ADD: Frequency Attention and Multi-View Based Knowledge Distillation to Detect Low-Quality Compressed Deepfake Images

Le Minh Binh,[1] **Simon Woo,** [2]

[1] Department of Computer Science and Engineering, Sungkyunkwan University, South Korea
[2] Department of Applied Data Science, Sungkyunkwan University, South Korea
bmle@g.skku.edu, swoo@g.skku.edu

Abstract

Despite significant advancements of deep learning-based forgery detectors for distinguishing manipulated deepfake images, most detection approaches suffer from moderate to significant performance degradation with low-quality compressed deepfake images. Because of the limited information in low-quality images, detecting low-quality deepfake remains an important challenge. In this work, we apply frequency domain learning and optimal transport theory in knowledge distillation (KD) to specifically improve the detection of low-quality compressed deepfake images. We explore transfer learning capability in KD to enable a student network to learn discriminative features from low-quality images effectively. In particular, we propose the Attention-based Deepfake detection Distiller (ADD), which consists of two novel distillations: 1) frequency attention distillation that effectively retrieves the removed high-frequency components in the student network, and 2) multi-view attention distillation that creates multiple attention vectors by slicing the teacher's and student's tensors under different views to transfer the teacher tensor's distribution to the student more efficiently. Our extensive experimental results demonstrate that our approach outperforms state-of-the-art baselines in detecting low-quality compressed deepfake images.

Introduction

Recently, facial manipulation techniques using deep learning methods such as deepfakes have drawn considerable attention (Rossler et al. 2019; Pidhorskyi, Adjeroh, and Doretto 2020; Richardson et al. 2020; Nitzan et al. 2020). Moreover, deepfakes have become more realistic and sophisticated, making it difficult to be distinguished by human eyes (Siarohin et al. 2020). And it has become much easier to generate such realistic deepfakes than before. Hence, such advancements and convenience enable even novices to easily create highly realistic fake faces for simple entertainment. However, these fake images raise serious security, privacy, and social concerns, as they can be abused for malicious purposes, such as impersonation (Catherine 2019), revenge pornography (Cole 2018) , and fake news propagation (Quandt et al. 2019).

To address such problems arising from deepfakes, there have been immense research efforts in developing effective deepfake detectors (Dzanic, Shah, and Witherden 2019; Rossler et al. 2019; Wang et al. 2019; Li and Lyu 2018; Khayatkhoei and Elgammal 2020; Zhang, Karaman, and Chang 2019). Most approaches utilize the deep learning-based approaches, where generally they perform well if there are a large amount of high-resolution training data. However, the performances of these approaches drop dramatically (by up to 18% (Dzanic, Shah, and Witherden 2019; Rossler et al. 2019)) for compressed low-resolution images due to lack of available pixel information to sufficiently distinguish fake images from real ones. In other words, because of the compression, subtle differences and artifacts such as sharp edges in hairs and lips that can be possibly leveraged for differentiating deepfakes can also be removed. Therefore, there still remains an important challenge to effectively detect low-quality compressed deepfakes, which frequently occur on social media and mobile platforms in bandwidth-challenging and storage-limited environments.

In this work, we propose the Attention-based Deepfake detection Distiller (ADD). Our primary goal is to detect low-quality (LQ) deepfakes, which are less explored in most previous studies but plays a pivotal role in real-world scenarios. First, we assume there are high-quality (HQ) images are readily available, similar to the settings in other studies (Rossler et al. 2019; Wang et al. 2019; Li and Lyu 2018; Khayatkhoei and Elgammal 2020; Zhang, Karaman, and Chang 2019; Dzanic, Shah, and Witherden 2019). And, we use knowledge distillation (KD) as an overarching backbone architecture to detect low-quality deepfakes. While most of the existing knowledge distillation methods aim to reduce the student size for model compression applications or improve the performance of lightweight deep learning models (Hinton, Vinyals, and Dean 2015; Tian, Krishnan, and Isola 2019; Huang and Wang 2017; Passalis and Tefas 2018), we hypothesize that a student can learn lost distinctive features of low-quality compressed images from a teacher that is pre-trained on high-quality images for deepfake detection. We first lay out the following two major challenges associated with detecting the LQ compressed deepfakes, and provide the intuitions of our approaches to overcome these issues:

1) Loss of high-frequency information. As discussed, while lossy image compression algorithms make changes visually unnoticeable to humans, they can significantly reduce

DNNs' deepfake detection capability by removing the fine-grained artifacts in high-frequency components. To investigate this phenomenon more concretely, we revisit Frequency Principal (F-Principal) (Xu, Zhang, and Xiao 2019), which describes the learning behavior of general DNNs in the frequency domain. F-Principal states that general DNNs tend to learn dominant low-frequency components first and then capture high-frequency components during the training process (Xu, Zhang, and Xiao 2019). For example, to illustrate this issue, Fig. 1 is provided to indicates that most of the lost information during compression is from high-frequency components. As a consequence, general DNNs shift their attention in later training epochs to high-frequency components, which now represent intrinsic characteristics of objects in each individual image rather than discriminative features. This learning process increases the variance of DNNs' decision boundaries and induces overfitting, thereby degrading the detection performance. A trivial approach to tackle the overfitting is applying the early stopping method; however, fine-grained artifacts of deepfakes can be subsequently omitted, especially when they are highly compressed. To overcome this issue, we propose the novel frequency attention distiller, which guides the student to effectively recover the removed high-frequency components in low-quality compressed images from the teacher during training.

2) Loss of correlated information. In addition, under heavy compression, crucial features and pixel correlations that not only capture the intra-class variations, but also characterize the inter-class differences are also degraded. In particular, these correlations are essential for CNNs' ability to learn the features at the local filters, but they are significantly removed in the compressed input images. Recent studies (Wang et al. 2018; Hu et al. 2018) have empirically demonstrated that training DNNs that are able to capture this correlated information can successfully improve their performances. Therefore, in this work, we focus on improving the lost correlations by proposing a novel multi-view attention, inspired by the work of Bonneel *et al.* (Bonneel et al. 2015), and contrastive distillation (Tian, Krishnan, and Isola 2019). The element-wise discrepancy between the teacher's and student's tensors that ignores the relationship within local regions of pixels, or channel-wise attention that only considers a single dimension of backbone features. On the other hand, our proposed method ensures that our model attends to output tensors from multiple views (slices) using Sliced Wasserstein distance (SWD) (Bonneel et al. 2015). Therefore, our multi-view attention distiller guides the student to mimic its teacher more efficiently through a geometrically meaningful metric based on SWD. In summary, we present our overall Attention-based Deepfake detection Distiller (ADD), which consists of two novel distillations (See Fig. 2): 1) frequency attention distillation that effectively retrieves the removed high-frequency components in the student network, and 2) multi-view attention distillation that creates multiple attention vectors by slicing the teacher's and student's tensors under different views to transfer the teacher tensor's distribution to the student more efficiently. Our contributions are summarized as follows:

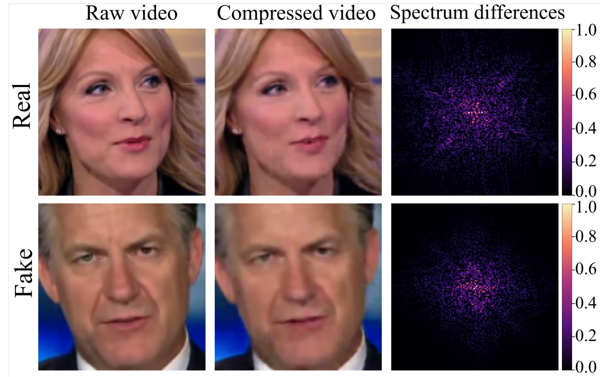

Figure 1: Degradation of high-frequency components after compression of real and fake videos. Sample fake face frames are taken from the NeuralTextures dataset in FaceForensics++ (Rossler et al. 2019). Left column: Sample faces from raw videos. Middle column: Sample faces from c40-compressed videos. Right column: Normalized spectrum differences in the frequency domain after applying Discrete Fourier Transform (DFT) to raw and compressed images. The concentrated differences at the center are the highest frequency components.

- We propose the novel *frequency attention distillation*, which effectively enables the student to retrieve high-frequency information from the teacher.
- We develop the novel *multi-view attention distillation* with contrastive distillation for the student to efficiently mimic the teacher while maintaining pixel correlations from the teacher to the student through SWD.
- We demonstrate that our approach outperforms well-known baselines, including attention-based distillation methods, on different low-quality compressed deepfake datasets.

Related Work

Deepfake Detection. Deepfake detection has recently drawn significant attention, as it is related to protecting personal privacy. Therefore, there has been a large number of research works to identify such deepfakes (Rossler et al. 2019; Li et al. 2020, 2019; Jeon et al. 2020; Rahmouni et al. 2017; Wang et al. 2019; Li and Lyu 2018). Li *et al.* (Li et al. 2020) tried to expose the blending boundaries of generated faces and showed the effectiveness of their method, when applied for unseen face manipulation techniques. Self-training with L2-starting point regularization was introduced by Jeon *et al.* (Jeon et al. 2020) to detect newly generated images. However, the majority of prior works are limited to high-quality (HQ) synthetic images, which are rather straightforward to detect by constructing binary classifiers with a large amount of HQ images.

Knowledge Distillation (KD). Firstly introduced by Hinton *et al.* (Hinton, Vinyals, and Dean 2015), KD is a training technique that transfers acquired knowledge from a pre-trained teacher model to a student model for model com-

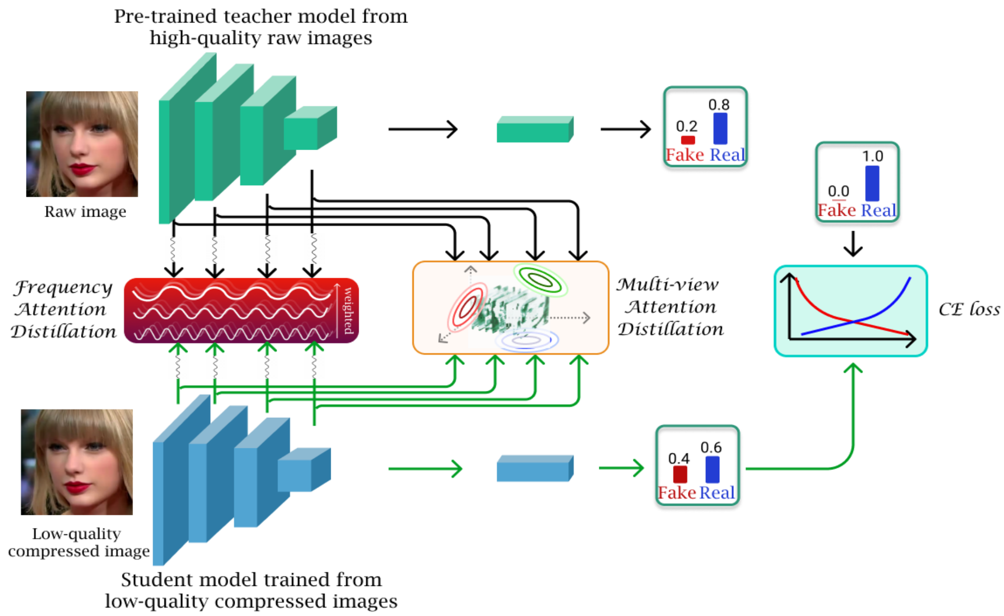

Figure 2: Illustration of our proposed Attention-based Deepfake detection Distiller (ADD) distillation framework. First, a low-quality compressed image and its corresponding raw image are used as an input to the student and pre-trained teacher model, respectively. The student model is trained with two novel distillers: 1) frequency attention distiller and 2) multi-view attention distiller. The frequency attention distiller creates a weighted loss, which focuses more on the degraded high-frequency components. The multi-view attention distiller slices student's and teacher's tensors by different random views to spawn multiple attention vectors. Green arrows indicate the flows of gradient decent updates to train the student's parameters.

pression applications. However, many existing works (Yim et al. 2017; Tian, Krishnan, and Isola 2019; Huang and Wang 2017; Passalis and Tefas 2018) applied different types of distillation methods to conventional datasets, *e.g.*, ImageNet, PASCAL VOC 2007, and CIFAR100, but not for deepfake datasets. On the other hand, Zhu *et al.* (Zhu et al. 2019) used FitNets (Romero et al. 2014) to train a student model that is able to detect low-resolution images, which is similar to our method in that the teacher and the student learn to detect high and low-quality images, respectively. However, their approach coerces the student to mimic the penultimate layer's distribution from the teacher, while it does not possess rich features at the lower layers.

In order to encourage the student model to mimic the teacher more effectively, Zagoruyko and Komodakis (Zagoruyko and Komodakis 2016) proposed the activation-based attention transfer, similar to FitNets, but their approach achieves better performance by creating spatial attention maps. Our multi-view attention method inherits from this approach but carries more generalization ability by not only exploiting spatial attention (in width and height dimension), but also introducing attention features from random dimensions using Radon transform (Helgason 2010). Thus, our approach pushes the student's backbone features closer to the teacher's.

In addition, inspired by InfoNCE loss (Oord, Li, and Vinyals 2018), Tian *et al.* (Tian, Krishnan, and Isola 2019) proposed contrastive representation distillation (CRD), which formulates the contrastive learning framework and motivates the student network to drive samples from positive pairs closer, and push away those from negative pairs. Although CRD achieves superior performance to those of previous approaches, it requires a large memory buffer to save embedding features of each sample. This is restrictive when training size and embedding space become larger. Instead, we directly sample positive and negative images in the same mini-batch and apply the contrastive loss to embedded features, similar to the Siamese network (Bromley et al. 1994).

Frequency Domain Learning. In the field of media forensics, several approaches (Jiang et al. 2020; Khayatkhoei and Elgammal 2020; Dzanic, Shah, and Witherden 2019) showed that discrepancies of high-frequency's Fourier spectrum are effective clues to distinguish CNN-based generated images. Frank *et al.* (Frank et al. 2020) and and Zhang *et al.* (Zhang, Karaman, and Chang 2019) utilized the checkerboard artifacts (Odena, Dumoulin, and Olah 2016) of the frequency spectrum caused by up-sampling components of generative neural networks (GAN) as effective features in detecting GAN-based fake images. Nevertheless, their detection performances were greatly degraded when the training synthesized images are compressed, becoming low-quality. Quian *et al.* proposed an effective frequency-based forgery detection method, named F^3Net, which decomposes an input image to many frequency components, collaborating with local frequency statistics on a two-streams network. The F^3Net, however, doubles the number of parameters from its backbone.

Wasserstein Distance. Induced by the optimal transport theory, Wasserstein distance (WD) (Villani 2008), and its variations have been explored in training DNNs to learn a particular distribution thanks to Wasserstein's underlying geometrically meaningful distance property. In fact, WD-based applications cover a wide range of fields, such as to improve generative models (Arjovsky, Chintala, and Bottou 2017; Deshpande, Zhang, and Schwing 2018), learn the distribution of latent space in autoencoders (Kolouri et al. 2018; Xu et al. 2020), and match features in domain adaptation tasks (Lee et al. 2019).In this work, we utilize the Wasserstein metric to provide the student geometrically meaningful guidance to efficiently mimic the teacher's tensor distribution. Thus, the student can learn the true tensor distribution, even though its input features are partially degraded through high compression.

Our Approach

Our Attention-based Deepfake detection Distiller (ADD) is consisted of the following two novel distillations (See Fig. 2): 1) frequency attention distillation and 2) multi-view attention distillation.

Frequency Attention Distillation

Let $f_{\mathcal{S}}$ and $f_{\mathcal{T}}$ be the student and the pre-trained teacher network. By forwarding a low-quality compressed input image and its corresponding raw image through $f_{\mathcal{S}}$ and $f_{\mathcal{T}}$, respectively, we obtain features $\mathcal{A}_{\mathcal{S}}$ and $\mathcal{A}_{\mathcal{T}} \in \mathbb{R}^{C\times W\times H}$ from its backbone network, which have C channels, the width of W, and the height of H. To create frequency representations, Discrete Fourier Transform (DFT) $\mathfrak{F} : \mathbb{R}^{C\times W\times H} \rightarrow \mathbb{C}^{C\times W\times H}$ is applied to each channel as follows:

$$\mathfrak{F}_{\mathcal{A}_{\mathcal{S}/\mathcal{T}}}(c,u,v) = \sum_{x=1}^{W}\sum_{y=1}^{H} \mathcal{A}_{\mathcal{S}/\mathcal{T}}(c,x,y)\cdot e^{-i2\pi(\frac{ux}{W}+\frac{vy}{H})}, \tag{1}$$

where c, x and y denote the c_{th}, x_{th} and y_{th} slice in the channel, the width and height dimension of $\mathcal{A}_{\mathcal{S}}$ and $\mathcal{A}_{\mathcal{T}}$, respectively. Here, for convenience, we use the notation $\mathfrak{F}_{\mathcal{A}_{\mathcal{S}/\mathcal{T}}}$ to denote that the function is independently applied for both student's and teacher's backbone features. Then, the value at (u, v) on each single feature-map $\mathfrak{F}_{\mathcal{A}_{\mathcal{S}/\mathcal{T}}}(c, :, :)$ indicates the coefficient of a basic frequency component. The difference between a pair of corresponding coefficients from the teacher and the student represents the "absence" of that student's frequency component. Next, let $d : \mathbb{C}^2 \rightarrow \mathbb{R}^+$ be a metric that assesses the distance between two input complex numbers and supports stochastic gradient descent. Then, the frequency loss between the teacher and student can be defined as follows:

$$\mathcal{L}_{\mathcal{FR}} = \sum_{c=1}^{C}\sum_{u=1}^{W}\sum_{v=1}^{H} w(u,v)\cdot d\big(\mathfrak{F}_{\mathcal{A}_{\mathcal{S}}}(c,u,v), \mathfrak{F}_{\mathcal{A}_{\mathcal{T}}}(c,u,v)\big), \tag{2}$$

where $w(u, v)$ is an attention weight at (u, v). In this work, we utilize the exponential of the difference across channels between the teacher and student as the weight in the following way:

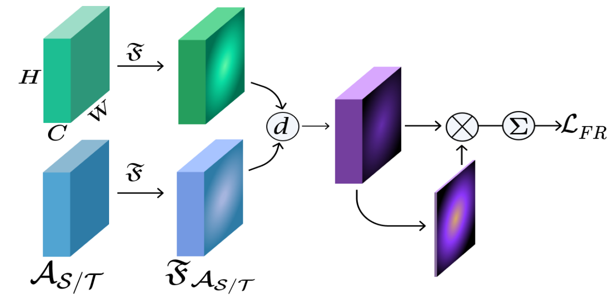

Figure 3: Illustration of our frequency attention distiller. The $\mathfrak{F}$ function is applied to each channel of the input tensor. Distance metric d calculates the discrepancy of the corresponding coefficients of each frequency component from the teacher and the student. Finally, the attention map is obtained by averaging the element-wise differences across channels.

$$w(u,v) = exp\big(\gamma_{\mathcal{FR}}\cdot\frac{1}{C}\sum_{c=1}^{C} d(\mathfrak{F}_{\mathcal{A}_{\mathcal{S}}}(c,u,v), \mathfrak{F}_{\mathcal{A}_{\mathcal{T}}}(c,u,v))\big), \tag{3}$$

where $\gamma_{\mathcal{FR}}$ is a positive hyper-parameter that governs the exponential cumulative loss, as the student's removed frequency increases. This design of attention weights ensures that the model focuses more on the losing high-frequency, and makes Eq. 2 partly similar to *focal loss* (Lin et al. 2017). Figure 3 visually illustrates our frequency loss.

Multi-view Attention Distillation

Sliced Wasserstein Distance. The p-Wasserstein distance between two probability measures μ and ν (Villani 2008) with their corresponding probability density functions P_μ and P_ν in a probability space $(\Omega, \mathcal{P}(\Omega))$ and $\Omega \subset \mathbb{R}^d$, is defined as follows:

$$W_p(P_\mu, P_\nu) = \Big(\inf_{\pi\in\Pi(\mu,\nu)} \int_{\Omega\times\Omega} \psi(x,y)^p d\pi(x,y)\Big)^{1/p}, \tag{4}$$

where $\Pi(\mu, \nu)$ is a set of all transportation plans π, which has the marginal densities P_μ and P_ν, respectively, and $\psi : \Omega\times\Omega \rightarrow \mathbb{R}^+$ is a transportation cost function. Equation 4 searches for an optimal transportation plan between μ and ν, which is also known as Kantorovitch formulation (Kantorovitch 1958). In the case of one-dimensional probability space, *i.e.*, $\Omega \subset \mathbb{R}$, the closed-form solution of the p-Wasserstein distance is:

$$W_p(P_\mu, P_\nu) = \Big(\int_0^1 \psi\big(F_\mu^{-1}(\kappa), F_\nu^{-1}(\kappa)\big)^p d\kappa\Big)^{1/p}, \tag{5}$$

where F_μ and F_ν are the cumulative distribution functions of P_μ and P_ν, respectively.

A variation of Wasserstein distance, inspired by the above closed-form solution, is Sliced Wasserstein distance (SWD) that deploys multiple projections from a high dimensional distribution to various one-dimensional marginal distributions and calculates the optimal transportation cost for each projection. In order to construct these one-dimensional marginal distributions, we use the Radon transform (Helgason 2010), which is defined as follows:

$$\mathcal{R}_{P_\mu}(t,\theta)=\int_{\Omega}\mu(x)\delta(t-\langle\theta,x\rangle)dx, \forall\theta\in\mathcal{S}^{d-1}, \forall t\in\mathbb{R}, \tag{6}$$

where δ denotes the Diract delta function, $\langle\cdot,\cdot\rangle$ is the Euclidean inner-product, and $\mathcal{S}^{d-1}\subset\mathbb{R}^d$ is the d-dimensional unit sphere. Thus, we denote $\mathcal{R}_\theta\mu$ as a 1-D marginal distribution of μ under the projection on θ. The Sliced 1-Wasserstein distance is defined as follows:

$$SW_1(P_\mu,P_\nu)=\int_{\mathcal{S}^{d-1}}W_1(\mathcal{R}_\theta\mu,\mathcal{R}_\theta\nu)d\theta. \tag{7}$$

Now, we can calculate the Sliced Wasserstein distance by optimizing a series of 1-D transportation problems, which have the closed-form solution that can be computed in $\mathcal{O}(Nlog(N))$ (Rabin et al. 2011). In particular, by sorting $\mathcal{R}_\theta\mu$ and $\mathcal{R}_\theta\nu$ in ascending order using two permutation operators τ_1 and τ_2, respectively, the SWD can be approximated as follows:

$$SWD(P_\mu,P_\nu)\approx\sum_{k=1}^{K}\sum_{i=1}^{N}\psi(\mathcal{R}_{\theta_k}\mu_{\tau_1[i]},\mathcal{R}_{\theta_k}\nu_{\tau_2[i]}), \tag{8}$$

where K is the number of uniform random samples θ using Monte Carlo method to approximate the integration of θ over the unit sphere $\mathcal{S}^{d-1}$ in Eq. 7.

Multi-view Attention Distillation. Let $P_\mathcal{A}$ be the square of $\mathcal{A}$ after being normalized by the Frobenius norm, *i.e.*, $P_\mathcal{A}=\frac{\mathcal{A}^{\circ 2}}{\|\mathcal{A}\|_F^2}$, where $\circ$ denotes the Hadamard power (Bocci, Carlini, and Kileel 2016). Consequently, we are now able to consider $P_\mathcal{A}$ as a discrete probability density function over $\Omega=\mathbb{R}^{C\times W\times H}\subset\mathbb{R}^3$, where $P_\mathcal{A}(c,x,y)$ indicates the density value at the slice c_{th}, x_{th} and y_{th} of the channel, the width and height dimension, respectively. To avoid replicating the element-wise differences, we additionally need to bin the projected vectors into g groups before applying distillation. One important property of our multi-view attention is that different values of θ provide different attention views (slices) of $\mathcal{A_S}$ and $\mathcal{A_T}$. For instance, with $\theta=(1,0,0)$, we achieve the channel-wise attention that was introduced by Chen *et al.* (Chen et al. 2017). Or, we can produce an attention vector in the width and height dimension, when θ becomes close to $(0,1,0)$ and $(0,0,1)$, respectively. With this general property, a student can pay full attention to its teacher's tensor distribution instead of some pre-defined constant attention views.

Figure 4 pictorially illustrates our overall multi-view attention distillation, and we summarize our multi-view attention in Algorithm in the supplementary materials. In order to encourage the semantic similarity of samples' representation from the same class and discourage that of those from different classes, we further apply the contrastive loss for each instance, which inspired by the CRD distillation framework of Tian *et al.* (Tian, Krishnan, and Isola 2019) . Thus, our

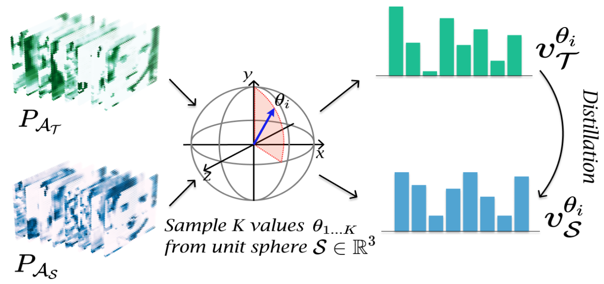

Figure 4: Detailed illustration of our multi-view attention distillation. Two backbone features of the teacher (top) and the student (bottom). After normalization, obtained features are projected on a random sample θ_i, then two attention vectors, $v_\mathcal{T}^{\theta_i}$ and $v_\mathcal{S}^{\theta_i}$, are obtained by sorting the projection images and binning them into g groups. Multiple values of θ provides us multiple attention views on the two tensors.

overall multi-view attention loss is defined as follows:

$$\begin{aligned}\mathcal{L}_{\mathcal{MV}}=&\gamma_{\mathcal{MV}}\times SWD(P_{\mathcal{A_S},P_{\mathcal{A_T}}})+\\&\eta_{\mathcal{MV}}\times\big[SWD(P_{\mathcal{A_S}},P_{\mathcal{A}_\mathcal{T}^+})+\\&\max(\Delta-SWD\big(P_{\mathcal{A_S}},P_{\mathcal{A}_\mathcal{T}^-}),0)\big],\end{aligned} \tag{9}$$

where $\mathcal{A}_\mathcal{T}^+$ and $\mathcal{A}_\mathcal{T}^-$ are the random instance's representation that belong to the same and the opposite class of $A_\mathcal{S}$ at the teacher, respectively. And Δ is a margin that manages the discrepancy of negative pairs and $\gamma_{\mathcal{MV}}$, and $\eta_{\mathcal{MV}}$ are scaling hyper-parameters.

Overall Loss Function

The overall distillation loss in our KD framework is formulated as follows:

$$\mathcal{L}_{Distill}(A_\mathcal{S},A_\mathcal{T})=\underbrace{\alpha\cdot\mathcal{L}_{\mathcal{FR}}}_{\text{frequency attention}}+\underbrace{\beta\cdot\mathcal{L}_{\mathcal{MV}}}_{\text{multi-view attention}}, \tag{10}$$

where α and β are hyper-parameters to balance the contribution of frequency attention distiller and multi-view attention distiller, respectively. Our attention loss is parameter-free and is independent from model architecture design, and it can be directly added to any detector model's conventional loss (*e.g.*, cross-entropy loss). Also, the frequency attention requires computational complexity in $\mathcal{O}\big(CWH\cdot(log(W)+log(H))\big)$ for one backbone feature, where $\mathcal{O}\big(WH\cdot(log(W)+log(H))\big)$ is the complexity of 2-D Fast Fourier Transform applied for one channel. On the other hand, the average-case complexity of multi-view attention is $\mathcal{O}\big(KN\cdot log(N)\big)$, where $\mathcal{O}\big(N\cdot log(N)\big)$ is the complexity of 1-D closed-form solution as mentioned above, K is the number of random samples θ, and N is the number of elements in one backbone feature, *i.e.*, $N=CWH$. Our end-to-end Attention-based Deepfake detection Distiller (ADD) pipeline is presented in Fig. 2.

Experiment

Datasets

Our proposed method is evaluated on five different popular deepfake benchmark datasets: NeuralTextures (Thies, Zollhöfer, and Nießner 2019) (NT), Deepfakes (Community 2017), Face2Face (Thies et al. 2016) (F2F), FaceSwap (Community 2016) (FS), and FaceShifter (Li et al. 2019) (FSr). Every dataset has 1,000 videos generated from 1,000 real human face videos by Rössler *et al.* (Rossler et al. 2019). These videos are compressed into two versions: medium compression (c23) and high compression (c40), using the H.264 codec with a constant rate quantization parameter of 23, and 40, respectively. Each dataset is randomly divided into training, validation, and test set consisting of 720, 140, and 140 videos, respectively. We randomly select 64 frames from each video and obtain 92,160, 17,920, and 17,920 images for training, validation, and test set, respectively. Then, we utilize the Dblib (King 2009) to detect the largest face in every single frame and resize them to a square image of 128×128 pixels.

Experiment Settings

In our experiments, we use Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 10^{-8}$. The learning rate is 2×10^{-4}, which follows one cycle learning rate schedule (Smith and Topin 2019) with a mini-batch size of 144. In every epoch, the model is validated 10 times to save the best parameters using validation accuracy. Early stopping is applied, when the validation performance does not improve after 10 consecutive times. We use ResNet50 (He et al. 2016) as our backbone to implement our proposed distillation framework. In Eq. 2, we define d as the square of the modulus of the difference between two complex number, *i.e.*, $d(c_1, c_2) = |c_1 - c_2|^2$, which satisfies the properties of a general distance metric: non-negative, symmetric, identity, and triangle inequality.

The number of binning groups g is equal to a half of the number of channels of $\mathcal{A}_{\mathcal{S}}$. Our hyper-parameters settings $\{\gamma_{\mathcal{FR}} = 1, \gamma_{\mathcal{MV}} = 100, \eta_{\mathcal{MV}} = 50, \Delta = 0.012, \alpha = 1,\}$ are kept the same, while β is fine-tuned on each dataset in the range of 16 to 23 through the experiments. The experiments are conducted on two TITAN RTX 24GB GPUs with Intel Xeon Gold 6230R CPU @ 2.10GHz.

Results

Our experimental results are presented in Table 1. We use Accuracy score (ACC) and Recall at 1 (R@1), which are described in detail in the supplementary materials . We compare our ADD method, with both distillation and non-distillation baselines. For a fair comparison between different methods, the same low resolution is used at 128×128 pixels as mentioned above is used throughout the experiments.

Non-distillation Methods. We reproduce two highest-score deepfake detection benchmark methods: 1) the method proposed by Rössler *et al.* (Rossler et al. 2019), which used Xception model, and 2) the approach by Dogonadze *et al.* (Dogonadze, Obernosterer, and Hou 2020), which employed Inception ResNet V1 pre-trained on the VGGFace2 dataset (Cao et al. 2018). These are the two best performing publicly available deepfake detection methods[1]. Additionally, we use the F^3Net, which is a frequency-based deepfake detection introduced by Quian *et al.* (Qian et al. 2020) for evaluation. The F^3Net is deployed on two streams of XceptionNet as described in the paper. Finally, ResNet50 (He et al. 2016) is also included as a baseline to compare with distillation methods.

Distillation Baseline Methods. As there has not been much research that deploys KD for deepfake detection, we further integrate other three well-known distillation architectures in the ResNet50 backbone to perform comparisons, including: FitNet (Romero et al. 2014), Attention Transfer (Zagoruyko and Komodakis 2016) (AT) and Non-local (Wang et al. 2018) (NL). Each of these methods is fine-tuned on the validation set to achieve its best performance.

First, comparing ours with the non-distillation baselines, we can observe that our method improves the detection accuracy from 1% to 6% across all five datasets for both compression data types. On average, our approach outperforms the other three distillation methods, and is superior on the highest compressed (c40) datasets. The model with FitNet loss, though it has a small improvement, does not have competitive results due to retaining insufficient frequency information. The attention module and non-local module also provide compelling results. However, they do not surpass our methods because of the lower attention dimension and frequency information shortage.

Ablation Study and Discussions

Effects of Attention Modules. We investigate the quantitative impact of the frequency attention and multi-view attention on the final performance. In the past, the NeuralTextures (NT) dataset has shown to be the most difficult to differentiate by both human eyes and DNN (Rossler et al. 2019). Hence, we conduct our ablation study on the c40 highly NT compressed dataset. The results are presented in Table 2. We can observe that frequency attention improves about 6.76% of the accuracy. Multi-view attention with contrastive loss provides a slightly better result than that of without contrastive at 68.14% and 67.01%, respectively. Finally, combining the frequency attention and multi-view attention distillation with contrastive loss significantly improves the accuracy up to 68.53%. The results of our ablation study demonstrate that each proposed attention distiller has a different contribution to the student's ability to mimic the teacher, and they are compatible when integrated together to achieve the best performance.

Sensitivity of Attention Weights (**α and β**). We conduct an experiment on the sensitivity of the frequency attention weight α and multi-view attention weight β on the five different datasets. The detailed results are presented in the supplementary materials. The result shows that by changing the value of α and β, the performance of our method continuously outperform the baseline results, indicating that our approach is less sensitive to both α and β.

[1]http://kaldir.vc.in.tum.de/faceforensics_benchmark/

Datasets	Models	Medium compression (c23)		High compression (c40)	
		ACC	R@1	ACC	R@1
NeuralTextures	Rössler *et al.*	76.36	57.24	56.75	51.88
	Dogonadze *et al.*	78.03	77.13	61.12	48.01
	F^3Net	77.91	77.39	61.95	32.35
	ResNet50	86.25	82.75	60.27	53.06
	FitNet - ResNet50	86.26	84.83	66.01	57.28
	AT - ResNet50	85.21	84.99	62.61	43.50
	NL - ResNet50	88.26	86.95	65.65	46.82
	ADD - ResNet50 (*ours*)	**88.48**	**87.53**	**68.53**	**58.42**
DeepFakes	Rössler *et al.*	97.42	96.96	92.43	82.39
	Dogonadze *et al.*	94.67	94.39	93.97	93.52
	F^3Net	96.26	95.84	93.06	93.00
	ResNet50	96.34	95.90	92.89	91.18
	FitNet - ResNet50	97.28	97.78	93.68	93.34
	AT - ResNet50	97.37	**98.72**	95.11	94.35
	NL - ResNet50	98.42	98.21	93.09	94.35
	ADD - ResNet50 (*ours*)	**98.67**	98.09	**95.50**	**94.59**
Face2Face	Rössler *et al.*	91.83	91.02	80.21	77.42
	Dogonadze *et al.*	89.34	88.73	83.44	81.00
	F^3Net	95.52	95.40	81.48	79.31
	ResNet50	95.60	94.77	83.94	79.88
	FitNet - ResNet50	95.91	96.16	83.48	78.99
	AT - ResNet50	96.80	96.84	83.55	78.72
	NL - ResNet50	96.44	96.64	83.69	82.04
	ADD - ResNet50 (*ours*)	**96.82**	**97.14**	**85.42**	**83.54**
FaceSwap	Rössler *et al.*	95.49	95.36	88.09	87.67
	Dogonadze *et al.*	93.33	92.78	90.02	89.10
	F^3Net	95.74	95.65	89.58	88.90
	ResNet50	92.46	90.85	88.91	86.52
	FitNet - ResNet50	97.29	96.29	89.16	90.13
	AT - ResNet50	97.66	97.27	89.75	90.41
	NL - ResNet50	97.34	96.95	91.86	90.78
	ADD - ResNet50 (*ours*)	**97.85**	**97.34**	**92.49**	**92.13**
FaceShifter	Rössler *et al.*	93.04	93.16	89.20	87.12
	Dogonadze *et al.*	89.80	89.36	82.03	79.96
	F^3Net	95.10	95.02	89.13	88.69
	ResNet50	94.89	93.88	89.56	88.48
	FitNet - ResNet50	**96.63**	95.95	90.16	89.36
	AT - ResNet50	96.32	**96.76**	88.28	89.45
	NL - ResNet50	96.24	95.28	90.04	87.71
	ADD - ResNet50 (*ours*)	96.60	95.84	**91.64**	**90.27**

Table 1: Experimental results of our proposed method and other seven different baseline approaches on five different deepfake datasets. The best results are highlighted in bold.

Experiment with Other Backbones. Table 3 shows the results with three other backbones: ResNet18 and ResNet34 (He et al. 2016), and EfficientNet-B0 (Tan and Le 2019). We set up the hyper-parameters of the four DNNs as the same for ResNet50, except $\gamma_{\mathcal{FR}}$ is changed to $1e^{-3}$ for EfficientNet-B0. Our distilled model improves the detection accuracy of all five datasets in different compression quality, up to 7%, 5.8%, and 7.1% with ResNet18, ResNet34, and EfficientNet-B0 backbone compared to their baselines, respectively.

Grad-CAM (Selvaraju et al. 2017). Using Grad-CAM, we provide visual explanations regarding the merits of training a LQ deepfake detector with our ADD framework. The gallery of Grad-CAM visualization is included in the supplementary material. First, our ADD is able to correct the facial artifacts' attention of the LQ detector to resemble its teacher trained on raw datasets. Second, the ADD vigorously instructs the student model to neglect the background noises and activate the facial areas as its teacher does when encountering facial images in complex backgrounds. Meanwhile, the baseline model which is solely trained on LQ datasets steadily makes wrong predictions with high confidence by activating non-facial areas and is deceived by complex backgrounds.

Model	ACC(%)
ResNet (baseline)	60.27
Our ResNet (FR)	67.03
Our ResNet (MV w/o contrastive)	67.01
Our ResNet (MV w/ contrastive)	68.14
Our ResNet (FR+MV)	**68.53**

Table 2: The effect of each single attention module on the final results experimented on NeuralTextures dataset.

		ResNet18		**ResNet34**		**EfficientNet-B0**	
	ADD	✗	✓	✗	✓	✗	✓
NT	*c23*	81.8	84.3	82.6	84.3	81.2	83.5
	c40	67.3	67.5	58.4	63.5	60.5	67.6
DF	*c23*	97.5	97.7	92.0	97.8	96.5	97.5
	c40	89.2	94.7	93.4	94.6	90.0	92.5
F2F	*c23*	94.2	95.7	94.2	94.9	94.1	96.7
	c40	85.0	85.3	83.2	83.4	77.4	80.3
FS	*c23*	90.2	96.0	92.4	96.8	92.6	95.3
	c40	84.5	91.5	88.6	90.6	83.4	87.5
FSr	*c23*	93.2	97.0	95.6	97.8	93.8	95.1
	c40	89.2	92.2	89.3	91.5	84.0	85.2

Table 3: Classification accuracy $(\%)$ of ResNet18, ResNet34 and EfficientNet-B0 baseline and their integration with our ADD training framework.

Conclusion

In this paper, we proposed a novel Attention-based Deepfake detection Distillations (ADD), exploring frequency attention distillation and multi-view attention distillation in a KD framework to detect highly compressed deepfakes. The frequency attention helps the student to retrieve and focus more on high-frequency components from the teacher. The multi-view attention, inspired by Sliced Wasserstein distance, pushes the student's output tensor distribution toward the teacher's, maintaining correlated pixel features between tensor elements from multiple views (slices). Our experiments demonstrate that our proposed method is highly effective and achieves competitive results in most cases when detecting extremely challenging highly compressed challenging LQ deepfakes.

Acknowledgments

This work was partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2019-0-00421, AI Graduate School Support Program (Sungkyunkwan University)), (No. 2019-0-01343, Regional strategic industry convergence security core talent training business) and the Basic Science Research Program through National Research Foundation of Korea (NRF) grant funded by Korea government MSIT (No. 2020R1C1C1006004). Also, this work was supported in part by the Institute of Information and Communications Technology Planning and Evaluation (IITP) Grant funded by the Korea Government (MSIT) (Artificial Intelligence Innovation Hub) under Grant 2021-0-02068. Lastly, this work was supported partly by the Institute of Information and Communications Technology Planning and Evaluation (IITP) grant funded by the Korea Government (MSIT) (No. 2022-0-00688, AI Platform to Fully Adapt and Reflect Privacy-Policy Changes).

References

Arjovsky, M.; Chintala, S.; and Bottou, L. 2017. Wasserstein generative adversarial networks. In *International conference on machine learning*, 214–223. PMLR.

Bocci, C.; Carlini, E.; and Kileel, J. 2016. Hadamard products of linear spaces. *Journal of Algebra*, 448: 595–617.

Bonneel, N.; Rabin, J.; Peyré, G.; and Pfister, H. 2015. Sliced and radon wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51(1): 22–45.

Bromley, J.; Guyon, I.; LeCun, Y.; Säckinger, E.; and Shah, R. 1994. Signature verification using a" siamese" time delay neural network. *Advances in neural information processing systems*, 737–737.

Cao, Q.; Shen, L.; Xie, W.; Parkhi, O. M.; and Zisserman, A. 2018. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, 67–74. IEEE.

Catherine, S. 2019. Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case. https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402. Accessed: 2021-01-01.

Chen, L.; Zhang, H.; Xiao, J.; Nie, L.; Shao, J.; Liu, W.; and Chua, T.-S. 2017. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5659–5667.

Cole, S. 2018. We Are Truly Fucked: Everyone Is Making AI-Generated Fake Porn Now. https://www.vice.com/en/article/bjye8a/reddit-fake-porn-app-daisy-ridley. Accessed: 2021-01-01.

Community, D. 2017. DeepFakes. https://github.com/deepfakes/faceswap. Accessed: 2021-01-01.

Community, F. 2016. FaceSwap. https://github.com/MarekKowalski/FaceSwap. Accessed: 2021-01-01.

Deshpande, I.; Zhang, Z.; and Schwing, A. G. 2018. Generative modeling using the sliced wasserstein distance. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3483–3491.

Dogonadze, N.; Obernosterer, J.; and Hou, J. 2020. Deep Face Forgery Detection. *arXiv preprint arXiv:2004.11804*.

Dzanic, T.; Shah, K.; and Witherden, F. 2019. Fourier Spectrum Discrepancies in Deep Network Generated Images. *arXiv preprint arXiv:1911.06465*.

Frank, J.; Eisenhofer, T.; Schönherr, L.; Fischer, A.; Kolossa, D.; and Holz, T. 2020. Leveraging frequency analysis for deep fake image recognition. In *International Conference on Machine Learning*, 3247–3258. PMLR.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Helgason, S. 2010. *Integral geometry and Radon transforms*. Springer Science & Business Media.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Hu, H.; Gu, J.; Zhang, Z.; Dai, J.; and Wei, Y. 2018. Relation networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3588–3597.

Huang, Z.; and Wang, N. 2017. Like what you like: Knowledge distill via neuron selectivity transfer. *arXiv preprint arXiv:1707.01219*.

Jeon, H.; Bang, Y.; Kim, J.; and Woo, S. S. 2020. T-GD: Transferable GAN-generated Images Detection Framework. *arXiv preprint arXiv:2008.04115*.

Jiang, L.; Dai, B.; Wu, W.; and Loy, C. C. 2020. Focal Frequency Loss for Generative Models. *arXiv preprint arXiv:2012.12821*.

Kantorovitch, L. 1958. On the translocation of masses. *Management science*, 5(1): 1–4.

Khayatkhoei, M.; and Elgammal, A. 2020. Spatial Frequency Bias in Convolutional Generative Adversarial Networks. *arXiv preprint arXiv:2010.01473*.

King, D. E. 2009. Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research*, 10: 1755–1758.

Kolouri, S.; Pope, P. E.; Martin, C. E.; and Rohde, G. K. 2018. Sliced Wasserstein auto-encoders. In *International Conference on Learning Representations*.

Lee, C.-Y.; Batra, T.; Baig, M. H.; and Ulbricht, D. 2019. Sliced wasserstein discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10285–10295.

Li, L.; Bao, J.; Yang, H.; Chen, D.; and Wen, F. 2019. Faceshifter: Towards high fidelity and occlusion aware face swapping. *arXiv preprint arXiv:1912.13457*.

Li, L.; Bao, J.; Zhang, T.; Yang, H.; Chen, D.; Wen, F.; and Guo, B. 2020. Face x-ray for more general face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5001–5010.

Li, Y.; and Lyu, S. 2018. Exposing deepfake videos by detecting face warping artifacts. *arXiv preprint arXiv:1811.00656*.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, 2980–2988.

Nitzan, Y.; Bermano, A.; Li, Y.; and Cohen-Or, D. 2020. Face identity disentanglement via latent space mapping. *ACM Transactions on Graphics (TOG)*, 39(6): 1–14.

Odena, A.; Dumoulin, V.; and Olah, C. 2016. Deconvolution and checkerboard artifacts. Distill (2016).

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Passalis, N.; and Tefas, A. 2018. Learning deep representations with probabilistic knowledge transfer. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 268–284.

Pidhorskyi, S.; Adjeroh, D. A.; and Doretto, G. 2020. Adversarial latent autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14104–14113.

Qian, Y.; Yin, G.; Sheng, L.; Chen, Z.; and Shao, J. 2020. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *European Conference on Computer Vision*, 86–103. Springer.

Quandt, T.; Frischlich, L.; Boberg, S.; and Schatto-Eckrodt, T. 2019. Fake news. *The international encyclopedia of Journalism Studies*, 1–6.

Rabin, J.; Peyré, G.; Delon, J.; and Bernot, M. 2011. Wasserstein barycenter and its application to texture mixing. In *International Conference on Scale Space and Variational Methods in Computer Vision*, 435–446. Springer.

Rahmouni, N.; Nozick, V.; Yamagishi, J.; and Echizen, I. 2017. Distinguishing computer graphics from natural images using convolution neural networks. In *2017 IEEE Workshop on Information Forensics and Security (WIFS)*, 1–6. IEEE.

Richardson, E.; Alaluf, Y.; Patashnik, O.; Nitzan, Y.; Azar, Y.; Shapiro, S.; and Cohen-Or, D. 2020. Encoding in style: a stylegan encoder for image-to-image translation. *arXiv preprint arXiv:2008.00951*.

Romero, A.; Ballas, N.; Kahou, S. E.; Chassang, A.; Gatta, C.; and Bengio, Y. 2014. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*.

Rossler, A.; Cozzolino, D.; Verdoliva, L.; Riess, C.; Thies, J.; and Nießner, M. 2019. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1–11.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, 618–626.

Siarohin, A.; Lathuilière, S.; Tulyakov, S.; Ricci, E.; and Sebe, N. 2020. First order motion model for image animation. *arXiv preprint arXiv:2003.00196*.

Smith, L. N.; and Topin, N. 2019. 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, 1100612. International Society for Optics and Photonics.

Tan, M.; and Le, Q. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, 6105–6114. PMLR.

Thies, J.; Zollhöfer, M.; and Nießner, M. 2019. Deferred neural rendering: Image synthesis using neural textures. *ACM Transactions on Graphics (TOG)*, 38(4): 1–12.

Thies, J.; Zollhofer, M.; Stamminger, M.; Theobalt, C.; and Nießner, M. 2016. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2387–2395.

Tian, Y.; Krishnan, D.; and Isola, P. 2019. Contrastive representation distillation. *arXiv preprint arXiv:1910.10699*.

Villani, C. 2008. *Optimal transport: old and new*, volume 338. Springer Science & Business Media.

Wang, R.; Ma, L.; Juefei-Xu, F.; Xie, X.; Wang, J.; and Liu, Y. 2019. Fakespotter: A simple baseline for spotting ai-synthesized fake faces. *arXiv preprint arXiv:1909.06122*, 2.

Wang, X.; Girshick, R.; Gupta, A.; and He, K. 2018. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7794–7803.

Xu, H.; Luo, D.; Henao, R.; Shah, S.; and Carin, L. 2020. Learning autoencoders with relational regularization. In *International Conference on Machine Learning*, 10576–10586. PMLR.

Xu, Z.-Q. J.; Zhang, Y.; and Xiao, Y. 2019. Training behavior of deep neural network in frequency domain. In *International Conference on Neural Information Processing*, 264–274. Springer.

Yim, J.; Joo, D.; Bae, J.; and Kim, J. 2017. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4133–4141.

Zagoruyko, S.; and Komodakis, N. 2016. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*.

Zhang, X.; Karaman, S.; and Chang, S.-F. 2019. Detecting and simulating artifacts in gan fake images. In *2019 IEEE International Workshop on Information Forensics and Security (WIFS)*, 1–6. IEEE.

Zhu, M.; Han, K.; Zhang, C.; Lin, J.; and Wang, Y. 2019. Low-resolution visual recognition via deep feature distillation. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 3762–3766. IEEE.