IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received January 29, 2022, accepted February 13, 2022, date of publication February 16, 2022, date of current version March 4, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3152029

Convolutional Neural Network Based on Diverse Gabor Filters for Deepfake Recognition

AHMED H. KHALIFA[1], NAWAL A. ZAHER[2], ABDALLAH S. ABDALLAH[3], (Member, IEEE), AND MOHAMED WALEED FAKHR[1]
[1]Department of Computer Engineering, Arab Academy for Science, Technology and Maritime Transport, Cairo 11799, Egypt
[2]Department of Electronics and Communications Engineering, Arab Academy for Science, Technology and Maritime Transport, Cairo 21913, Egypt
[3]School of Engineering, Penn State Erie-The Behrend College, Erie, PA 16563, USA

Corresponding author: Nawal A. Zaher (nawalzaher@aast.edu)

ABSTRACT Media synthesis and manipulation has reached unprecedented levels of realism owing to the proliferation of deep learning. Deepfake has been the de-facto tool for media manipulation. Although this technology has potential in the entertainment industry, its threats include political manipulation and bypassing biometric security systems. As a result, deepfake detection has garnered widespread attention among research communities. The intuition is to use deep learning to fix the problems created by deep learning. Although convolutional neural networks have shown their dominance in the filed of pattern recognition, the receptive field-model size dilemma still persists along with the lack of interpretation for such models. While the traditional Gabor function was proposed to fix these problems, it can only generate limited linear Gabor filters which makes it optimal for limited data and applications. The contribution of this paper is quadruple: (i) proposing a unified Gabor function capable of generating linear, elliptical, and circular Gabor filters. (ii) leveraging the back-propagation learning framework to incorporate the proposed function in convolutional neural networks and generate adaptive Gabor filters. (iii) presenting a dual scale large receptive field network for deepfake image recognition. (iv) demonstrating where the proposed model stands in terms of performance and architecture size compared to state-of-the-art models. The proposed model is evaluated on four benchmark datasets: Celeb-DF (v2), DeepFake Detection Challenge Preview, FaceForensics++ and Wilddeepfake. Experimental results show that the proposed adaptive Gabor filters reduce the model size by 64.9% compared to adaptive weighted filters without performance reduction.

INDEX TERMS Compact neural networks, image classification, image forensics, learnable filters, pattern recognition.

I. INTRODUCTION

The recent developments in deep generative models (DGMs), particularly variational autoencoders [1] and Generative Adversarial Networks [2], has enabled media synthesis and manipulation to reach unprecedented levels of realism. DGMs have impacted different fields including medical imaging [3], digital forensics [4] and art production [5]. However, the dark side of DGMs have been perceived with the emergence of ‘deepfake’ which is an infamous technology that employs DGMs to superimpose face images of a target person over that of a source person as shown in Fig. 1. Although public figures were the first targets of deepfake due to the abundant availability of their images online, it is currently possible for attackers to digitally impersonate any individual with acquiring a single image. The threats of deepfake include fake pornographic production, political manipulation, and bypassing biometric security systems. Since the risks of deepfake outweigh its benefits, deepfake detection models have become indispensable tools for distinguishing fabricated from authentic media.

While a deepfake image could fool human eyes, videos have noticeable key points to distinguish deepfake from genuine. In [7], the detection model is based on out-of-sync audio. Another identifier for videos is the inconsistencies between consecutive frames [8]. Furthermore, the average blinking time was found to be longer in deepfake videos than that of real videos [9], [10]. While sequence-based models based on the aforementioned elements achieved promising results, they suffer from two drawbacks. First, current and upcoming deepfake generators tend to improve these artifacts to come up with more robust models that could fool

The associate editor coordinating the review of this manuscript and approving it for publication was Zahid Akhtar.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

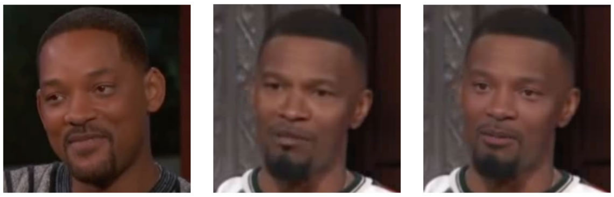


FIGURE 1. Deepfake frames from Celeb-DF (v2) dataset [6]. Frames from left to right belong to target, source and fake, respectively.

established detectors. Second, sequence-based detectors cannot be applied to deepfake images due to the lack of temporal information. Conversely, image-based detectors are applicable to manipulated videos through frame analysis and score fusion.

Although early image-based methods focused on salient artifacts for deepfake image detection, it was found that such methods tend not to generalize as well to samples spawned from unknown generators with latent artifacts compared to convolutional neural networks (CNNs) [11], [12]. CNNs have shown their dominance in the field of pattern recognition with Adaptive Weighted Filters (AWFs) being the fundamental component. However, the learnable weights for a single AWF of size $k \times k$ is k^2. As a result, there is a parabolic increase in the number of learnable parameters with the filter size for a constant number of filters. In order to alleviate this problem, Simonyan *et al.* proposed to use consecutive filters with small receptive fields (i.e., filter size) instead of a filter with large receptive field [13]. While this replacement was necessary to avoid the explosive increase in architecture size, it was not suitable for all recognition problems [14]. In addition to the increase in number of parameters with receptive fields, another problem is that AWFs lack interpretation.

Despite the particularity of each problem, they both stem from the foundation of CNNs that resides in AWFs of convolutional layers. Prior to deep learning, visual descriptors for image classification were extracted through hand-crafted methods characterized by their effectiveness and interpretability. Hence, incorporating these methods to deep learning will concurrently solve these problems. While traditional predefined filters such as Sobel, Schmid and Gabor were considered in CNNs, Gabor filters became a common choice lately [15]–[17]. The preference of Gabor filter comes from the following points. From a neurophysiological perspective, studies have revealed that the response of Gabor filters is equivalent to that of receptive fields of simple cells in the primary visual cortex [15]. From a signal processing perspective, Gabor filters are capable of extracting informative and discriminative joint spatial-spectral features. Furthermore, it was shown that low-level layers of CNNs tend to redundantly learn Gabor filters [17].

In this paper, we introduce parameter α to the Gabor function that controls the geometrical shape of Gabor stripes (i.e., axial ratio). Therefore, the Gabor function is capable of generating diverse filters including linear, circular and elliptical Gabor filters. Note that we consider both the real and imaginary components of the proposed function by considering the phase-induced form of the Gabor function [17]. Moreover, we adopt a back-propagation learning framework to enable the generation adaptive Gabor filters (AGFs). While AGFs produced from the proposed function could be applied in a variety of vision-related applications, we propose a compact architecture based on dual scale large receptive fields and self-attention for deepfake image recognition to demonstrate the effectiveness of AGFs. Three well-known deepfake datasets are used to evaluate the proposed architecture. In addition to comparing the proposed architecture to state-of-the-art image recognition models in terms of performance and model size, we further show that the utilization of AGFs instead of AWFs in the proposed architecture reduces the architecture size by 64.9%. The main contributions of this paper are outlined as follows:

- Different from previous work [16], [17] that leverage linear Gabor filters for pattern recognition, we develop a unified Gabor function capable of producing linear, circular and elliptical filters.
- In order to incorporate the proposed function in data-driven models, we utilize the back-propagation framework to enable the learnability of function parameters.
- We present a compact architecture for deepfake image recognition. The architecture leverages dual scale convolution with high receptive fields and self-attention mechanism.
- We evaluate the proposed architecture on three datasets and compare it to other state-of-the-art models in terms of performance and architecture size.

The remainder of this paper is organized as follows. Section II provides a brief review of the relevant related work in both Gabor-based CNNs and deepfake image recognition. Section III introduces the proposed Gabor function, provides the learning framework to generate AGFs, and presents the proposed architecture for deepfake image recognition. Section IV demonstrates the experimental results. Finally, Section V concludes the paper.

II. RELATED WORK

A. GABOR-BASED CNNs

Initially, predefined Gabor filters with fixed parameters were introduced in CNNs based on the observation that some weighted filters in AlexNet redundantly learn Gabor filters [19]. The objective was to modulate the learnable weighted filters aiming to enhance the deep feature representations with steerable orientation and scale capacities. This approach has proven to enhance the recognition performance with a perceptual reduction in the architecture size. Motivated by [19], Jiang *et al.* explored different architectures with varying depth for fast and efficient facial expression recognition [20]. In order to extract distinctive feature at different scales and orientations from limited training data, a combination of both fixed Gabor ensemble filter and AWFs were proposed for hyperspectral image classification [21]. In [22], a series of Gabor filters replaced the weighted filters in the first layer of CNN to enhance the overall classification performance. Furthermore, the spatial frequency and scale parameters of the filters are optimally obtained through coarse search in a small predefined subset of the parameters space and backward propagation is applied for fine tuning.

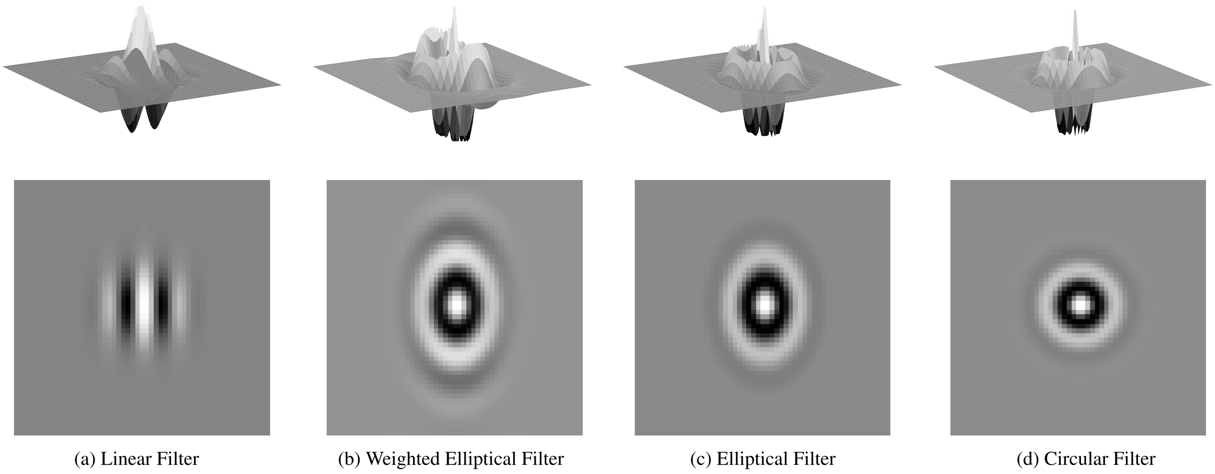
Meng *et al.* proposed a training procedure for Gabor filters in the first layer of CNNs based on the multipopulation genetic algorithm. The proposed procedure showed reduction in computational time and storage requirements [23]. In [24], Yuan et. al developed a regularizer loss function for a learnable Gabor convolution module. The module is used as a pre-processing tool and the features are passed to the ResNet-50 architecture for person re-identification. In [15], Zhang *et al.* addressed the difficulty of adjusting the parameters of Gabor filters via adaptive learning. In addition, the relation between the scale and frequency of the filter was leveraged to converge to the optimal values. The model demonstrated superior performance in finger-vein recognition. Stimulated by the fact that Gabor features assist in mitigating the negative effects introduced by the lack of training data, Liu *et al.* introduced naive Gabor networks [16]. In naive Gabor networks, CNNs strictly learn traditional Gabor filters to reduce the number of involved parameters and constrain the solution space. In addition, the offset phase parameters was not ignored in order to extract both local low-frequency and high-frequency features. The naive Gabor network was applied in hyperspectral image classification and showed superior performance with a small training set. Despite of the work done in Gabor-based CNNs, only linear Gabor filters (i.e., traditional) were incorporated in CNNs offering architectures with low diversity and adaptability to highly complex data.

B. DEEPFAKE IMAGE RECOGNITION

One of the first attempts to detect deepfake images involved exploiting the mesoscopic properties of images [25]. Two architectures were proposed to detect tampering: Meso-4 and MesoInception-4. Meso-4 consists of four layers of successive convolutions and pooling. These layers are followed by a dense network with one hidden layer. MesoInception-4, which is based on Meso-4, is formulated by replacing the first two convolutional layers by a variant of the inception module. Despite the lack of physical and mathematical interpretation for the proposed solutions, MesoNets and their variations have shown promising results in deepfake detection. In [26], capsule networks were proposed to detect forged images. Faces are detected and scaled to 128 × 128 in the pre-processing phase and a segment of VGG-19 extracts latent features. These features are distributed to three primary capsules and statistical pooling is utilised for forgery detection. Finally, outputs of the three capsules are dynamically routed to the output capsules. The two output capsules, one for real images and one for fake images, indicate the authenticity of images. Li *et al.* observed that the DeepFake generation pipeline produces face warping artifacts [12]. These artifacts result from resolution inconsistency. Motivated by pre-existing architectures such as VGG16, ResNet50, ResNet101 and ResNet152, these models were trained to detect face warping artifacts. In [27], Kim *et al.* proposed a combination of content and trace feature extractors to expose deepfake images. The content feature extractor utilizes ResNet-18 pre-trained model, while the trace feature extractor employs multi-channel constrained convolution. Furthermore, the features from both extractors are aggregated and connected to a fully connected layer to produce the classification result.

Owing to the fact that face manipulation methods share a common blending step, Li *et al.* proposed an image representation called face X-ray [28]. This representation was able to show the blending boundary for fake images without relying on specific facial artifacts, making it effective in image forgery detection. On the other hand, in [29], a patch-wise consistency learning approach was proposed for deepfake image detection. The module requires minor modification depending on the utilized backbone architectures. In [30], Feng *et al.* incorporated triplet loss function in the feature extraction stage of the deep learning model followed by a linear classification network to discriminate the learned contrastive features between real and fake face images. Due to the growth in number of face manipulation methods, octave convolution and an attention-based fusion module were proposed for mining intrinsic clues in channel difference image and spectrum image [31]. In addition, they designed an alignment module to enhance generalizability. In order to detect the convolutional traces left by GANs in fake images, Guarnera *et al.* developed an approach based on the Expectation-Maximization algorithm to detect the fingerprints of fake images [32].

Multiple methods started to leverage the frequency domain for deepfake recognition more recently. Durall *et al.* utilized discrete Fourier transform and applied azimuthal averaging to reduce the amount of features without losing relevant information [33]. Additionally, three different classifiers were used for classification comparison: SVM, logistic regression, and K-means clustering. In [34], it was shown that frequency representation can be used to easily identify severe artifacts. In addition, they demonstrated that transformed images via discrete cosine transform were linearly separable, while classification on raw pixel images required nonlinear models. Motivated by frequency-aware forgery clues, Qian *et al.* developed a Frequency-Aware Decomposition (FAD) for adaptive partition of input images according to a set of learnable frequency filters [35]. Moreover, Local Frequency Statistics (LFS) were extracted to describe the statistical discrepancy between real and fake samples. Both clues from FAD and LFS were learned by a cross-attention powered two-stream network. In [36], it was observed that cumulative up-sampling in face forgery techniques resulted in plain changes in the phase spectrum. Therefore, they designed a novel spatial-phase shallow learning approach that utilized the spatial image and phase spectrum to capture the up-sampling artifacts of face forgery. Wang *et al.* managed



(a) Linear Filter (b) Weighted Elliptical Filter (c) Elliptical Filter (d) Circular Filter

FIGURE 2. Visualization of some special cases of the proposed Gabor function at $\omega = \pi/4$, $\sigma = 5$, $\theta = 0$, $\phi = 0$. (a) $\alpha = 0$, $\gamma = 0.5$. (b) $\alpha = \gamma = 0.5$. (c) $\alpha = \gamma^2 = 0.5$. (d) $\alpha = \gamma = 1$. The top row indicates 3-D representation and the bottom row shows the corresponding 2-D filters.

to capture the subtle artifacts at different scales through the utilization of transformer models [37]. In addition to the multi-scale transformer that detects local inconsistency at different spatial levels, frequency information is leveraged to enhance the robustness of the model to image compression.

III. METHOD

A. DIVERSE GABOR FUNCTION

The conventional 2-D Gabor function is defined as a Gaussian function multiplied by a sinusoidal plane wave. Therefore, filters produced from the traditional function have the shape of parallel linear stripes, which represent the sinusoidal wave, encompassed by a Gaussian envelope. Hereafter, traditional Gabor filters are referred to as linear Gabor filters (LGFs). Other variants of Gabor filters such as Circular Gabor Filters (CGFs) and Elliptical Gabor Filters (EGFs) have been designed for hand-crafted texture segmentation and analysis of ring-like shapes [38], [39]. However, a single image could contain different shapes. For example, while the outline human face is linear, the outline of eyelid is elliptical and that of the iris is circular. In order to fully utilize all these shape for effective feature extraction, we develop a unified phase-induced diverse Gabor function $G_\Psi(x, y)$ by introducing the novel parameter α, as follows:

$$\mathfrak{G}_\Psi(x, y) = K \exp(jP) \tag{1}$$

$$K = \frac{\gamma}{2\pi\sigma^2} \exp\left(-\frac{x_r^2 + \gamma^2 y_r^2}{2\sigma^2}\right) \tag{2}$$

$$P = \omega\sqrt{x_r^2 + \alpha y_r^2} + \phi \tag{3}$$

$$x_r = x\cos\theta + y\sin\theta \quad y_r = y\cos\theta - x\sin\theta \tag{4}$$

where $\Psi = \{\gamma, \sigma, \theta, \omega, \alpha, \phi\}$ defines the set of parameters described in Table 1. K defines the Gaussian function, x_r and y_r refer to the axis transformation. The real and imaginary components of the complex function can be expressed by

$$\Re\{\mathfrak{G}_\Psi(x, y)\} = K \cos P \tag{5}$$

$$\Im\{\mathfrak{G}_\Psi(x, y)\} = K \sin P \tag{6}$$

Generally, the phase offset is ignored and only the real component is considered for practical applications. However, it has been proven that both real and imaginary components are needed for low and high frequency analysis [40]. Therefore, we consider the real cosine component with phase offset ϕ, since the imaginary sine component could be generated from the real cosine component at $\phi = -\pi/2$. Hereafter, we refer to the phase-induced real component by $G = K \cos P$.

TABLE 1. Diverse Gabor parameters.

Parameter	Description	Range
α	diversity ratio of the stripes	$[0, 1]$
γ	aspect ratio of the envelope	$(0, 1]$
θ	orientation of the normal to the stripes	$[0, \pi)$
σ	standard deviation of the envelope	$(0, 4\pi]$
ϕ	phase offset	$[0, 2\pi)$
ω	angular frequency	$(0, \pi)$

The incorporation of parameter α provides greater diversity, making LGF and CGF special cases of the proposed function when $\alpha = 0$ and $\alpha = 1$, respectively. Furthermore, weighted EGFs can be generated in addition to EGFs by changing the values of α and γ. The difference is that the former has varying amplitude while the latter has constant amplitude along the ellipse. However, we refer to both as EGFs for convenience and the same goes for CGFs. Fig. 2 illustrates special cases of the diverse Gabor function by varying the values of α and γ.

B. LEARNING FRAMEWORK

A distinct difference between AWFs and the proposed AGFs is the reduced hypothesis space. This can be inferred from the range of values for each parameter. Since the objective function used in the deepfake image recognition model is the differentiable binary cross-entropy loss, the parameter set Ψ can be optimized via back-propagation. The update of parameters can be expressed by

$$\psi = \psi - \eta \frac{\partial \mathcal{L}}{\partial \psi}, \qquad \psi \in \Psi \tag{7}$$

By applying the chain rule

$$\frac{\partial \mathcal{L}}{\partial \psi} = \frac{\partial \mathcal{L}}{\partial G} \times \frac{\partial G}{\partial \psi}, \qquad \psi \in \Psi \tag{8}$$

The gradient with respect to each parameter is defined as follows:

$$\frac{\partial G}{\partial \alpha} = -\frac{\omega y_r^2}{2\sqrt{x_r^2 + \alpha y_r^2}} K \sin P \tag{9}$$

$$\begin{aligned}\frac{\partial G}{\partial \gamma} &= \left[\frac{1}{2\pi\sigma^2} - \frac{\gamma^2 y_r^2}{2\pi\sigma^4}\right] \exp\left(-\frac{x_r^2 + \gamma^2 y_r^2}{2\sigma^2}\right) \cos P \\ &= \left[\frac{1}{\gamma} - \frac{\gamma y_r^2}{\sigma^2}\right] G\end{aligned} \tag{10}$$

$$\begin{aligned}\frac{\partial G}{\partial \sigma} &= \left[\left(\frac{\gamma}{2\pi\sigma^2}\right)\left(\frac{x_r^2 + \gamma^2 y_r^2}{\sigma^3}\right) - \frac{\gamma}{\pi\sigma^3}\right] \\ &\quad \times \exp\left(-\frac{x_r^2 + \gamma^2 y_r^2}{2\sigma^2}\right) \cos P \\ &= \left[\frac{x_r^2 + \gamma^2 y_r^2}{\sigma^3} - \frac{2}{\sigma}\right] G\end{aligned} \tag{11}$$

$$\frac{\partial G}{\partial \phi} = -K \sin P \tag{12}$$

$$\frac{\partial G}{\partial \omega} = -\sqrt{x_r^2 + \alpha y_r^2}\, K \sin P \tag{13}$$

In order to simplify the derivation of θ, the relation in Eq. (4) is leveraged as follows:

$$\frac{\partial x_r}{\partial \theta} = y \cos\theta - x \sin\theta = y_r \tag{14}$$

$$\frac{\partial y_r}{\partial \theta} = -(x \cos\theta + y \sin\theta) = -x_r \tag{15}$$

$$\frac{\partial G}{\partial \theta} = \left[\frac{\gamma^2 - 1}{\sigma^2} \cos P + \frac{\omega(\alpha - 1)}{\sqrt{x_r^2 + \alpha y_r^2}} \sin P\right] x_r y_r K \tag{16}$$

The advantage of incorporating the proposed function in convolutional layers is threefold. First, it will significantly reduce the number of parameters should the need for large receptive field (i.e. filter size) arise. Traditionally, the input to a convolutional layer of N weighted filters of each $k \times k$ size is assumed to consist of C channels. As a result, the number of parameters for a convolutional layer based on AWFs is $k^2N(C + 1)$. For a constant N and C, there is a parabolic increase in the number of parameters with respect to k. However, in case of a convolutional layer incorporated with the diverse Gabor function, the cardinality of its parameter set $|\Psi|$ is 6. As a result, the number of learnable parameters will be $6N(C + 1)$ for a convolutional layer based on AGFs. In this case, the number of parameters is independent on the filter size which allows the utilization of large receptive field filters without increase in the number of learnable parameters. However, it is noteworthy that AGFs will be very inefficient for convolutional models used in attention mechanisms that utilize 1×1 filters [41]. Second, AGFs have higher interpretability compared to AWFs since they are generated from a mathematical function with a predefined set of parameters ψ. Third, while the proposed function can be used to construct deep architectures that solely consist of convolutional layers based on AGFS, it can also be simultaneously used with other convolutional layers that are based on AWFs within a single architecture as in the proposed dual scale large receptive field network (DSLRFN) for deepfake image recognition.

C. DEEPFAKE RECOGNITION

Although the proposed function and learning framework could be applied to any visual application, we focus on the pervasive problem of deepfake image recognition. The overall model is shown in Fig. 3. Since deepfake manipulates the facial region, face detection is an essential step in deepfake recognition. A multi-task cascaded convolutional network (MTCNN) is adopted since it provides satisfactory performance-runtime trade-off [42]. The extracted facial region is then passed to DSLRFN for classification. The proposed DSLRFN consists of four main blocks. The face image goes through a dual scale convolutional block that utilizes the proposed AGFs with large receptive fields. If small receptive fields were to replace the large receptive fields, 5 and 7 convolution layers of 3 × 3 filters would replace the 11 × 11 and 15 × 15 layers, respectively. Therefore, AGFs do not only reduced the architecture size but also reduces its depth. The objective of dual scale block is the extraction of feature at different scale. Furthermore, features are aggregated using element-wise maximum instead of a conventional concatenation layer for feature space reduction. Note that padding is used in the 15 × 15 convolutional layer to produce feature maps of the same size. The second block represents a self-attention mechanism. The block takes feature maps produced from a max pooling layer, produces attention maps, and multiplies attention maps to the feature maps in order to refine the most germane segments of the feature maps. The third block is a high-level embedding block. It takes feature maps from the second pooling layer in order to produce highly abstract features used for classification. The final block is the classifier block which takes high-level features and uses a single fully connected layer to produce the classification result. A batch normalization layer is used after every convolutional layer for training stabilization by reducing the covariant shift [43]. Note that AGFs are only used in dual scale convolution blocks. Furthermore, the number of filters in each convolutional layer is 32, resulting in compact network that consists of 17,013 parameters.

IV. EXPERIMENTS

A. DATASETS

The experiments were conducted on four benchmark datasets: Celeb-DF (v2) (CD2) [6], DeepFake Detection Challenge Preview (DFDC) [44], FaceForensics++ (FF++) [45]

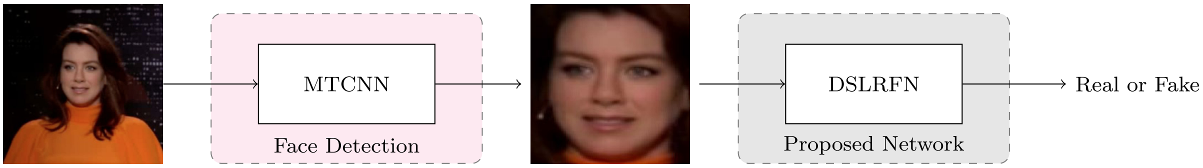


FIGURE 3. Overall model.

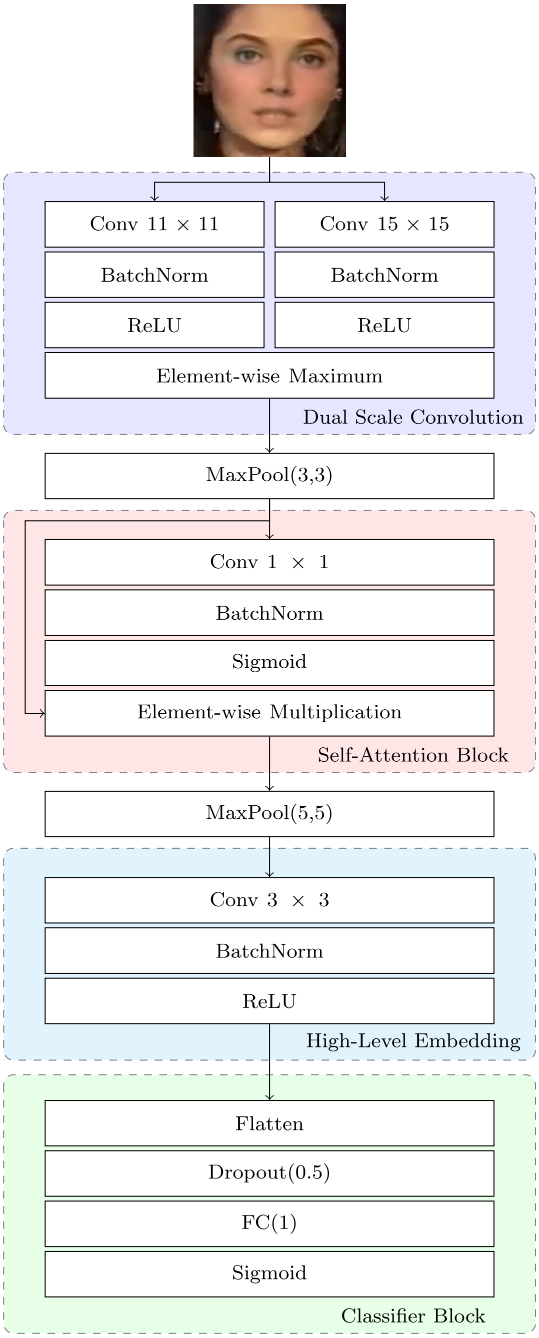


FIGURE 4. Proposed architecture.

and WildDeepfake (WDF) [46]. CD2 contains 5639 fake videos generated from 890 genuine videos collected from the internet. The standard testing split consists of 520 videos (179 real and 341 fake). The fake videos in CD2 were produced by a single generator. DFDC consists of 1131 real videos acquired from paid actors and 4105 fake videos produced by two unknown generators. The standard testing split of DFDC is 775 videos (276 real and 499 fake). FF++ is a superset that consists of 1000 real videos that were manipulated to produce 4000 tampered videos by four different methods: Deepfake, FaceSwap, Face2Face, and NeuralTextures. FF++ has a standard train:validation:test split of 720:140:140. Finally, WDF is one of the most recent datasets that consists of 707 videos. Note that all the videos, whether real or fake, were found in the wild. Hence, the number of manipulation methods in this database is unknown. Therefore, WDF is only used to test the generalizability of the models against samples found in the wild.

B. IMPLEMENTATION DETAILS

Conventionally, image-based deepfake detectors exploit a number of still frames from a given video in order to avoid redundancy and high computational complexity, especially since videos in the considered datasets are short and take place in stationary environments. Existing models select key frames, first few frames, or sample one frame per second of the video. This work selects equally separated frames from each video for the sake of variety following [31]. Additionally, MTCNN is used to detect 224 × 224 facial regions in the selected frames. Furthermore, DSLRFN is implemented in PyTorch using the binary cross-entropy loss and Adam optimizer with an initial learning rate of 0.001 for 120 epochs and batch size of 128 and the learning rate is reduced on plateau by a factor of 0.1 with a patience of 8.

C. EVALUATION

Three widely used metrics for deepfake detection are considered for evaluation: accuracy (ACC), Area Under receiver operating characteristic Curve (AUC) and Equal Error Rate (EER). Note that image-level evaluation is performed since the model operates on images [46]. For comparison, we consider the following state-of-the-art architectures ResNet-18 [47], DenseNet-121 [48], MobileNetV2 [49], EfficientNetB0 [50], and MesoNets [25]. In addition, we consider the replacement of AGFs by conventional AWFs in the proposed DSLRFN. Note that the pre-trained ImageNet weights were not used and all these models were trained on the same data as the proposed network for fair evaluation. Tables 2, 3, 4 show the performance on FF++, CD2 and DFDC, respectively. Furthermore, we evaluate model generalizability on WDF. Owing to the fact that DFDC has shown

TABLE 2. Performance on FF++ dataset. Bold and underlined scores represent the best and second best scores, respectively.

Method	DeepFake			FaceSwap			Face2Face			NeuralTexture		
	ACC	AUC	EER	ACC	AUC	EER	ACC	AUC	EER	ACC	AUC	EER
ResNet-18	91.56	97.94	7.14	91.12	97.34	8.75	92.46	97.59	7.32	79.02	86.01	20.62
DenseNet-121	90.94	97.45	9.20	**92.46**	98.02	**7.05**	92.95	98.05	6.16	78.79	86.71	21.52
MobileNet-V2	91.12	97.28	8.75	92.10	97.75	7.95	91.79	97.57	7.77	77.81	85.44	21.96
EfficientNetB0	90.62	96.86	9.38	92.28	**98.08**	**7.05**	92.63	97.64	7.68	73.44	82.24	25.36
Meso-4	91.92	97.67	8.12	90.89	96.63	9.38	92.50	97.71	7.05	80.62	**89.27**	**18.93**
MesoInception-4	92.54	**98.29**	7.41	91.21	97.32	8.04	91.88	97.24	8.12	79.46	88.01	20.09
DSLRFN (AWFs)	93.30	98.22	6.43	91.96	96.94	8.12	**93.75**	**98.32**	6.07	77.10	85.01	22.59
DSLRFN (AGFs)	**93.62**	98.12	**6.16**	92.28	97.54	7.59	93.44	98.08	**5.89**	**80.71**	88.60	19.38

TABLE 3. Performance on Celeb-DF (v2) dataset. Bold and underlined scores represent the best and second best scores, respectively.

Method	ACC ↑	AUC ↑	EER ↓
ResNet-18	**87.03**	93.91	13.82
DenseNet-121	86.14	**94.62**	**12.79**
MobileNet-V2	83.75	90.93	17.44
EfficientNetB0	84.61	91.59	16.12
Meso-4	86.29	94.03	13.93
MesoInception-4	86.20	94.04	13.37
DSLRFN (AWFs)	86.10	93.60	13.71
DSLRFN (AGFs)	86.43	94.16	13.15

TABLE 4. Performance on DeepFake Detection Challenge Preview dataset. Bold and underlined scores represent the best and second best scores, respectively.

Method	ACC ↑	AUC ↑	EER ↓
ResNet-18	73.42	81.91	25.15
DenseNet-121	71.85	81.84	26.15
MobileNet-V2	70.49	80.65	26.21
EfficientNetB0	**74.70**	**85.24**	**23.26**
Meso-4	67.51	82.20	26.67
MesoInception-4	67.57	81.76	25.58
DSLRFN (AWFs)	72.14	83.83	24.11
DSLRFN (AGFs)	72.74	84.48	23.51

TABLE 5. Ablation study of the dual-scale block and the self-attention mechanism.

Method	CD2	DFDC	FF++
DSLRFN	94.16	84.48	95.59
DSLRFN w/o Dual Scale	91.67	82.82	95.37
DSLRFN w/o Attention	93.58	82.81	95.13

TABLE 6. Generalization on Wild DeepFake dataset.

Method	ACC ↑	AUC ↑	EER ↓
ResNet-18	64.16	72.40	33.91
DenseNet-121	64.44	71.49	34.69
MobileNet-V2	68.21	72.32	31.83
EfficientNetB0	64.06	70.45	34.77
Meso-4	67.53	73.71	32.57
MesoInception-4	67.19	73.65	34.23
DSLRFN (AWFs)	64.74	72.18	33.74
DSLRFN (AGFs)	67.19	73.11	32.71

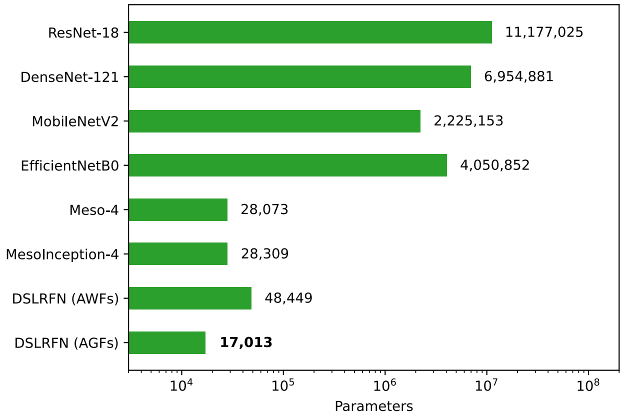


FIGURE 5. Comparison of model size in terms of the number of parameters.

the best generalizability results out of the three datasets used for training, the scores reported in Table 6 is when DFDC is used. An ablation study is conducted, results shown in Table 5, to show the importance of the dual-scale block and the self-attention mechanism.

D. DISCUSSION

In addition to the satisfactory performance of DSLRFN compared to state-of-the-art architectures, it is considered the most compact only utilizing a few number of parameters as shown in Fig. 5. Furthermore, AGFs reduced the architecture size by 64.9% compared to AWFs. It is noteworthy that this reduction is due to the large receptive field of filters in the first layer. If larger receptive fields are used, the size will be further

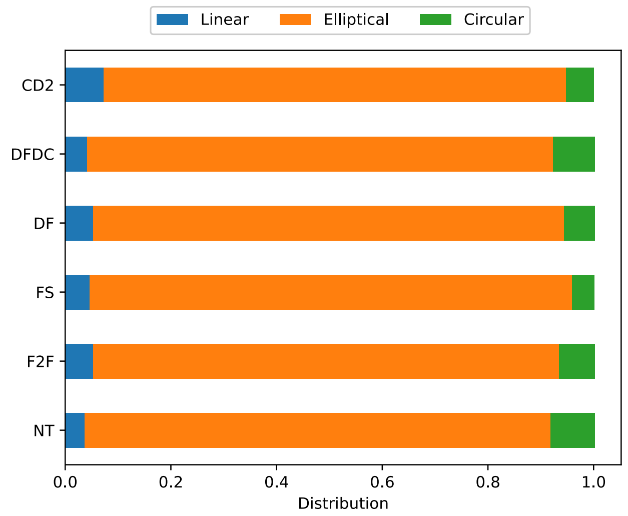


FIGURE 6. Categorical distribution of Gabor filters for each dataset.

reduced. In contrast to architectures based on AWFs, AGFs can be directly interpreted since all of the filters are based on a single mathematical function with a parameter set Ψ. Moreover, AGFs showed greater generalizability compared to AWFs as shown in Table 6. Since we introduced the parameter α that controls the shape of Gabor stripes, we show the distribution of α for AGFs along each dataset as shown in Fig. 6. It is clear that EGFs are widely utilized compared to LGFs and CGFs. Note that previously proposed Gabor models such as naive Gabor networks [21] and deep Gabor networks [24] can be viewed as special cases of the proposed AGFs. Therefore, AGFs provides a more generalizable framework compared to previous Gabor-based CNNs.

V. CONCLUSION

In this paper, we proposed a unified Gabor function capable of producing linear, elliptical, and circular Gabor filters. The proposed function is applicable to images that has diverse shapes compared to the limited traditional Gabor function. A back-propagation learning framework was adopted to allow the adaptability of the proposed function in CNNs. In contrast to conventional adaptive weighted filters, adaptive Gabor filters enable the utilization of large receptive field without parabolic increase in the number of learnable parameters. While deep architectures could be designed solely using the proposed function, it could also be used with adaptive weighted filters within the same architecture. While the underlying function could be applied to a variety of visual pattern recognition problem, a dual-scale large receptive field network (DSLRFN) was developed for deepfake image recognition. DSLRFN consists of a dual scale convolution, self-attention mechanism, high-level embedding block, and a simple classifier. The proposed architecture demonstrated its performance compared to other state-of-the-art models on CD2, DFDC, FF++ and WDF datasets with a substantially smaller number of parameters.

REFERENCES

[1] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2013, pp. 1–14.

[2] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[3] Y. Ma, J. Liu, Y. Liu, H. Fu, Y. Hu, J. Cheng, H. Qi, Y. Wu, J. Zhang, and Y. Zhao, "Structure and illumination constrained GAN for medical image enhancement," *IEEE Trans. Med. Imag.*, vol. 40, no. 12, pp. 3955–3967, Dec. 2021.

[4] S. Duan, Z. Chen, Q. M. J. Wu, L. Cai, and D. Lu, "Multi-scale gradients self-attention residual learning for face photo-sketch transformation," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 1218–1230, 2021.

[5] R. Yi, M. Xia, Y.-J. Liu, Y.-K. Lai, and P. L. Rosin, "Line drawings for face portraits from photos using global and local structure based GANs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3462–3475, Oct. 2021.

[6] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A large-scale challenging dataset for DeepFake forensics," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3204–3213.

[7] S. Agarwal and H. Farid, "Protecting world leaders against deep fakes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops*, Jun. 2019, pp. 38–45.

[8] D. M. Montserrat, H. Hao, S. K. Yarlagadda, S. Baireddy, R. Shao, J. Horvath, E. Bartusiak, J. Yang, D. Guera, F. Zhu, and E. J. Delp, "Deepfakes detection with automatic face weighting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 1–9.

[9] Y. Li, M.-C. Chang, and S. Lyu, "In ictu oculi: Exposing AI created fake videos by detecting eye blinking," in *Proc. IEEE Int. Workshop Inf. Forensics Secur. (WIFS)*, Dec. 2018, pp. 1–7.

[10] T. Jung, S. Kim, and K. Kim, "DeepVision: Deepfakes detection using human eye blinking pattern," *IEEE Access*, vol. 8, pp. 2169–3536, 2020.

[11] Y. Mirsky and W. Lee, "The creation and detection of deepfakes: A survey," *ACM Comput. Surv.*, vol. 54, no. 1, pp. 1–41, Jan. 2021.

[12] Y. Li and S. Lyu, "Exposing deepfake videos by detecting face warping artifacts," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Nov. 2019, pp. 46–52.

[13] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[14] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun, "Large kernel matters—Improve semantic segmentation by global convolutional network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4353–4361.

[15] Y. Ma, Y. Luo, and Z. Yang, "PCFNet: Deep neural network with predefined convolutional filters," *Neurocomputing*, vol. 382, pp. 32–39, Mar. 2020.

[16] Y. Zhang, W. Li, L. Zhang, X. Ning, L. Sun, and Y. Lu, "Adaptive learning Gabor filter for finger-vein recognition," *IEEE Access*, vol. 7, pp. 159821–159830, 2019.

[17] C. Liu, J. Li, L. He, A. J. Plaza, S. Li, and B. Li, "Naive Gabor networks for hyperspectral image classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 1, pp. 376–390, Jan. 2021.

[18] H. T. Le, S. L. Phung, P. B. Chapple, A. Bouzerdoum, C. H. Ritz, and L. C. Tran, "Deep Gabor neural network for automatic detection of mine-like objects in sonar imagery," *IEEE Access*, vol. 8, pp. 2169–3536, 2020.

[19] S. Luan, C. Chen, B. Zhang, J. Han, and J. Liu, "Gabor convolutional networks," *IEEE Trans. Image Process.*, vol. 27, no. 9, pp. 4357–4366, Sep. 2018.

[20] P. Jiang, B. Wan, Q. Wang, and J. Wu, "Fast and efficient facial expression recognition using a Gabor convolutional network," *IEEE Signal Process. Lett.*, vol. 27, pp. 1954–1958, 2020.

[21] K.-K. Huang, C.-X. Ren, H. Liu, Z.-R. Lai, Y.-F. Yu, and D.-Q. Dai, "Hyperspectral image classification via discriminant Gabor ensemble filter," *IEEE Trans. Cybern.*, early access, Feb. 5, 2021, doi: 10.1109/TCYB.2021.3051141.

[22] J. Bai, Y. Zeng, Y. Zhao, and F. Zhao, "Training a v1 like layer using Gabor filters in convolutional neural networks," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2019, pp. 1–8.

[23] F. Meng, X. Wang, F. Shao, D. Wang, and X. Hua, "Energy-efficient Gabor kernels in neural networks with genetic algorithm training method," *Electronics*, vol. 8, no. 1, pp. 1–18, 2019.

[24] Y. Yuan, J. Zhang, and Q. Wang, "Deep Gabor convolution network for person re-identification," *Neurocomputing*, vol. 378, pp. 387–398, Feb. 2020.

[25] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: A compact facial video forgery detection network," in *Proc. IEEE Int. Workshop Inf. Forensics Secur. (WIFS)*, Dec. 2018, pp. 1–7.

[26] H. H. Nguyen, J. Yamagishi, and I. Echizen, "Capsule-forensics: Using capsule networks to detect forged images and videos," in *Proc. ICASSP - IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 2307–2311.

[27] E. Kim and S. Cho, "Exposing fake faces through deep neural networks combining content and trace feature extractors," *IEEE Access*, vol. 9, pp. 123493–123503, 2021.

[28] L. Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wen, and B. Guo, "Face X-ray for more general face forgery detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5000–5009.

[29] T. Zhao, X. Xu, M. Xu, H. Ding, Y. Xiong, and W. Xia, "Learning self-consistency for deepfake detection," 2020, *arXiv:2012.09311*.

[30] D. Feng, X. Lu, and X. Lin, "Deep detection for face manipulation," in *Proc. Int. Conf. Neural Inf. Process.*, 2020, pp. 316–323.

[31] Y. Yu, R. Ni, and Y. Zhao, "Mining generalized features for detecting AI-manipulated fake faces," 2020, *arXiv:2010.14129*.

[32] L. Guarnera, O. Giudice, and S. Battiato, "Fighting deepfake by exposing the convolutional traces on images," *IEEE Access*, vol. 8, pp. 165085–165098, 2020.

[33] R. Durall, M. Keuper, F.-J. Pfreundt, and J. Keuper, "Unmasking DeepFakes with simple features," 2019, *arXiv:1911.00686*.

[34] J. Frank, T. Eisenhofer, L. Schönherr, A. Fischer, D. Kolossa, and T. Holz, "Leveraging frequency analysis for deep fake image recognition," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2020, pp. 3247–3258.

[35] Y. Qian, G. Yin, L. Sheng, Z. Chen, and J. Shao, "Thinking in frequency: Face forgery detection by mining frequency-aware clues," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 86–103.

[36] H. Liu, X. Li, W. Zhou, Y. Chen, Y. He, H. Xue, W. Zhang, and N. Yu, "Spatial-phase shallow learning: Rethinking face forgery detection in frequency domain," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 772–781.

[37] J. Wang, Z. Wu, J. Chen, and Y.-G. Jiang, "M2TR: Multi-modal multi-scale transformers for deepfake detection," 2021, *arXiv:2104.09770*.

[38] J. Zhang, T. Tan, and L. Ma, "Invariant texture segmentation via circular Gabor filters," in *Proc. Int. Conf. Pattern Recognit. (ICPR)*, 2002, pp. 901–904.

[39] G.-H. Hu, "Automated defect detection in textured surfaces using optimal elliptical Gabor filters," *Optik-Int. J. Light Electron Opt.*, vol. 126, no. 14, pp. 1331–1340, 2015.

[40] L. He, J. Li, A. Plaza, and Y. Li, "Discriminative low-rank Gabor filtering for spectral–spatial hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 3, pp. 1381–1395, Mar. 2017.

[41] N. Bonettini, E. Daniele Cannas, S. Mandelli, L. Bondi, P. Bestagini, and S. Tubaro, "Video face manipulation detection through ensemble of CNNs," 2020, *arXiv:2004.07676*.

[42] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.

[43] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015, *arXiv:1502.03167*.

[44] B. Dolhansky, R. Howes, B. Pflaum, N. Baram, and C. Canton Ferrer, "The deepfake detection challenge (DFDC) preview dataset," 2019, *arXiv:1910.08854*.

[45] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner, "FaceForensics++: Learning to detect manipulated facial images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1–11.

[46] B. Zi, M. Chang, J. Chen, X. Ma, and Y.-G. Jiang, "WildDeepfake: A challenging real-world dataset for deepfake detection," 2021, *arXiv:2101.01456*.

[47] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[48] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269.

[49] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.

[50] M. Tan and Q. Len, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2019, pp. 1–11.

AHMED H. KHALIFA received the B.Sc. degree in electronics and communications engineering from the Arab Academy for Science, Technology and Maritime Transport, Egypt, where he is currently pursuing the M.Sc. degree in computer engineering. His research interests include computer vision and wireless communications.

NAWAL A. ZAHER received the B.Sc. and M.Sc. degrees in electronics and communications engineering from the Arab Academy for Science, Technology and Maritime Transport (AASTMT), Cairo, Egypt, in 2010 and 2013, respectively, and the Ph.D. degree in engineering and applied science from Aston University, Birmingham, U.K., in 2018. From 2010 to 2018, she worked as a Teaching Assistant in communications engineering with AASTMT, where she has been an Assistant Professor, since 2018. Her research interests include signal processing, image processing, and machine learning.

ABDALLAH S. ABDALLAH (Member, IEEE) received the Ph.D. degree from the Bradley Department of Electrical and Computer Engineering, Virginia Tech, in 2016. He has experiences working with the telecommunication industry giants INTEL and BROADCOM, as a Wireless Research Intern. He also worked with INTEL Company as an Embedded Systems Software Engineer, from 2016 to 2017, before he joined The Pennsylvania State University as a Faculty Member, in September 2017. He is currently an Assistant Professor of electrical and computer engineering with Penn State Erie-The Behrend College. His research interests include wireless networks, image processing, and video streaming over HTTP. He focuses on applying machine learning techniques to address critical problems in wireless networks, the Internet of Things (IoT), and signal identification and recognition systems. He consistently serves as a Regular Reviewer for several IEEE journals, including IEEE ACCESS.

MOHAMED WALEED FAKHR received the Ph.D. degree in neural networks and machine learning from the University of Waterloo, Canada, in 1993. He then joined the Speech Research Laboratory, NORTEL, Montreal, Canada, for five years, where he was a Researcher investigating and implementing different speech processing, speech recognition, language modeling, and statistical error analysis techniques, and has two patents with NORTEL. Since 1999, he has been a Professor with the Arab Academy for Science and Technology, Cairo, Egypt, with three years sabbatical at the University of Bahrain. He has been doing research in the areas of time series forecasting, deep neural networks, natural language processing, and privacy-preserving computing.