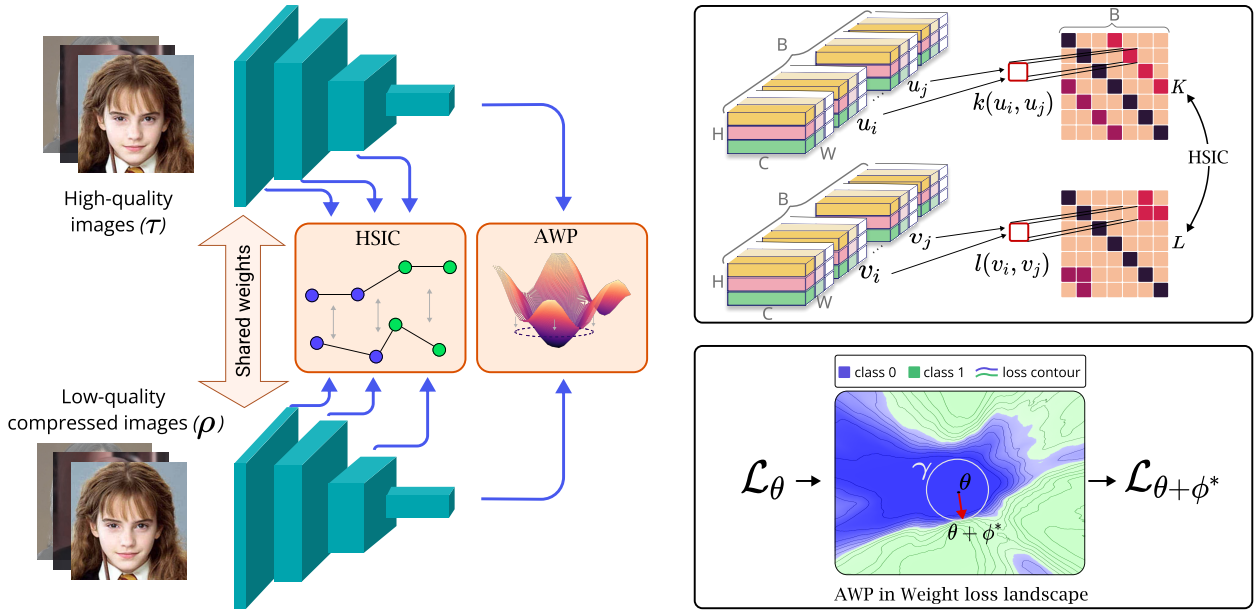

Figure 2. **Overview of our QAD framework**. A mini-batch of images from different quality modalities (*e.g.*, two in this diagram) is forwarded through a **single universal model**. Although it is one model, we pictorially split it into two branches for the reader's understanding. After training, we obtain one universal model that regulates different qualities. **Top-right**: The HSIC geometrically maximizes the dependency between images from various quality modalities at different resolutions, supporting high-level output predictions. **Bottom-right**: Through searching for the worst-case parameters' corruption and compensating for input corruption (compression), the AWP flattens the model's weight loss landscape, making the model robust under varying input compression.

challenge is how we can further lower the sensitivity of f_θ to x at various compression settings and push their representations close to each other.

Classification loss. We can construct a robust model under input corruption by flattening the weight loss landscape. In particular, we apply the *Adversarial Weight Perturbation* [56] to search for the worst-case perturbations ϕ^* of the model weights at every training step. Thereafter, optimizing the perturbed model via $\mathcal{L}(f_{\theta+\phi^*}, y)$ can enable it to be more robust under varying input image corruptions/distortions, which represent the varying qualitiies of deepfake inputs (See Fig. 2—Bottom-right panel). Furthermore, the classification loss in Eq. 2 is upper bounded by this new loss function due to the worst-case perturbations.

Collaborative loss. With respect to the collaborative learning loss, Eq. 3 shares similar characteristics with recent knowledge distillation research [18, 52]. However, our goal is to develop and train a single universal model, not having a teacher-student relationship. Nevertheless, we argue that the gap between raw and compressed image representations can be minimized more efficiently by regularizing their discrepancy at the low-level representations. Moreover, enforcing the similarity of the pairwise representations at the intermediate layers with an input difference of $|\delta_c|$ can collapse layers' weights to zero or can lead a deep model to remember training data instead of learning discriminative features. Therefore, we relax this constraint by maximizing the kernel dependency in a mini-batch of data between the raw and compressed image representations by the *Hilbert-Schmidt Independence Criterion* (HSIC). Using HSIC, we can instead enforce the geometrical structures of mini-batch of raw data and the compressed data to be similar, so that we can still effectively detect different-quality deepfakes in a single model. From a mutual information perspective, maximizing the kernel dependency can enforce the mutual information of the learned representations of different compression ratios, thus regularizing the detector to be more generalized (See Fig. 2—Top-right panel).

3.2. Details of our methods

3.2.1 Weight loss landscape flattening

Recent studies [21, 55] suggest that searching for a flatter minima can improve the generalization ability of the model. To achieve this, we propose using Adversarial Weight Perturbation [56] to identify flat local minima of the empirical risk. The worst-case perturbations ϕ^* of the model weights which increases the loss dramatically is formulated as:

$$\phi^* = \underset{\phi \in \mathcal{B}_p(\theta,\gamma)}{\arg\max}\, \mathcal{L}(f_{\theta+\phi}(x), y), \tag{4}$$

where $\mathcal{B}_p(\theta, \gamma) = \{\upsilon \in \Theta : \| \theta - \upsilon \|_p \leq \gamma\}$ is the feasible region of any perturbation ϕ. The AWP adds the worst-case perturbation to the model weight, so that $\mathcal{L}(f_{\theta+\phi^*}(x), y)$

becomes the supremum value in $\mathcal{B}_p(\theta, \gamma)$. Therefore, optimizing $\mathcal{L}(f_{\theta+\phi^*}(x), y)$ pushes θ adjust its values such that the loss landscape is more flatten with the same capability $\mathcal{B}_p(\theta, \gamma)$. As a result, f become more stable under input image's changes.

Similar to adversarial example perturbation [35], ϕ^* is generated by projected gradient method as follows:

$$\phi^* \leftarrow \Pi_\theta^\gamma \left(\phi + \eta \frac{\nabla \mathcal{L}(f_{\theta+\phi}(x), y)}{\| \nabla \mathcal{L}(f_{\theta+\phi}(x), y) \|} \| \theta \| \right), \quad (5)$$

where Π_θ^γ is an operator that projects its input into the feasible region $B_p(\theta, \gamma)$, and $\eta \in \mathbb{R}$ is the step size. In fact, we empirically find that using a **one-step** projection for ϕ^* is sufficient for the model's robustness under the image corruptions formed by compression. By adding ϕ^* in Eq. 5 to θ, it is straightforward to bound $\mathcal{L}(f(x_c), y)$ in Eq. 2.

3.2.2 Intra-model collaborative learning

Hilbert-Schmidt Independence Criterion (HSIC). Let $\mathcal{T}$ and $\mathcal{G}$ be two separable Reproducing Kernel Hilbert Spaces (RKHS) on metric spaces $\mathcal{U}$ and $\mathcal{V}$, respectively. HSIC measures the dependency between two random variables U and V from a joint distribution on $\mathcal{U}$ and $\mathcal{V}$, by evaluating the cross-covariance of the nonlinear transformations of the two random variables:

$$HSIC(U, V) = \| \mathbb{E}[\zeta(U)\psi(V)^T] - \mathbb{E}\zeta(U)\mathbb{E}\psi(V)^T \|_{HS}^2, \quad (6)$$

where $\| \cdot \|_{HS}$ is the Hilbert-Schmidt norm, which becomes the Frobenius norm in finite dimensions. And, $\zeta : \mathcal{U} \to \mathcal{T}$ and $\psi : \mathcal{V} \to \mathcal{G}$ are nonlinear mapping functions. With appropriate transformation ζ and ψ, HSIC is a dependence test, which can identify the *nonlinear dependencies* between U and V as follows: $HSIC(U, V) = 0 \Leftrightarrow U \perp V$.

Also, inner products in $\mathcal{T}$ and $\mathcal{G}$ are formed by positive definite kernel functions: $k(u, u') = \langle \zeta(u), \zeta(u') \rangle_\mathcal{T}$ and $l(v, v') = \langle \psi(v), \psi(v') \rangle_\mathcal{G}$. And, let (U', V') and (U'', V'') be independent copies of (U, V), then Eq. 6 can be expressed as follows:

$$\begin{aligned} HSIC(U, V) &= \mathbb{E}[k(U, U')l(V, V')] \\ &- 2\mathbb{E}[k(U, U')]\mathbb{E}[l(V, V'')] + \mathbb{E}[k(U, U')]\mathbb{E}[l(V, V')]. \end{aligned} \quad (7)$$

Estimation of HSIC. The empirical estimation of HSIC with an bias of $\mathcal{O}(\frac{1}{n})$ using n samples *i.i.d* drawn $\{(u_i, v_i)\}_{i=1}^n$ from the joint distribution (U, V) is provided as follows [13]:

$$\widehat{HSIC}(U, V) = \frac{1}{(n-1)^2} \mathbf{tr}(KHLH), \quad (8)$$

where $K_{i,j} = k(u_i, u_j)$, and $L_{i,j} = l(v_i, v_j)$ are kernel matrices for the kernels k, and l, respectively, and $H_{i,j} = \delta_{i,j} - \frac{1}{n}$ is a centering matrix. Regarding the kernel functions, Theorem 4 in [13] suggested that an universal kernel, such as Laplace and Gaussian RBF kernel, can guarantee the HSIC to detect any dependency between U and V.

Algorithm 1 QAD: Quality-Agnostic Deepfake detection.

Require: DNN f parameterized by θ, training dataset $\mathcal{D}$ with M quality modalities $\mathcal{T} = \{r, c_1, ...c_{M-1}\}$. Learning rate α_l and mini-batch size of B. Model weight perturbation size γ, step size η, and the number of steps K. Layers of f to apply $HSIC$ L.

while not converged **do**
  for mini-batch $(X = [X_\tau]_{\tau \in \mathcal{T}}, Y) \in \mathcal{D}$ **do**
    `# One-step AWP`
    $\mathcal{L}_1 = \mathcal{L}(f_{\theta+\phi}(X), Y)/MB$
    $\phi \leftarrow \Pi_\theta^\gamma \left(\phi + \eta \frac{\nabla \mathcal{L}_1}{\|\nabla \mathcal{L}_1\|} \| \theta \| \right)$
    $\theta \leftarrow \theta + \phi$
    $\mathcal{L}_1 \leftarrow \mathcal{L}(f_\theta(X), Y)/MB$
    `# intermediate reps.`
    $[Z_l^\tau]_{l \in L}^{\tau \in \mathcal{T}} := f_\theta(X)$
    `# HSIC: dependence maximization`
    $\mathcal{L}_2 \leftarrow \sum_{\tau, \rho \in \mathcal{T}}^{\tau \neq \rho} \mathcal{L}_{col}(\tau, \rho)$
    `# Overall loss Eq. (10)`
    $\mathcal{L}_{QAD} \leftarrow \mathcal{L}_1 + \alpha \times \mathcal{L}_2$
    $\theta \leftarrow \theta - \alpha_l \cdot \nabla_\theta \mathcal{L}_{QAD}$
    $\theta \leftarrow \theta - \phi$
  end for
end while

HSIC for maximizing the geometrical similarity. Let τ and ρ be two different qualities of deepfakes, (e.g., raw vs. compressed). Consider the l-th layer of the learning network f, we denote the learning features of a mini-batch of B images from τ and ρ are $Z_l^\tau = \{u_i\}_B$, and $Z_l^\rho = \{v_i\}_B$, respectively, where $u_i, v_i \in \mathbb{R}^{H \times W \times C}$ and H, W and C are the height, width, and channel number. Our regularization aims to maximize the dependency between Z_l^τ and Z_l^ρ via a mini-batch of representations. In other words, we try to minimize the following loss:

$$\mathcal{L}_{col}(\tau, \rho) = -\sum_{l \in L} \widehat{HSIC}(Z_l^\tau, Z_l^\rho), \quad (9)$$

where L is a predetermined collection of layers to apply the collaborative loss. And, the computational complexity for calculating Eq. 9 is $\mathcal{O}(B^2 L)$, which can be reduced to $\mathcal{O}(BL)$ when applying random Fourier features [40].

3.3. End-to-end training loss

Given a training mini-batch B that include all M quality modalities $\mathcal{T} = \{r, c_1, ...c_{M-1}\}$, the overall collaborative

Model	Test Set AUC (%)							
	NT	DF	F2F	FS	FSH	CDFv2	FFIW10K	*Avg*
Video Compression (raw + c23 + c40 of test set)								
MesoNet [1]$^{\diamondsuit}$	70.24	93.72	94.15	85.17	96.00	80.52	94.56	*87.77*
Rössler *et al.* [43] $^{\diamondsuit}$	89.64	99.05	97.89	98.83	98.50	97.49	**99.17**	*97.22*
F^3Net [38] $^{\diamondsuit}$	86.79	98.73	96.32	97.82	97.45	95.06	97.94	*95.73*
MAT [61]$^{\diamondsuit}$	86.79	98.73	96.32	97.82	97.45	95.06	97.94	*95.73*
Fang & Lin [9]	89.30	98.98	97.33	98.43	98.66	96.58	98.94	*96.89*
SBIs [45]†	78.33	95.19	79.74	80.37	80.48	-	-	*82.82*
BZNet [29]†	80.12	98.81	94.10	97.71	-	-	-	91.01
ADD [28]†	86.26	96.23	90.62	95.57	95.94	-	-	*92.92*
QAD-R (***ours***)	91.25	**99.54**	98.34	99.01	99.12	98.36	99.10	*97.82*
QAD-E (***ours***)	**94.92**	99.53	**98.94**	**99.27**	**99.12**	**98.38**	99.16	***98.47***

Table 1. **Classification performance in the quality-agnostic setting with video compression of test set**. The methods are trained using one of three approaches: simultaneously with three modalities (raw + c23 + c40), individually with each of the three modalities, or with a mid-level of compression (c23) to prevent performance degradation resulting from lossy compression. In the inference phase, ***video compression*** is applied to the input. The best results are highlighted in **bold**. † and ♢ indicate results were obtained from methods' pre-trained weights and published code, respectively.

Model	Test Set AUC (%)							
	NT	DF	F2F	FS	FSH	CDFv2	FFIW10K	*Avg*
Random Image Compression (JPEG on raw of test set)								
MesoNet [1]$^{\diamondsuit}$	70.23	92.02	88.32	82.60	91.84	81.12	91.87	*85.43*
Rössler *et al.* [43]$^{\diamondsuit}$	69.89	98.62	94.97	96.66	96.76	96.98	98.81	*93.24*
F^3Net [38]$^{\diamondsuit}$	70.95	97.89	92.83	96.34	94.72	95.44	97.19	*92.19*
MAT [61] $^{\diamondsuit}$	69.53	98.96	**95.53**	97.99	96.97	98.21	98.91	*93.73*
Fang & Lin [9]	75.49	98.32	94.63	97.64	97.28	96.67	98.39	*94.06*
SBIs [45]†	77.75	97.83	82.05	86.10	85.42	-	-	*85.83*
BZNet [29]†	**79.00**	98.77	95.23	97.92	-	-	-	92.73
ADD [28]†	75.84	96.83	92.23	95.24	96.00	-	-	*91.23*
QAD-R (***ours***)	75.18	98.86	93.72	98.52	98.18	98.51	**98.96**	*94.56*
QAD-E (***ours***)	76.27	**99.20**	94.44	**98.69**	**98.60**	**98.52**	98.86	***94.94***

Table 2. **Classification performance in the quality-agnostic setting with image compression of test set**. The training approach resembles that of Table 1's setting, however, in the inference phase, ***random image compression*** is applied to the input.

learning loss in our QAD framework is formulated as:

$$\mathcal{L}_{QAD} = \frac{1}{MB} \sum_{\tau \in \mathcal{T}, i \in B} \mathcal{L}_{\phi^*}(x_{\tau,i}, y_i) + \alpha \sum_{\tau,\rho \in \mathcal{T}}^{\tau \neq \rho} \mathcal{L}_{col}(\tau, \rho), \tag{10}$$

where α is a hyper-parameter to balance contribution of each loss. It is worth noting that our QAD training loss is parameter-free, and is not affected by the order of the modalities. Further, unlike other model-based collaborative learning [47], our QAD does not derive any sub-models. In other words, it can be integrated with any backbone, *i.e.*, ResNet50, and introduces no extra computation at the inference time. Note that Theorem 1 still holds when replacing the classification loss $\mathcal{L}(f(x), y)$ with any cross-entropy based loss, since $\mathcal{L}(f(x), y)$ is bounded by the cross-entropy loss. Finally, we present our end-to-end algorithm for optimizing Eq. 10 in Algorithm 1, and its pictorial illustration in Fig. 2.

4. Experimental Results

4.1. Dataset and pre-processing

For evaluating our proposed method, we experiment with *seven* different popular deepfake benchmark datasets: NeuralTextures (NT) [50], Deepfakes (DF) [4], Face2Face (F2F) [51], FaceSwap (FS) [5], FaceShifter (FSH) [30], CelebDFV2 (CDFv2) [59], and Face Forensics in the Wild (FFIW10K) [62]. Besides the raw version, these videos are also compressed into two types: medium (c23) and high (c40), utilizing the H.264 codec and constant rate quantization parameters of 23 and 40, respectively. These effectively

Method	w/ prior infor.	#params	Test Set AUC (%)							
			NT	DF	F2F	FS	FSH	CDFv2	FFIW10K	*Avg*
BZNet [29]† [×3]	Y	22M × 3	91.01	99.30	96.90	98.82	-	-	-	96.51
ADD [28]† [×3]	Y	23.5M × 3	89.08	99.25	96.53	98.21	98.25	-	-	96.26
RESNET50 [×3]	Y	23.5M × 3	88.96	99.26	97.04	98.63	98.71	97.09	98.58	*96.90*
QAD-R (***ours***)	N	23.5M × 1	88.85	99.42	97.77	98.83	98.93	97.56	98.93	*97.18*
EFFICIENTNET-B1[×3]	Y	6.5M × 3	87.63	99.05	96.72	98.16	97.95	96.70	98.54	*96.39*
QAD-E (***ours***)	N	**6.5M × 1**	**92.25**	**99.46**	**98.30**	**99.08**	**98.90**	**97.50**	**99.01**	***97.79***

Table 3. **Classification performance in the quality-aware setting with video compression of test set**. Except for our model, each model is trained with three modalities: raw, c23, and c40, respectively (denoted [×3]). In the inference phase, while our QAD uses ***one single*** pre-trained model, other methods use their ***corresponding*** pre-trained model (*e.g.*, pre-trained RESNET-50 on raw) to detect a given testing input (*e.g.*, raw). Reported performances are averaged score of the three modalities.

result in different quality of deepfakes, and details of these datasets are provided in our Supp. Material.

4.2. Experimental Settings

The models are trained with the Adam optimizer [26] with a learning rate of 2e-3, scheduled by one-cycle strategy [46] in 32 epochs. We use a mini-batch size of 64. In every training epoch, the model is evaluated *ten* times, and we save the best one based on the validation accuracy. Regarding the backbone network, we use the RESNET-50 [17] (QAD-R) and EFFICIENTNET-B1 [49] (QAD-E) with their default input size of 224 × 224 and 240 × 240, respectively. The backbone models utilize pre-trained weights from IMAGENET dataset [7]. Our hyper-parameters settings $\{\sigma = 6, \alpha = 0.004, \gamma = 0.002\}$ are obtained by fine-tuning on RESNET50 with NeuralTextures dataset and are kept the same throughout all datasets, whereas that of EFFICIENTNET-B1 are $\{\sigma = 6, \alpha = 0.002, \gamma = 0.006\}$.

4.3. Results

This section reports the results of our QAD and other baselines under two scenarios: 1) *quality-agnostic* setting, which represents no model has prior knowledge of the input images' quality, and 2) *quality-aware* setting, which baselines are required to know inputs' quality information.

4.3.1 Quality-agnostic models

We use the popular deepfake detection benchmark methods on our datasets: 1) **MesoNet** with Inception layer by [1], 2) **Xception** model proposed by [43], 3) F^3**Net** [38], 4) **MAT** [61] - a multi-attention deepfake detector, 5) a deviation of the method proposed by [9] to the ***instance-based* collaborative learning**, SBIs [45] - a self-blended method using real image only during training, 7) **ADD** [28] - a knowledge distillation-based approach for detecting low-quality deepfakes, and 8) **BZNet** [29] - a super-resolution approach for improving detection of low-quality deepfakes. Each method has a different training approach to defend against performance degradation caused by lossy compression. The first five methods are trained with a mixture of the three data quality types (*raw+c23+c40*). SBIs is trained with the mid-level of video compression (*c23*), which is commonly adopted in many works. Meanwhile, ADD and BZNet models are trained on *raw*, *c23*, and *c40*, respectively; however, in the inference phase, they are blindly tested over the entire dataset without the prior knowledge of quality types, and we report their average performance. In the test set, we include both video compression and random JPEG image compression [2].

The results for the video compression are presented in Table 1, where our QAD outperforms other SOTA baselines across multiple benchmark datasets. Notably, we achieve a significant improvement in AUC score of up to 5.28% for heavily compressed datasets, such as NeuralTextures (89.64% vs 94.92%). We also surpassed previous works on various deepfake datasets by 0.44% to 1.05% points, with the exception of Deepfakes and FFIW10K datasets, which are easy to detect even when compressed. Compared to the collaborative learning baseline by [9], which is a comparative benchmark, our QAD still gains a decent improvement on average, up to 0.93% and 1.54% points with QAD-R and QAD-E, respectively. Finally, our QAD-E models achieved the highest score on average, reaching 98.47%.

Regarding the random image compression experiment, the results are provided in Table 2. Although BZNet is marginally outperform our model on face-reenactment deepfakes (NT and F2F), our method still achieves the best performance with the highest scores on *five* over *seven* datasets. On average, our enhancements show decent improvements, with margins of 0.5% and 0.88% of QAD-R and QAD-E, respectively, compared to the second-best competitor (Fang & Lin).

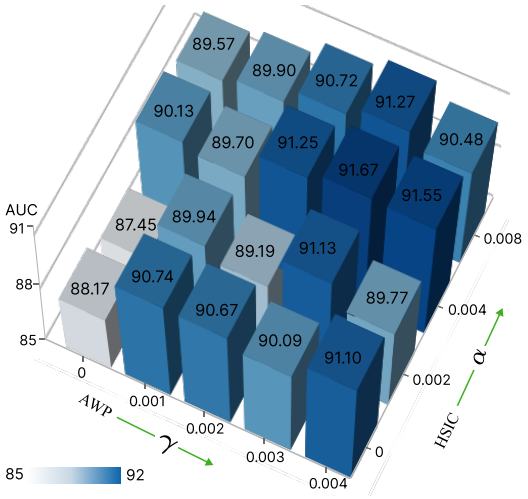

Figure 3. Model's performance versus α and γ on the NeuralTextures.

Model / loss		ResNet-50	
		ACC (%)	AUC (%)
Baseline		78.8	88.2
Coll. loss	Soft-label	77.0	84.0
	Pairwise loss	79.7	89.1
	Center loss	79.8	88.9
	HSIC	**80.3**	**90.1**
Adv. loss	AWP-KL	80.9	89.4
	AWP-XE	**81.7**	**90.7**
QAD *(ours)*		**82.2**	**91.3**

Table 4. Performance (ACC & AUC) of ResNet50 integrated with different losses.

4.3.2 Quality-aware models

In this experiment, we compare our models with quality-aware benchmark baselines. In particular, beside ResNet-50 and EfficientNet-B1, for each of the *raw*, *c23*, and *c40* datasets, we implement ADD [28] and BZNet [29] models. Since ADD and BZNet are the best performing methods, in which they utilized knowledge distillation and super-resolution approaches, respectively, for detecting deepfakes in different qualities. Hence, we only include them in this experiment. In the inference phase, the performance of these models is validated with the prior knowledge of the input image's quality, *i.e.*, *c40* images are evaluated by the same quality *c40* pre-trained models. Meanwhile, our universal QAD is **blindly** evaluated without such prior knowledge. We integrate our QAD on ResNet50 and EfficientNet-B1 and present their performance in Table 3. As we can obbserve, our QAD-E model performs slightly better or on par with ResNet-50, BZNet, and ADD models, despite having only **one-third of the number of parameters** and **no prior knowledge of input image quality**. Moreover, when integrating with EfficientNet-B1, QAD-E achieves a new SOTA performance with an improvement of up to 0.89% points (97.79% vs. 96.90%), with a modest number of parameters (6.5M).

QAD	NT	DF	F2F	FS	FSH	CDFv2	FFIW10K	*Avg*
ResNet18								
✗	88.73	98.93	98.05	98.06	98.67	97.09	98.72	*96.89*
✓	**91.38**	**99.32**	**98.32**	**99.19**	**98.94**	**97.97**	**99.10**	***97.75***
ResNet34								
✗	88.26	99.01	97.98	98.67	98.88	96.61	98.97	*96.91*
✓	**92.87**	**99.30**	**98.37**	**99.10**	**99.21**	**98.48**	**99.14**	***98.07***
EfficientNet-B0								
✗	86.12	99.29	97.93	98.21	98.42	97.81	98.80	*96.65*
✓	**91.99**	**99.33**	**98.66**	**99.15**	**99.00**	**98.38**	**99.12**	***97.95***

Table 5. Performance (AUC) of ResNet18, ResNet34, and EfficientNet-B0 baseline and their integration with our QAD training framework.

4.4. Ablation studies

4.4.1 α and γ of our loss

We investigate the sensitivities of our QAD with respect to α and γ, and summarize the results of our analysis in Fig. 3. In this study, we experiment with ResNet50 on the NeuralTextures dataset, which is the hardest dataset to detect when compressed. And, we vary the values of the hyperparameters $\alpha \in \{0.002, 0.004, 0.008\}$ and $\gamma \in \{0.001, 0.002, 0.003, 0.004\}$, where the value at $(\alpha, \gamma) = (0.0, 0.0)$ indicate the baseline. The results suggest that when α is greater than 0.002, the performance of our model is high and stable, surpassing current SOTA with any α greater than 0.002. Additionally, increasing γ generally improves performance. Note that, as we did not tune the hyper-parameters to optimize the test accuracy, Section 4.3's hyper-parameters are not the best, despite outperforming all current methods on the datasets.

4.4.2 Selection of losses

We study different alternatives for the *collaborative learning loss* and the *adversarial weight perturbation* approach. In particular, for the collaborative learning loss, we apply the loss function that was introduced by [47], in which they aggregate the logits of different views, combining them with their true labels to generate the soft labels. Besides, we replace our HSIC regularization with intermediate pairwise loss (Eq. 3) and center loss. Regarding the adversarial weight perturbation, we further apply the KL divergence between the representations of raw and compressed images to perturb the model's weights.

We report the results in Table 4, where we observe that the soft label loss fails to improve the baselines due to a

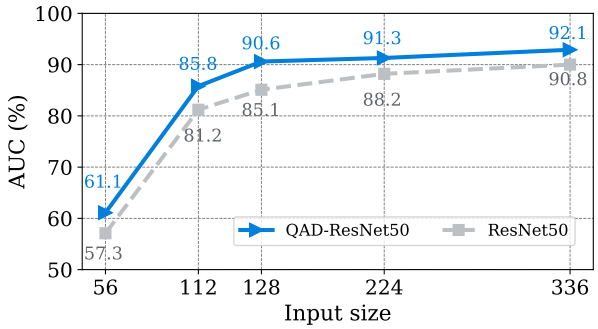

Figure 4. Performance (AUC) of our proposed method at different input resolutions with NeuralTextures dataset.

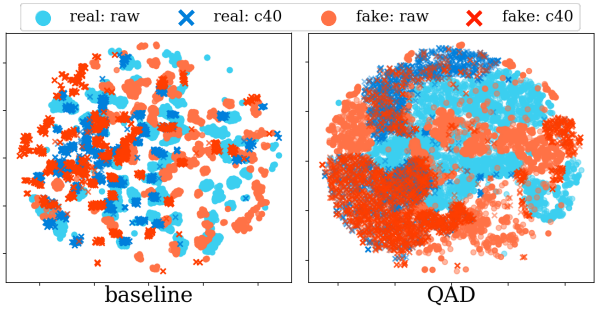

Figure 5. t-SNE visualisation of baseline and our QAD.

lack of low-level representation agreement between different image qualities. While both pairwise and center loss slightly improve the baselines and are unstable with different model architecture, our HSIC consistently achieves the best performance by relaxing the instance-base constraint. Meanwhile, we can observe the model's performance drops in terms of both ACC and AUC, when replacing our AWP with cross-entropy loss with KL divergence. Generally, this experiment shows that using pairwise differences of various quality image representations at the output, such as soft label, pairwise constraint, or AWP-KL, for optimizing the model can hinder its convergence to the optimal parameters.

4.4.3 Experiment with different backbones

Table 5 shows the comparability of our QAD with *three* different backbone networks: ResNet18, ResNet34, and EfficientNet-B0. The hyperparameter settings are kept the same for ResNet50 and EfficientNet-B1. As shown in Table 5, our QAD consistently improves the baselines, from 0.86% to 1.3% points on average of seven deepfake datasets.

4.4.4 Performance at different input scales

Unlike other classification tasks, a notable factor that substantially affects the deepfake detection performance, which is omitted by most previous works [43, 38, 61], is the input size of faces. We resize the input images from 56 to 336 and demonstrate how it impacts our QAD in comparison with ResNet50 baseline. The experiment is performed with the NeuralTextures dataset, and its results are reported in Fig. 4. We note that our proposed QAD and the baselines consistently improve their performance, when increasing the input size. Beside, our method also keep its staging improvement across the input resolutions compared to ResNet50 baseline.

4.4.5 Feature distribution visualization

To verify the consistency of invariant representation upon the input quality, we draw the feature distribution of EfficientNet-B1 and our QAD-EfficientNet-B1 pretrained on NeuralTextures (with raw, c23, and c40 datasets) with t-SNE [53]. The results are shown in Fig. 5. As observed, our QAD model's representations are less dispersed both in terms of intra-class and inter-quality. This experiment demonstrates that traditional cross-entropy loss trained with multiple input quality are confused due to the low-level constraints, while our QAD enables the model to achieve more generalization regardless of input quality.

5. Conclusion

Most deep learning-based deepfake detectors use a single model for each video quality, leaving an unsolved practical issue of their generalizability for detecting different quality of deepfakes. In this work, we propose a universal deepfake detection framework (QAD). Using intra-model collaborative learning, we minimize the geometrical differences of images in various qualities at different intermediate layers by the HSIC module. Moreover, our adversarial weight perturbation (AWP) module is directly applied to the model's parameters to provide its robustness against input image compression. Extensive experiments show that our QAD achieves competitive detection accuracy and marks the new SOTA results on various deepfake datasets without prior knowledge of input image quality.

Acknowledgements. This work was partly supported by Institute for Information & communication Technology Planning & evaluation (IITP) grants funded by the Korean government MSIT: (No. 2022-0-01199, Graduate School of Convergence Security at Sungkyunkwan University), (No. 2022-0-01045, Self-directed Multi-Modal Intelligence for solving unknown, open domain problems), (No. 2022-0-00688, AI Platform to Fully Adapt and Reflect Privacy-Policy Changes), (No. 2021-0-02068, Artificial Intelligence Innovation Hub), (No. 2019-0-00421, AI Graduate School Support Program at Sungkyunkwan University), and (No. RS-2023-00230337, Advanced and Proactive AI Platform Research and Development Against Malicious deepfakes).

References

[1] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Mesonet: a compact facial video forgery detection network. In *2018 IEEE international workshop on information forensics and security (WIFS)*, pages 1–7. IEEE, 2018. 6, 7

[2] Alexander Buslaev, Vladimir I Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A Kalinin. Albumentations: fast and flexible image augmentations. *Information*, 11(2):125, 2020. 7

[3] Dan Andrei Calian, Florian Stimberg, Olivia Wiles, Sylvestre-Alvise Rebuffi, András György, Timothy A Mann, and Sven Gowal. Defending against image corruptions through adversarial augmentations. In *International Conference on Learning Representations*, 2021. 2

[4] DeepFakes Community. Deepfakes. https://github.com/deepfakes/faceswap, 2017. Accessed: 2021-01-01. 6

[5] FaceSwap Community. Faceswap. https://github.com/MarekKowalski/FaceSwap, 2016. Accessed: 2021-01-01. 6

[6] Oscar de Lima, Sean Franklin, Shreshtha Basu, Blake Karwoski, and Annet George. Deepfake detection using spatiotemporal convolutional networks. *arXiv preprint arXiv:2006.14749*, 2020. 2

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 7

[8] Tarik Dzanic, Karan Shah, and Freddie D Witherden. Fourier spectrum discrepancies in deep network generated images. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pages 3022–3032, 2020. 1, 2

[9] Shijie Fang and Tong Lin. Intra-model collaborative learning of neural networks. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2021. 3, 6, 7

[10] FBI. Deepfakes and stolen pii utilized to apply for remote work positions. https://www.ic3.gov/Media/Y2022/PSA220628, June 2022. Accessed: 2022-07-01. 1

[11] Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *International Conference on Machine Learning*, pages 3247–3258. PMLR, 2020. 2

[12] Daniel Greenfeld and Uri Shalit. Robust learning with the hilbert-schmidt independence criterion. In *International Conference on Machine Learning*, pages 3759–3768. PMLR, 2020. 3

[13] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, pages 63–77. Springer, 2005. 3, 5

[14] Hui Guo, Shu Hu, Xin Wang, Ming-Ching Chang, and Siwei Lyu. Eyes tell all: Irregular pupil shapes reveal gan-generated faces. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2904–2908. IEEE, 2022. 2

[15] Qiushan Guo, Xinjiang Wang, Yichao Wu, Zhipeng Yu, Ding Liang, Xiaolin Hu, and Ping Luo. Online knowledge distillation via collaborative learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11020–11029, 2020. 3

[16] Alexandros Haliassos, Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Lips don't lie: A generalisable and robust approach to face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5039–5049, 2021. 2

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 7

[18] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *stat*, 1050:9, 2015. 3, 4

[19] Daniel Hsu, Ziwei Ji, Matus Telgarsky, and Lan Wang. Generalization bounds via distillation. *arXiv preprint arXiv:2104.05641*, 2021. 3

[20] Juan Hu, Xin Liao, Jinwen Liang, Wenbo Zhou, and Zheng Qin. Finfer: Frame inference-based deepfake detection for high-visual-quality videos. In *Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence*, pages 951–959, 2022. 1

[21] Takashi Ishida, Ikko Yamane, Tomoya Sakai, Gang Niu, and Masashi Sugiyama. Do we need zero training loss after achieving zero training error? *arXiv preprint arXiv:2002.08709*, 2020. 4

[22] Hyeonseong Jeon, Young Oh Bang, Junyaup Kim, and Simon Woo. T-gd: Transferable gan-generated images detection framework. In *International Conference on Machine Learning*, pages 4746–4761. PMLR, 2020. 2

[23] Yonghyun Jeong, Doyeon Kim, Youngmin Ro, and Jongwon Choi. Frepgan: Robust deepfake detection using frequency-level perturbations. In *Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence*, pages 1060–1068, 2022. 1

[24] Yuan Ji, Xu Jia, Huchuan Lu, and Xiang Ruan. Weakly-supervised temporal action localization via cross-stream collaborative learning. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, page 853–861, New York, NY, USA, 2021. Association for Computing Machinery. 3

[25] Mahyar Khayatkhoei and Ahmed Elgammal. Spatial frequency bias in convolutional generative adversarial networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 7152–7159, 2022. 1, 2

[26] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 7

[27] Binh M Le and SS Woo. Exploring the asynchronous of the frequency spectra of gan-generated facial images. In *CEUR Workshop Proceedings*, volume 3084. CEUR-WS, 2021. 2

[28] Binh M Le and Simon S Woo. Add: Frequency attention and multi-view based knowledge distillation to detect low-quality

compressed deepfake images. In *Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence*, pages 122–130, 2022. 1, 2, 3, 6, 7, 8

[29] Sangyup Lee, Jaeju An, and Simon S Woo. Bznet: Unsupervised multi-scale branch zooming network for detecting low-quality deepfake videos. In *Proceedings of the ACM Web Conference 2022*, pages 3500–3510, 2022. 1, 2, 6, 7, 8

[30] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. Faceshifter: Towards high fidelity and occlusion aware face swapping. *arXiv preprint arXiv:1912.13457*, 2019. 2, 6

[31] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face x-ray for more general face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5001–5010, 2020. 1, 2

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

[33] Yazhe Li, Roman Pogodin, Danica J Sutherland, and Arthur Gretton. Self-supervised learning with kernel dependence maximization. *Advances in Neural Information Processing Systems*, 34:15543–15556, 2021. 3

[34] Wan-Duo Kurt Ma, JP Lewis, and W Bastiaan Kleijn. The hsic bottleneck: Deep learning without back-propagation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5085–5092, 2020. 3

[35] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *International Conference on Learning Representations (ICLR)*, 2018. 5

[36] Awais Muhammad, Fengwei Zhou, Chuanlong Xie, Jiawei Li, Sung-Ho Bae, and Zhenguo Li. Mixacm: Mixup-based robustness transfer via distillation of activated channel maps. *Advances in Neural Information Processing Systems*, 34:4555–4569, 2021. 3

[37] Xiong Peng, Feng Liu, Jingfen Zhang, Long Lan, Junjie Ye, Tongliang Liu, and Bo Han. Bilateral dependency optimization: Defending against model-inversion attacks. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2022. 3

[38] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *European Conference on Computer Vision*, pages 86–103. Springer, 2020. 1, 2, 6, 7, 9

[39] Thorsten Quandt, Lena Frischlich, Svenja Boberg, and Tim Schatto-Eckrodt. Fake news. *The international encyclopedia of Journalism Studies*, pages 1–6, 2019. 1

[40] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20, 2007. 5

[41] Nicolas Rahmouni, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Distinguishing computer graphics from natural images using convolution neural networks. In *2017 IEEE Workshop on Information Forensics and Security (WIFS)*, pages 1–6. IEEE, 2017. 2

[42] Alfréd Rényi. On measures of dependence. *Acta mathematica hungarica*, 10(3-4):441–451, 1959. 3

[43] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1–11, 2019. 1, 2, 6, 7, 9

[44] Andrey Sebyakin, Vladimir Soloviev, and Anatoly Zolotaryuk. Spatio-temporal deepfake detection with deep neural networks. In *International Conference on Information*, pages 78–94. Springer, 2021. 1, 2

[45] Kaede Shiohara and Toshihiko Yamasaki. Detecting deepfakes with self-blended images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18720–18729, 2022. 6, 7

[46] Leslie N Smith and Nicholay Topin. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, volume 11006, page 1100612. International Society for Optics and Photonics, 2019. 7

[47] Guocong Song and Wei Chai. Collaborative learning for deep neural networks. *Advances in neural information processing systems*, 31, 2018. 2, 3, 6, 8

[48] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. 3

[49] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 7

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

[51] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2387–2395, 2016. 6

[52] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. In *International Conference on Learning Representations*, 2019. 4

[53] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 9

[54] Run Wang, Lei Ma, Felix Juefei-Xu, Xiaofei Xie, Jian Wang, and Yang Liu. Fakespotter: A simple baseline for spotting ai-synthesized fake faces. *arXiv preprint arXiv:1909.06122*, 2, 2019. 1, 2

[55] Colin Wei, Sham Kakade, and Tengyu Ma. The implicit and explicit regularization effects of dropout. In *International conference on machine learning*, pages 10181–10192. PMLR, 2020. 4

[56] Dongxian Wu, Shu-Tao Xia, and Yisen Wang. Adversarial weight perturbation helps robust generalization. *Advances in Neural Information Processing Systems*, 33:2958–2969, 2020. 2, 4

[57] Guile Wu and Shaogang Gong. Peer collaborative learning for online knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 10302–10310, 2021. 3

[58] Yongxin Yang and Timothy Hospedales. Deep multi-task representation learning: A tensor factorisation approach. In *5th International Conference on Learning Representations*, 2017. 3

[59] Li Yuezun, Yang Xin, Sun Pu, Qi Honggang, and Lyu Siwei. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *IEEE Conference on Computer Vision and Patten Recognition (CVPR)*, 2020. 6

[60] Xu Zhang, Svebor Karaman, and Shih-Fu Chang. Detecting and simulating artifacts in gan fake images. In *2019 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6. IEEE, 2019. 2

[61] Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu. Multi-attentional deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2185–2194, 2021. 1, 2, 6, 7, 9

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