IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received July 30, 2021, accepted August 29, 2021, date of publication August 31, 2021, date of current version September 10, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3109441

Towards Explainable Ear Recognition Systems Using Deep Residual Networks

HAMMAM ALSHAZLY[1,2], CHRISTOPH LINSE[1], ERHARDT BARTH[1], (Member, IEEE), SAHAR AHMED IDRIS[3], AND THOMAS MARTINETZ[1], (Senior Member, IEEE)

[1]Institute for Neuro- and Bioinformatics, University of Lübeck, 23562 Lübeck, Germany
[2]Faculty of Computers and Information, South Valley University, Qena 83523, Egypt
[3]College of Industrial Engineering, King Khalid University, Abha 62529, Saudi Arabia

Corresponding author: Hammam Alshazly (hammam.alshazly@sci.svu.edu.eg)

This work was supported by the Deanship of Scientific Research at King Khalid University through the Research Groups Program under Grant RGP.1/95/42. The work of Christoph Linse was supported by the Bundesministeriums für Wirtschaft und Energie (BMWi) through the Mittelstand 4.0-Kompetenzzentrum Kiel Project.

ABSTRACT This paper presents ear recognition models constructed with Deep Residual Networks (ResNet) of various depths. Due to relatively limited amounts of ear images we propose three different transfer learning strategies to address the ear recognition problem. This is done either through utilizing the ResNet architectures as feature extractors or through employing end-to-end system designs. First, we use pretrained models trained on specific visual recognition tasks, inititalize the network weights and train the fully-connected layer on the ear recognition task. Second, we fine-tune entire pretrained models on the training part of each ear dataset. Third, we utilize the output of the penultimate layer of the fine-tuned ResNet models as feature extractors to feed SVM classifiers. Finally, we build ensembles of networks with various depths to enhance the overall system performance. Extensive experiments are conducted to evaluate the obtained models using ear images acquired under constrained and unconstrained imaging conditions from the AMI, AMIC, WPUT and AWE ear databases. The best performance is obtained by averaging ensembles of fine-tuned networks achieving recognition accuracy of 99.64%, 98.57%, 81.89%, and 67.25% on the AMI, AMIC, WPUT, and AWE databases, respectively. In order to facilitate the interpretation of the obtained results and explain the performance differences on each ear dataset we apply the powerful Guided Grad-CAM technique, which provides visual explanations to unravel the black-box nature of deep models. The provided visualizations highlight the most relevant and discriminative ear regions exploited by the models to differentiate between individuals. Based on our analysis of the localization maps and visualizations we argue that our models make correct prediction when considering the geometrical structure of the ear shape as a discriminative region even with a mild degree of head rotations and the presence of hair occlusion and accessories. However, severe head movements and low contrast images have a negative impact of the recognition performance.

INDEX TERMS Ear recognition, biometrics, deep residual networks, transfer learning, deep ensembles, visual explanations, explainable prediction.

I. INTRODUCTION

Personal identification based on biological characteristics, including physiological (e.g., face, iris, retina, fingerprints, etc.) or behavioral (e.g., voice, signature, gait, gesture, etc.) modalities, has established itself as the most convenient means of reliable and fast recognition of individuals. Nevertheless, biometric systems based on the physiological characteristics are found to have a high level of reliability due to their robustness against stress effects and being relatively more stable throughout the life of individuals.

Recently, research has opened up new biometrics for personal identification, such as the ear shape. A number of studies has been conducted to explore the unique characteristics of human ears as an appealing alternative for or addition to common biometrics. Compared to conventional biometric modalities such as faces and fingerprints, ears provide some unique features and are considered a rich source of information for human identification. The characteristics of human ears include a rich structure of textural features, stability over

The associate editor coordinating the review of this manuscript and approving it for publication was Vicente Alarcon-Aquino.

a long period of time, robustness against external factors such as aging and facial expressions and less intrusiveness and sensitivity to be captured. The existing of bilateral symmetry between the left and right ears of the same subjects makes the ear discriminative enough to distinguish different individuals robustly [1]–[4]. Given all these remarkable characteristics, ear recognition has become an active area of research with several potential applications in forensics, security, monitoring and surveillance.

Many studies have been conducted to build recognition systems that consider ear images for recognizing individuals. The early ear recognition techniques were constructed by employing handcrafted feature extraction (i.e., descriptor) methods to extract the discriminative information in the form of feature vectors from ear images. Subsequently, a traditional classifier was adopted for matching and classifying the resulting feature vectors to identify individuals. Based on the feature extraction method, the existing techniques are mainly classified into four categories including geometric, holistic, local and hybrid methods [5]. The techniques were designed to exploit specific image features and to handle image variations to achieve acceptable recognition performance. However, their recognition accuracy drops when testing on ear images with wide ranges of image variations [6]–[8]. Changes in illumination, viewing angles, noise, low contrast, and partial occlusions by hair, earrings, or accessories are commonly encountered challenges in real-life scenarios. Therefore, ear recognition experiments have to be conducted under such unconstrained conditions to obtain robust and reliable identification systems.

With recent advancements in computer vision, machine learning techniques, and particularly deep convolutional neural networks (CNNs) many recognition tasks are now solvable under unconstrained imaging conditions. Deep CNNs have been developing at a fast pace and have become a popular data driven learning strategy for various computer vision problems. They combine feature extraction and classification into one end-to-end model. Moreover, training deep neural networks comes with the additional advantage of learning the representation of the input data to suit the particular problem. The benefit of also learning the features leads to the high adaptiveness of deep learning strategies, but also comes with a cost. Training deep CNNs requires large amounts of data in order to alleviate a common phenomenon in machine learning which is called over-fitting. The issue of over-fitting arises if an identification system uses noise to memorize the training images without actually processing the underlying relationships in the data to differentiate individuals. Furthermore, public extensive datasets are still lacking in the field of ear recognition and, as a consequence, deep CNNs have not yet been utilized so extensively for the ear modality. Key approaches to tackle these limitations are to perform aggressive data augmentation, reduce the model size, apply regularization techniques or perform transfer learning using pretrained models from other large-scale datasets like ImageNet [9]. In this work we integrate these cues and propose three transfer learning strategies for domain adaptation to improve performance of ear recognition systems.

While deep CNNs achieve superior recognition performance in various vision tasks, they are often criticized for being black-boxes. Their nested nonlinear structure and the lacking decomposability into easily understandable components make them hard to interpret. Consequently, these CNN-based systems may fail without any warnings or explanations why they did fail, which is a problem especially for security related applications. In order to build confidence in the decisions made by our recognition system, we provide visual explanations, interpret what deep learning models actually learn and visualize how they make their predictions. The visual explanations can answer the most common questions about CNNs including: What type of features do they learn? Which ear regions are considered more discriminative? What causes a system to fail? Do deep CNNs extract discriminative features that make sense to humans? We apply Guided Gradient-weighted Class Activation Mapping (Grad-CAM) [10] and conduct extensive experiments using deep CNN models of varying depths to obtain useful insights and to provide rationale answers to these questions. This technique can generate class-discriminative and high-resolution visualizations, which are useful to justify the predictions of deep models. This will assist in getting useful insights on the internal representations of deep models and make the constructed recognition systems more transparent and easy to interpret. These are considered important factors for a meaningful integration of deep models in the biometric domains.

In this paper, our aims are to construct ear recognition systems, which work end-to-end and achieve the best performance, as well as to provide a sufficient level of transparency and explainability. Overall, the contributions of this work can be summarized as follows:

- We present deep residual learning-based personal identification models using ear images. We implement and adopt five deep ResNet architectures of gradually increasing depth to suit the ear recognition tasks. The models are pretrained on the ImageNet dataset. They are trained and tested on four benchmark datasets, which contain ear images acquired under different imaging conditions ranging from relatively unconstrained settings to fully uncontrolled conditions.
- We propose three transfer learning strategies in order to learn discriminative ear features and to achieve the best recognition accuracy. First, we investigate their performance when training the fully connected layer to engage the ear recognition task. Subsequently, we fine-tune entire models on ear datasets. Third, we utilize the extracted features from the fine-tuned models to train machine learning classifiers such as Support vector machines (SVMs). Finally, we combine different models to form one robust voting committee for better generalization and improving the overall recognition performance.

- The ear images from the AMIC and AWE datasets comprise different spatial resolutions and aspect ratios. Resizing them to a fixed input dimension may introduce unwanted geometric distortions. We investigate different input sizes for the networks to deal with the varying input size problem and preserve the aspect ratios of the ear images. We also apply two different data augmentation strategies to improve the generalization of models on the different datasets. Extensive experiments are conducted and the obtained results reveal the effectiveness of our configurations and adaptation strategies.
- We achieve state-of-the-art recognition results on the considered ear datasets through ensembles of heterogeneous ResNet models. Our deep ensembles achieve a rank-1 accuracy of up to 99.64% and 98.57% on two datasets with mild image variations, and up to 81.89% and 67.25% on two datasets with extensive image variations. The reported recognition performance advances the recently published results under both imaging conditions from 2% to 5%.
- We provide class-discriminative and high-resolution visual explanations from different ResNet-based models of various depths. The visualizations highlight the most discriminative ear regions that are responsible for making predictions. They also give useful insights on the failure cases and provide reasonable explanations.

The remaining sections of the paper are organized as follows. Section II reviews the recent work on constrained and unconstrained ear recognition. In Section III we describe deep residual networks. The applied transfer learning methods are explained in Section IV. The experimental setup, the ear datasets, performance evaluation metrics and the procedure of training the models are explained in Section V. The experimental results are reported and discussed in Section VI. Section VII provides visual explanations from different ear recognition models. Finally, the paper is concluded in Section VIII.

II. RELATED WORK

This section discusses the most relevant work on ear recognition techniques in the literature. We refer to approaches that base on ear images acquired under constrained and unconstrained imaging conditions and highlight their achieved recognition performances. For a comprehensive review and analysis of existing work of ear recognition refer to [5], [11], [12].

In [13] the authors experimented with different local texture descriptors for constructing robust human identification systems. The ear image features were extracted by the texture descriptors and then used for training various classifiers for identification. Experiments were conducted on the IIT Delhi-1 [14], IIT Delhi-2 [14], and USTB [15] ear datasets, which were collected under controlled imaging conditions with slight variations. The experiments revealed competitive results compared to other ear recognition strategies from the literature.

The authors of [16] defined a set of seven ear features and trained an efficient feed-forward artificial neural network to recognize individuals using ear images. The features were measured for a dataset with 51 right ears, which were captured in gray scale under the same angle (pose). Experiments were conducted with various training and test set sizes and various network configurations like different number of layers, number of neurons per layer and with and without the addition of noise. The obtained results indicated that the network can obtain a recognition accuracy of up to 95%.

In [17] local and global features of ear images were extracted in the frequency domain and combined for an improved recognition performance. The global features were first extracted by applying the Gabor-Zernike operator [18] to the entire image, whereas local features were extracted by applying the local phase quantization (LPQ) [19] descriptor. The global and local features were then combined using a genetic algorithm to find the optimum combination of features. Identification was performed using the nearest neighbor classifier with the Canberra distance. The experimental results on the IIT Delhi-1 [14], IIT Delhi-2 [14], and USTB [15] ear datasets achieved better performance compared with existing ear recognition methods.

A number of studies has been conducted for addressing the unconstrained ear recognition problem. In [20] an ear registration approach was proposed based on the scale invariant feature transform (SIFT) technique [21]. It attempted to create a homography transform between the gallery ears and the probe. The algorithm starts with segmenting the ear from the gallery as a preprocessing step and then analyzes each image to extract its SIFT feature points. For each probe image, the SIFT feature points are detected and for each point the gallery is searched to find correspondences. If four points are matched between the probe and the gallery then a perspective transform is calculated and the probe is registered. The images are then aligned and the distance between them is measured and the nearest galley image is used to identify the person. The technique was evaluated on a relatively unconstrained ear dataset and showed a certain degree of robustness with acceptable recognition accuracy of 96%.

In 2017 the first unconstrained ear recognition challenge (UERC) [22] was organized to assess the performance of ear recognition technology on a large-scale ear dataset. The assessment involves tightly cropped ear images that exhibit a wide range of variations in head movements (poses), illumination, image resolution, and occlusions. Eight ear recognition techniques were submitted and evaluated for the challenge. A comprehensive analysis was performed on the ability of the submitted approaches to cope with the various image variations in the data. The obtained results indicated the sensitivity of all tested approaches to changes in head poses. In 2019 another round of the UERC competition was organized for evaluating advancement in the ear recognition technology [23]. A comprehensive analysis was performed to assess the sensitivity of ear recognition models towards variations in image resolution, illumination, gender, ethnicity and obscuring parts of the ear by hair or earrings. Generally, the submitted approaches showed an improved recognition performance compared with the 2017 models. Nevertheless, the obtained results revealed negative impacts of using small image resolution and partial ear occlusions on the recognition performance.

In [24] the authors proposed a two-stage domain adaptation strategy to fine-tune deep CNN-based models for the unconstrained ear recognition problem. The authors performed an in-depth analysis of some factors including the dataset bias, illumination changes, aspect ratios and the impact of using data augmentation and alignment on the recognition performance. The experiments were conducted on the unconstrained UERC dataset [22] and revealed the efficacy of the proposed fine-tuning strategy. Further performance improvements were achieved by combining different deep CNN models along with data augmentation.

In [25] a framework that combines handcrafted and CNN-based features was proposed. The authors experimented with various feature combinations and reported improved performance when both features were combined together as they seem to be complementary to each other. The gain in recognition performance was also attributed to the normalization process they followed when conducting the experiments. Important remarks were concluded from the study with respect to the performance of both features and their role in solving the unconstrained ear recognition challenge.

Dodge *et al.* [26] investigated two different approaches to address the unconstrained ear recognition challenge. The first is a traditional feature-classifier pipeline approach and the other is a complete end-to-end system. They compared the performance of both approaches and noted that the features extracted from pretrained models combined with shallow classifiers achieve high performance in unconstrained ear recognition. On the other hand, constructing an end-to-end system by fine-tuning deep network architectures tends to over-fit due to the limited training data. In order to alleviate the later problem, the authors proposed an averaging ensemble of fine-tuned networks, which achieved the best recognition performance on the ear datasets used.

An experimental evaluation of several descriptor- and deep learning based ear recognition models was achieved in [27]. The authors studied the characteristics of the recognition techniques and the impact of various covariates including gender, ethnicity, accessories and head movements on the recognition performance. Identification experiments were carried out on the Annotated Web Ear (AWE) dataset [5]. The results indicated that the presence of accessories and head movements significantly affect the identification performance, whereas other covariates of gender and ethnicity only affected the performance to a limited extent.

Alshazly *et al.* [28] presented the first experimental study on the unconstrained EarVN1.0 dataset [29]. Different pretrained CNN models were fine-tuned using custom-sized inputs tailored specifically for each deep network. The fine-tuning strategy proved to achieve state-of-the-art recognition results with an accuracy above 93% using a single model. Moreover, an accuracy of 95.85% was achieved by an ensemble of deep models. The authors also provided visualizations of the learned features from different models and showed the ability of the models to distinguish between the different subjects.

In a recent study, a multi-modal biometric recognition system was proposed [30]. The system utilized images of the ear and face profiles to alleviate the shortcoming of using only a single biometric. The images of each biometric modality were first subjected to a feature extraction process using two independent histogram-based descriptors (LPQ [19] and local directional pattern (LDP) [31]). The extracted feature vectors from both descriptors were concatenated into a single high-dimensional feature vector, thereby combining complementary information from the spatial and frequency domains. Principle component analysis (PCA) was then applied on each combined feature vector for both modalities to reduce the dimensionality. The reduced feature vectors of both biometric modalities were concatenated through feature-level fusion. A kernel-based discriminative common vector (KDCV) approach was then applied on the combined feature set in order to select the most discriminative features. Personal identification was finally performed using the K-nearest neighbor classifier. Experiments were conducted on two benchmark datasets consisting of side face images, which are publicly available at the University of Notre Dame (UND), namely collection E (UND-E) [1] and collection J2 (UND-J2) [32]. The obtained results indicated an improved recognition performance of the multi-modal system compared to using any individual biometric modality.

We complement the body of existing work on ear recognition by experimenting on four benchmark datasets, which contain ear images acquired under constrained and unconstrained imaging conditions. We implement and adopt five variants of the ResNet architecture with increased depth and propose three different learning strategies. Also, we provide visual explanations for various models to uncover the black-box nature of deep networks and to make them more transparent.

III. DEEP RESIDUAL NETWORKS

Recent studies in deep learning have empirically evidenced that increasing the depth of neural networks has a significant influence on their success. The top performers [33], [34] on the ImageNet dataset exploit deeper models. Very deep networks can represent more complex functions and can learn features at different levels of abstractions. Thus, increasing the depth while taking care of over-fitting leads to improved performance. However, increasing the network depth by simply stacking more layers has two main drawbacks. First, training very deep networks becomes more difficult due to the vanishing gradient problem. When the gradient is back-propagated to earlier layers its value becomes small due to repeated multiplication and hence convergence of the

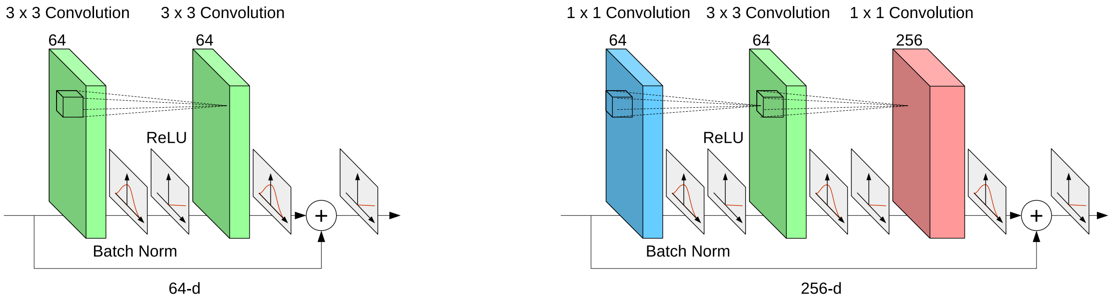


FIGURE 1. Two variants of the residual module building block. (a): The original residual module used in ResNet18 and ResNet34. (b): The bottleneck residual module utilized in ResNet50, ResNet101 and ResNet152.

TABLE 1. Configurations of the five ResNet architectures utilized in this study. The residual modules are shown in brackets and the number *n* of stacked modules is specified by the ×*n* next to the brackets.

Layer name	ResNet-18	ResNet-34	ResNet-50	ResNet-101	ResNet-152
Convolution	7 × 7, 64, stride 2				
Pooling	3 × 3, max pooling, stride 2				
Block (1)	$\begin{bmatrix}3\times3,64\\3\times3,64\end{bmatrix}\times 2$	$\begin{bmatrix}3\times3,64\\3\times3,64\end{bmatrix}\times 3$	$\begin{bmatrix}1\times1,64\\3\times3,64\\1\times1,256\end{bmatrix}\times 3$	$\begin{bmatrix}1\times1,64\\3\times3,64\\1\times1,256\end{bmatrix}\times 3$	$\begin{bmatrix}1\times1,64\\3\times3,64\\1\times1,256\end{bmatrix}\times 3$
Block (2)	$\begin{bmatrix}3\times3,128\\3\times3,128\end{bmatrix}\times 2$	$\begin{bmatrix}3\times3,128\\3\times3,128\end{bmatrix}\times 4$	$\begin{bmatrix}1\times1,128\\3\times3,128\\1\times1,512\end{bmatrix}\times 4$	$\begin{bmatrix}1\times1,128\\3\times3,128\\1\times1,512\end{bmatrix}\times 4$	$\begin{bmatrix}1\times1,128\\3\times3,128\\1\times1,512\end{bmatrix}\times 8$
Block (3)	$\begin{bmatrix}3\times3,256\\3\times3,256\end{bmatrix}\times 2$	$\begin{bmatrix}3\times3,256\\3\times3,256\end{bmatrix}\times 6$	$\begin{bmatrix}1\times1,256\\3\times3,256\\1\times1,1024\end{bmatrix}\times 6$	$\begin{bmatrix}1\times1,256\\3\times3,256\\1\times1,1024\end{bmatrix}\times 23$	$\begin{bmatrix}1\times1,256\\3\times3,256\\1\times1,1024\end{bmatrix}\times 36$
Block (4)	$\begin{bmatrix}3\times3,512\\3\times3,512\end{bmatrix}\times 2$	$\begin{bmatrix}3\times3,512\\3\times3,512\end{bmatrix}\times 3$	$\begin{bmatrix}1\times1,512\\3\times3,512\\1\times1,2048\end{bmatrix}\times 3$	$\begin{bmatrix}1\times1,512\\3\times3,512\\1\times1,2048\end{bmatrix}\times 3$	$\begin{bmatrix}1\times1,512\\3\times3,512\\1\times1,2048\end{bmatrix}\times 3$
Classification	adaptive average pooling				
	fully connected layer				

earlier layers is affected negatively. The second drawback is the performance degradation problem, as adding more layers to a network increases the parameter space that needs to be optimized and the training error can even increase [35], [36].

To address the above-mentioned problems, the authors in [37] introduced a deep residual learning framework that allows training very deep networks and the resulting new architectures were codenamed Residual Networks (ResNet). The layers are reformulated to learn a residual mapping with respect to the layer inputs. Let the desired mapping be denoted by $H(x)$ and the stacked layers fit a residual function $F(x)$ where $F(x) := H(x) - x$, where x is the input. Then, the desired mapping can be represented as the sum of the input and the residual function $H(x) = F(x) + x$. The operation $F(x)+x$ is carried out as illustrated in Figure 1. The shortcut connections simply perform identity mappings and their outputs are added to the outputs of the stacked layers.

The residual function $F(x)$ can have two convolutional layers as used in the shallower networks such as ResNet-18 and ResNet-34 or three convolutional layers as utilized in deeper ResNet variants, e.g. ResNet-50, ResNet-101 and ResNet-152. These two variants form the cornerstone of the ResNet architecture. Figure 1 depicts both variants and the two resulting residual modules. The module in Figure 1 (a) consists of two main branches. The first branch performs a series of 3 × 3 convolutions, batch normalization (BN) [38] and rectified linear unit (ReLU) as activation function [39]. The second branch is simply a shortcut, which connects the input of the module with the output of the first branch through element-wise addition. The sum is passed through another ReLU activation for increasing the non-linearity of the features. Figure 1 (b) depicts the other variant of the residual module, called bottleneck. The bottleneck module contains a simple extension, which adds an extra convolutional layer to the first branch of the module. The first layer consists of 1×1 filters, the second of 3×3 filters, and the third of 1×1 filters. The number of filters in the first two layers is $1/4$ the number of filters in the third layer. The bottleneck module performs better, in particular when training deeper networks.

A deep ResNet model is constructed by stacking multiple residual modules along with conventional convolution and pooling layers. Table 1 presents five ResNet architectures that are implemented and utilized in our study. The first layer performs convolution with 64 large kernels of size 7 × 7 and a stride of 2, followed by a max-pooling operation. Then, the architectures consist of four convolutional blocks, where

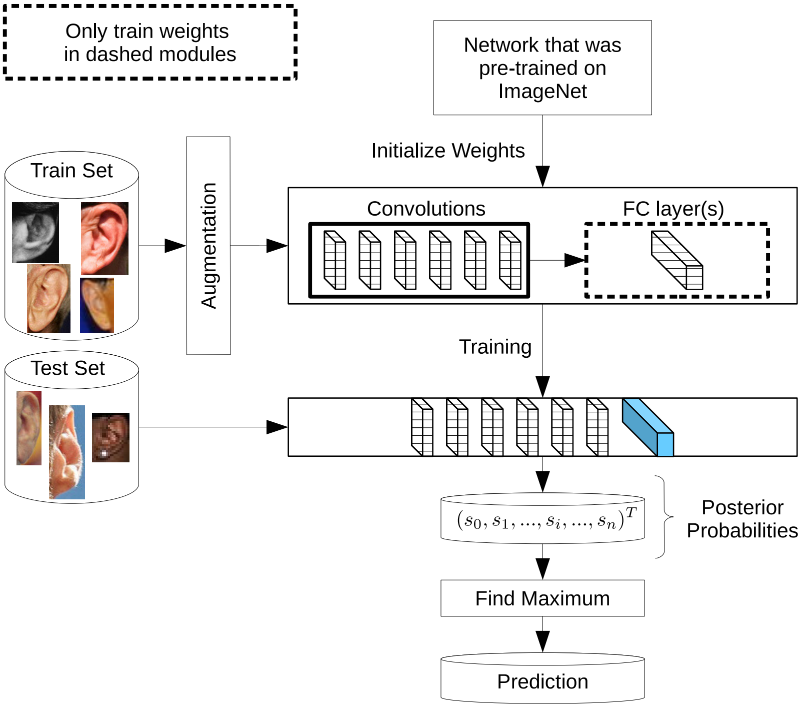


FIGURE 2. Feature extraction with pretrained networks. Only the weights in the fully connected part are adjusted during training.

each block has a specific number of stacked residual modules. The individual ResNet architectures differ in the type and number of modules in each block as can be seen in Table 1. After the four blocks a global average pooling (GAP) operation is performed to drastically reduce the spacial resolution of the feature maps to one element. During this process, the GAP layer reduces a feature map of dimensions $H \times W \times D$ to a size of $1 \times 1 \times D$. Finally, a fully connected layer with a softmax classifier is attached to perform the classification.

IV. TRANSFER LEARNING METHODS

Transfer learning is a powerful machine learning approach where models are developed for specific vision tasks and reused to initialize models for solving other tasks [40], [41]. In this paper we present three different transfer learning strategies applied to deep ResNet architectures to learn discriminative features from ear images. We describe these approaches in the following subsections in detail.

A. FEATURE EXTRACTION

Deep CNN-based systems learn different features at different layers. The nature of a layered architecture allows using pretrained models, such as ResNets, as fixed feature extractors. Feature extraction is accomplished by propagating the ear images forward through the network. The activations at an arbitrary layer are flattened and used as feature vectors for training traditional machine learning classifiers [42]–[45]. Table 2 shows the length of the feature vectors extracted from each of the ResNet variants. The table also describes other important characteristics like the model size and the number of trainable parameters.

TABLE 2. Comparison of the distinguishing characteristics of various ResNet models.

Model	Model size (MB)	Trainable parameters	Feature Size
ResNet18	42.83	11,227,812	512
ResNet34	81.39	21,335,972	512
ResNet50	90.46	23,712,932	2048
ResNet101	162.91	42,705,060	2048
ResNet152	222.58	58,348,708	2048

Our approach to perform feature extraction from pretrained ResNet architectures is to replace the fully connected layer with a new one, to initialize the new weights randomly and to fine-tune the weights of the new layer while freezing all other weights. This feature extraction method has two advantages. First, the network can be trained as an end-to-end identification system. Second, this enables us to utilize data augmentation, which can lead to better generalization. As Figure 2 shows, the images from the train set are propagated through the convolutional layers and multivariate feature vectors are obtained. They can be subsequently used to train the fully connected part of the network. As can be seen in the figure, the convolutional weights are not adjusted to the ear recognition problem. This strategy is called feature extraction.

Another modality of using feature extraction is shown in Figure 3. It makes use of traditional machine learning

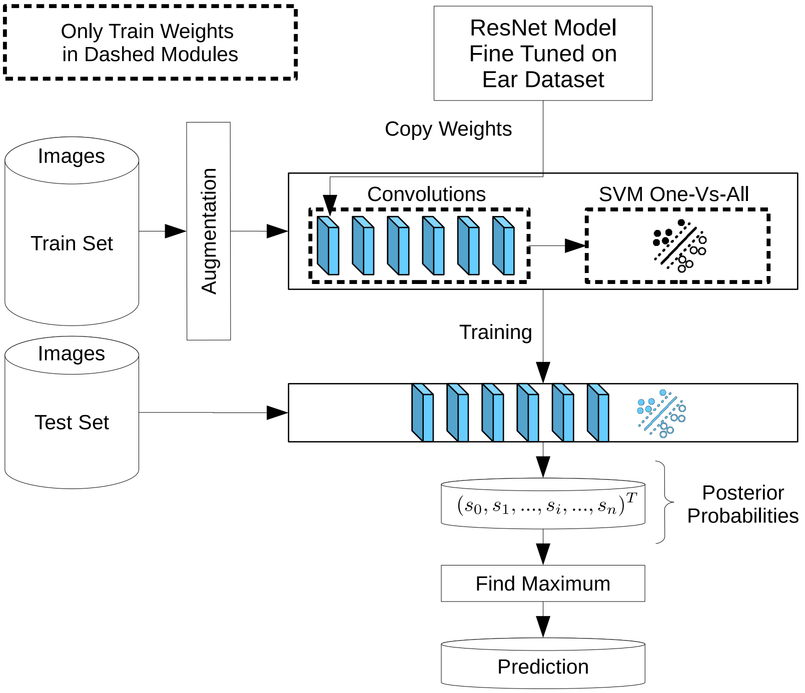


FIGURE 3. Feature extraction with fine-tuned networks and SVMs for classification.

classifiers such as SVMs [46]. The goal is to construct a hybrid model consisting of a CNN as a fine-tuned feature extractor and a set of SVMs to perform classification. As described in the diagram, the CNN has already been trained on ear images and therefore provides suitable representations of them. We train as many SVMs as there are individuals in each dataset in order to solve the multiclass-problem in a one-versus-all classification scenario. The training strategy is called feature extraction + SVM.

B. FINE-TUNING

Fine-tuning pretrained networks on the new task is another effective transfer learning strategy [47]. Applying fine-tuning allows to benefit from the robust and discriminative filters learned by top-performing CNNs on large-scale datasets as ImageNet and minimizes the domain divergence by adapting the filters for the target recognition task. However, this method requires us to perform network surgery. Figure 4 illustrates the fine-tuning process applied to the five pretrained ResNet architectures. The process starts with initializing the networks using the pretrained ImageNet weights. Then, we remove the last layer and add a fully connected layer with the suitable number of neurons matching the number of subjects in each dataset. The new fully connected layer is initialized with random weights. The process continues with training the fully connected layer to learn patterns from the highly discriminative convolutional layers, while freezing all other layers. Training the network is performed using a very small learning rate so that the new layer can start learning useful patterns from the previously learned layers. Once the fully connected layer is trained, we unfreeze the rest of the network and continue training the entire network until convergence. This strategy is called fine-tuning.

C. DEEP ENSEMBLES

Ensemble methods refer to the training of several models for the same task and then combining their predictions via averaging or voting in order to boost the performance. A common method to construct an ensemble of deep networks is to train several individual networks using different random initializations, and then averaging their predictions. Ensembles constructed by several deep networks are called deep ensembles [48], [49]. Deep ensembles have proven to substantially improve the classification performance of single models and they are top-performing approaches on the ImageNet competitions [50], [51]. It has also been found that a single large network may perform worse than an ensemble of several medium-size networks with the same total number of parameters [52].

In essence, constructing deep ensemble shares similarities with the multi-view learning approach which considers learning from multiple views of a particular dataset to improve the generalization performance [53]–[55]. The views can be acquired by various sensors or represented with different feature descriptors. For example, in case of images, texture information and color information represent two different features, which can be considered as two-view data.

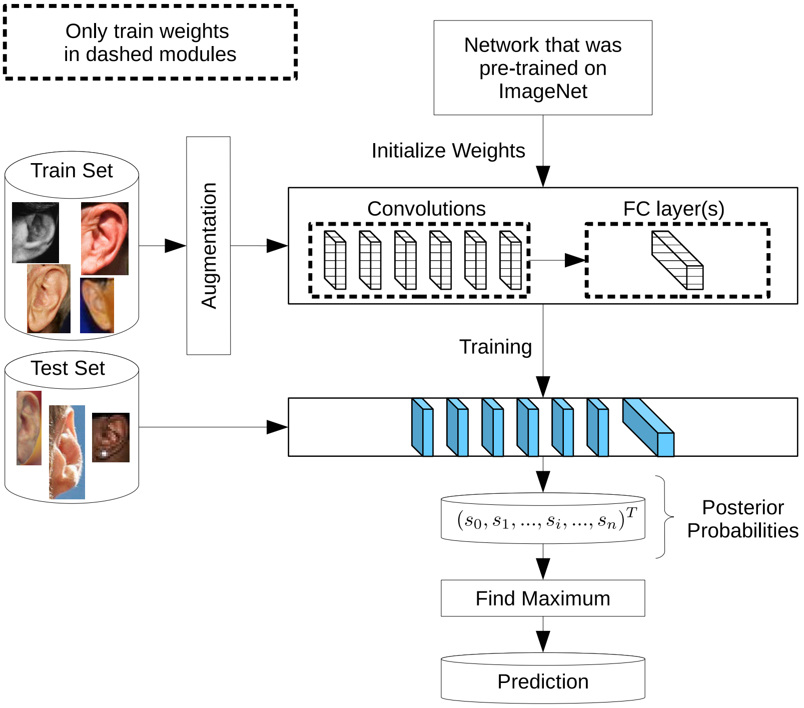


FIGURE 4. **The process of fine-tuning pretrained networks on ear images. The weights of all layers are adjusted during training.**

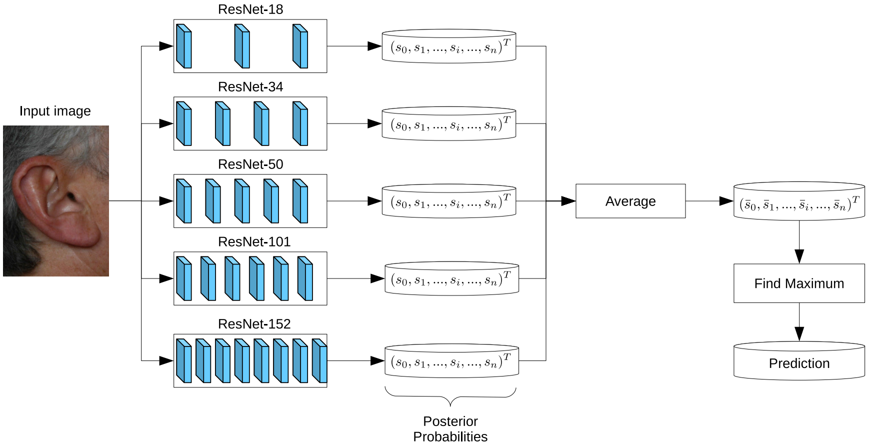


FIGURE 5. **Deep ensembles using fine-tuned ResNet models of various depths. The models are combined to generate a synergistic effect, which improves the performance over the single models.**

In multi-view learning the aim is to learn a specific function to model each view and is to jointly optimize all functions to boost the generalization performance. However, to benefit from multi-view learning it is necessary to have data from multiple and heterogeneous sources (views) that describe the given task. Moreover, dealing with a large number of views is a difficult task. In contrast, constructing deep ensemble could be achieved by independently training different networks of varying depths on the same data (single-view) and then combining the individual models to achieve better generalization performance.

In this work, we focus on constructing deep ensembles from various fine-tuned networks of varying depths. The various networks are combined to form one robust voting committee with potentially better generalization abilities. Figure 5 depicts the process, where the test images are fed into each network independently. The softmax function is applied to get the posterior probabilities (scores) for each class from the single votes. The posterior probabilities are then averaged and the mean scores are obtained. The maximal score belongs to the individual that is predicted by the ensemble.

V. EXPERIMENTAL SETUP

This section describes the ear image datasets and the data splitting procedure followed in our experiments. We also explain our two strategies of data augmentation to improve the generalization ability of our models. Finally, we depict the model training settings along with the specific configurations for each learning method.

A. EAR IMAGE DATASETS

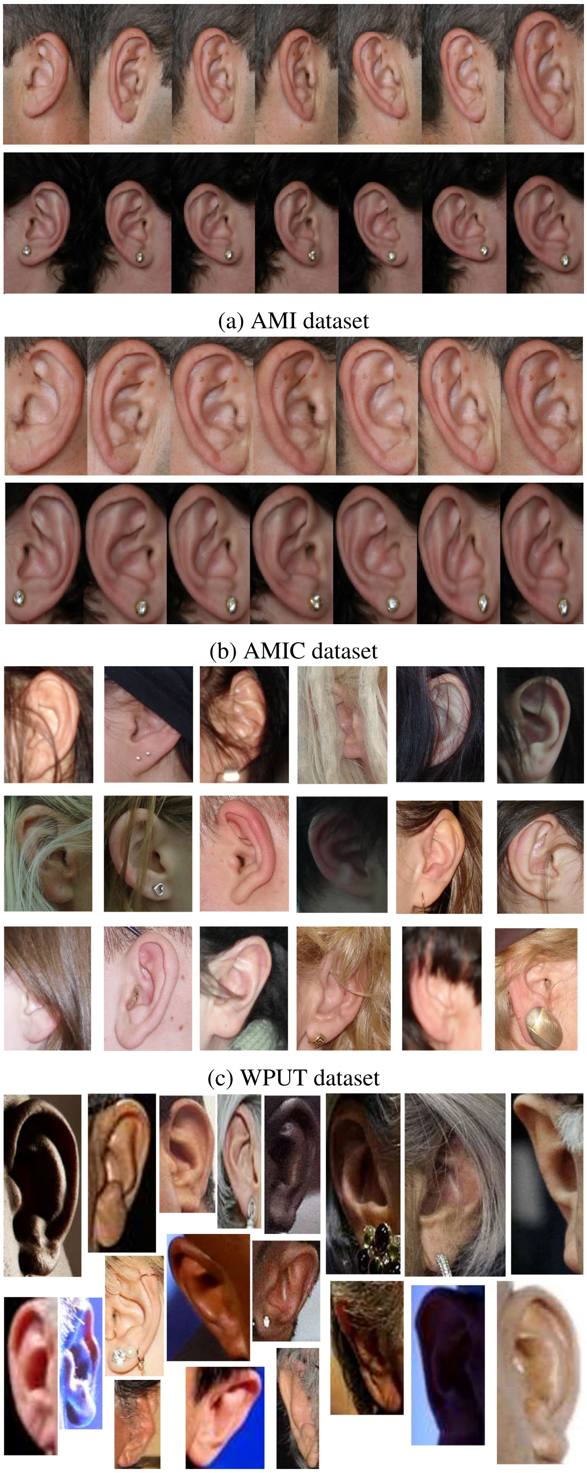
To evaluate the performance of the different models under each learning strategy we use four benchmark datasets with ear images of increasing difficulty from constrained to unconstrained image settings.

The first dataset is the Mathematical Analysis of Images (AMI) ear dataset [56]. It was collected from 100 subjects and has 700 ear images in total. Each subject has six images for the right ear and one image for the left one. Five images were collected for the right ear with different head poses such as looking forward, up, down, to the left, and to the right. The sixth image is also from the right ear with the subject facing forward but it is taken with a different focal length. The last image is a left side profile of the left ear with the subject facing forward. The subjects were photographed in an indoor environment with a Nikon D100 camera under consistent lighting conditions. The images have the same spatial resolution of 492 × 702 pixels. Example images from the AMI dataset are shown in Figure 6 (a) for one individual.

The second dataset is the AMIC ear dataset introduced in [57]. The dataset represents a tightly cropped version of the AMI dataset with the identical number of ear images and subjects. As a result of the cropping process, information such as skin texture or hair style is lost, which makes the AMIC dataset more challenging. Also, the resulting images have variable sizes ranging from 363 × 224 pixels to 492 × 702 pixels. Figure 6 (b) shows cropped ear images in Figure 6 (a) for the same individual.

The third ear image dataset comes from the West Pomeranian University of Technology (WPUT) [58], which contains images collected from males and females for the right and left ears. We use the cleaned version of the WPUT dataset considered in [57] that has 1960 images for 474 subjects and no duplicated images. Each subject has between 4 to 8 images. The images have a resolution of 380 × 500 pixels and were acquired under different illumination and viewing angles. Ears of some subjects are occluded by hair and accessories, which impose more challenges. Consequently, the WPUT dataset reflects real world scenarios of ear images taken under unconstrained conditions. Figure 6 (c) depicts example images from the WPUT dataset.

The last ear dataset is the Annotated Web Ear (AWE) database [5], which contains images collected from the internet. It represents one of the most challenging ear databases. It has 1000 ear images for 100 subjects and each subject has 10 images. The images exhibit a wide range of image variations such as variable spatial resolutions between 15 × 29 and 473 × 1022 pixels, various head poses, angles, varying


(a) AMI dataset

(b) AMIC dataset

(c) WPUT dataset

(d) AWE dataset

FIGURE 6. Examples of ear images from each dataset. The image variability starts from laboratory-like imaging conditions as in the AMI dataset and gradually increases to the fully unconstrained imaging conditions as in the AWE dataset.

illumination conditions, left and right ears, poor contrast, and in some cases major occlusions by earrings, accessories and hair. This dataset represents the in the wild scenario for evaluating ear recognition models. Figure 6 (d) presents sample images from the AWE ear dataset.

B. DATA SPLITTING AND EVALUATION METRICS

In order to conduct our recognition experiments we split each of the datasets into two disjoint sets: training and test set. 60% of ear images from each dataset is used for training, and the remaining 40% of images is used for testing. The training set is utilized for transfer learning, whereas the test set is utilized for reporting our results.

To analyze the characteristics of each recognition model we consider three standard quantitative metrics for performance evaluation. We also plot the Cumulative Match Characteristics (CMC) curves for each recognition experiment to summarize the performance of each model at different ranks.

The Cumulative Match Characteristics (CMC) curves are the most popular performance evaluation metrics for biometric identification methods. They are ranking-based metrics that show at which rank $R (R \leq N)$ a model returns the correct identity, where N is the number of subjects.

The Rank-1 (R1) recognition rate refers to the percentage of probe ear images for which the correct identity is returned as the top match from a gallery.

The Rank-5 (R5) recognition rate refers to the percentage of probe ear images for which the correct identity is returned within the top five matches from a gallery.

The Area under the curve (AUC) is an objective measure and an important evaluation metric to check the performance of identification models. A high AUC score indicates a high ability of the model to distinguish between the subjects.

C. DATA AUGMENTATION

Training deep CNNs with millions of parameters requires relatively large amounts of annotated training samples in order to overcome overfitting. Tackling these issues when experimenting with only few hundreds or thousands of ear images and few individuals is a challenging task. Data augmentation techniques are used to mitigate this problem and to introduce more variations to the training samples without any additional labeling costs. Various transformation steps are combined into a single preprocessing pipeline to generate image variants.

In this paper we employ two types and strategies of data augmentation. The first augmentation strategy uses strong random affine transformations like rotation and shearing and is found to give good results mainly on the AMI and AMIC datasets. The second one is less harsh considering affine transformations and is beneficial for the unconstrained datasets used in this work.

Moreover, some datasets such as the AMIC dataset contain images of different spatial resolutions. Therefore, dealing with the variable input size problem by normalizing images to a fixed image size is a must. In order to process images of various sizes and aspect ratios, the images are not aggressively deformed to fit a fixed size. Instead, they are scaled to fit into a canvas image of constant size, which defines the input size for the CNNs. The mean pixel value of the dataset is chosen as background color for the canvas image.

We give a detailed description of each strategy in the following subsections.

1) FIRST AUGMENTATION STRATEGY

The first data augmentation strategy makes strong use of rotation and shearing. Also, brightness and contrast are changed quite aggressively. The following ordered list shows image transformations applied to each training image.

- Scale the image randomly and paste it into the canvas, such that 70% to 100% of the canvas area is covered.
- Use the mean pixel value of ImageNet to fill the background of the canvas.
- Rotate the image randomly in the range of −45 and 45 degrees.
- Shear the image randomly up to 5% of the image width.
- Crop the image randomly to 90% to 100% of its original size, keeping the aspect ratio.
- Resize the image to canvas size.
- Blur the image with a probability of 50%. If so, blur with a Gaussian kernel of size 3.
- Mix the image with Gaussian noise of random amount.
- Modify image brightness randomly from −20% to 20%.
- Modify image contrast randomly from −40% to 40%.
- Modify image saturation randomly from −20% to 20%.
- Shift image hue from −5% to 5% of the entire color range.
- Flip the image horizontally with a probability of 50%.

2) SECOND AUGMENTATION STRATEGY

In the second data augmentation strategy some transformations are changed to be less extreme. Lowering the strength of random rotations and shearing empirically turned out to be beneficial for the unconstrained WPUT and AWE datasets, but seemed to have no influence on the AMI or AMIC datasets. The steps involved are:

- Scale the image randomly and paste it into the canvas, such that 80% to 100% of the canvas area is covered.
- Use the mean pixel value of ImageNet to fill the background of the canvas.
- Rotate the image randomly in the range of −20 and 20 degrees.
- Shear the image randomly up to 7.5% of the image width.
- Crop the image randomly to 90% to 100% of its original size, keeping the aspect ratio.
- Resize the image to canvas size.
- Blur the image with a probability of 20%. If so, blur with a Gaussian kernel either of size 3 (50% chance) or 4 (50% chance).
- Mix the image with Gaussian noise of random amount.

- Modify image brightness randomly from −20% to 20%.
- Modify image contrast randomly from −40% to 40%.
- Modify image saturation randomly from −20% to 20%.
- Shift image hue from −3% to 3% of the entire color range.
- Flip the image horizontally with a probability of 50%.

D. MODEL TRAINING

Three different training strategies are applied to analyze the recognition performance of the five ResNet architectures. The first learning strategy is called feature extraction and involves freezing all convolutional layers of the pretrained model and training the fully connected layer only. This strategy is explained in Section IV-A in detail. The second training strategy is called feature extraction + SVM and utilizes SVMs as final classifiers to work with the activation maps from the CNN. Multiple SVMs act together in an one-vs-all fashion to solve a multi-class recognition problem. The process is also described in Section IV-A. The third strategy performs fine-tuning of all layers of the pretrained ResNet model and is called fine-tuning. More information about this transfer learning strategy can be found in Section IV-B.

After training the ResNet variants, combinations of the deep models are integrated into ensembles. Section IV-C provides more detail of how the ensembles are created and how the networks vote together to make a final decision. Using this technique, various combinations of the five ResNet architectures are built and their performances are evaluated.

The learning rate is chosen to be a decremental step function. Training starts with an initial learning rate of 0.02 and is adjusted in distinct time steps measured in epochs. In a fixed scheduling interval the learning rate is multiplied with the constant factor 0.5. The batch size is kept constant for all models to promote comparability between the network architectures. For alleviating the issue of over-fitting, momentum is applied as regularization method. For preserving the image aspect ratios, the input size for the AMI and AMIC datasets are kept at 150 × 200 pixels. For the WPUT dataset the input size is chosen to be 170 × 220 pixels, because further reduction lead to a degradation of performance. The same argument applies to the choice of the image size for the AWE dataset, which is 130 × 250 pixels. The first augmentation strategy is applied on both, the AMI and the AMIC datasets. The second variant of data augmentation is applied on the WPUT and the AWE datasets, because too aggressive data augmentation worsens performance. Data augmentation turns out to be tremendously beneficial to the unconstrained datasets, since it boosts the R1 recognition rate of the ResNet-50 model on AWE for fine-tuning by about 20%.

Model training was carried out on a PC with Intel(R) Core(TM) i7-3770 CPU, 16 GB RAM and an Nvidia GTX 1080. When fine-tuning all layers during the second and third learning strategy, the models need 150 epochs and the learning rate is halved every 30 epochs. The first learning strategy needs 300 epochs until full convergence and the learning rate is halved every 60 epochs. A momentum of 0.9 is applied during stochastic gradient descent on the cross-entropy loss for all training strategies.

VI. RESULTS AND DISCUSSION

This section reports the experimental results of the ResNet-based models for the three different learning methodologies on four benchmark ear datasets. Table 3 summarizes the quantitative results in terms of R1, R5 and AUC. Moreover, the CMC curves under fine-tuning and feature-based SVM strategies for all assessed ResNet variants are plotted to show the performance differences in Figures 7, 8, 9 and 10. We also compare our obtained results with previous work from the literature when applicable.

A. FEATURE EXTRACTION

When only adjusting the fully connected layer of the ResNet models, which were pretrained on ImageNet, the obtained results are inferior, as can be seen in Table 3. Generally, the obtained results under the feature extraction strategy is far from being satisfactory on all datasets. One observes a significant drop in recognition rates for all models under this scenario on all datasets, proving the specific nature of the learned features from the ImageNet dataset and a necessity of domain adaptation of the learned features to suit the ear recognition task. Interestingly, the ImageNet features are not able to discriminate the AWE subjects at all, as the highest R1 is only 20% using ResNet-50. In addition, there does not seem to be a consistent relationship between model depth and performance for this learning strategy.

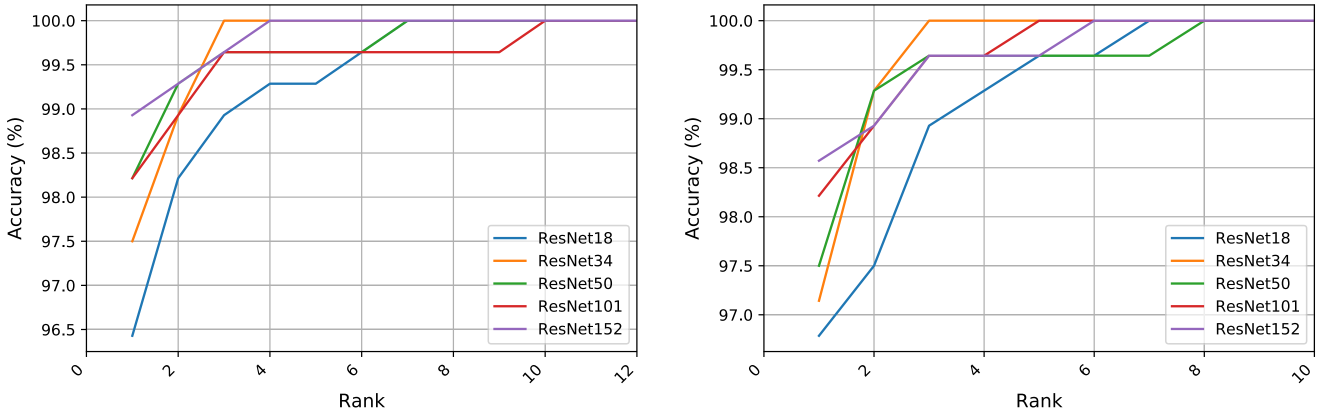
B. FINE-TUNING

During fine-tuning an improvement of around 30% in recognition performance is observed in comparison to feature extraction on all datasets. For the AMI dataset, ResNet-152 is a top performer with respect to all metrics achieving R1 accuracy of nearly 99%. It recognizes all ear images correctly within the top five matches with R5 of 100%. As a result, ResNet-152 improves on the previous state-of-the-art results of 97.84% on the AMI dataset presented in [57]. While the top R1 accuracy is 65% for the feature extraction strategy, the fine-tuning strategy offers almost no room for improvement on the AMI dataset. ResNet-152 is also the best performer on the AWE dataset targeting R1 of 62% and on the WPUT dataset with R1 of 78.83%. The previous top-performer on the WPUT dataset was presented in the paper [57] and achieved an accuracy of 79.08%. The previous single-model top-performer on the AWE dataset for ResNet variants was a ResNet-18 with an accuracy of 56.35% [26], which is about 6% below the present results. Moreover, our fine-tuned ResNet-152 achieves a similar R1 rate of 62% as the top-performer on the AWE dataset from [65], which was achieved by a fine-tuned SqueezeNet model and extensive data augmentation equivalent to 100-times of the original image (i.e., augmentation factor of 100).

The experiments substantiate the finding that ear recognition on the AMIC dataset is more difficult than on the AMI

TABLE 3. Comparison of quantitative performance metrics R1, R5, and AUC for different ResNet models under various learning strategies on each benchmark ear dataset. Results are given in percentages where the best value for each performance metric is highlighted in bold. We also compare our obtained results with the published work from the literature when applicable.

Strategy	Model	AMI			AMIC			WPUT			AWE		
		R1	R5	AUC	R1	R5	AUC	R1	R5	AUC	R1	R5	AUC
Feature Extraction	ResNet18	**65.00**	**87.50**	**97.26**	54.28	**79.28**	**95.75**	32.65	**58.16**	**94.84**	17.25	45.50	84.94
	ResNet34	58.92	83.57	96.49	48.21	75.71	94.19	29.46	49.36	93.35	19.50	46.25	**84.96**
	ResNet50	63.92	86.78	97.06	**54.64**	77.85	94.99	33.29	55.74	94.72	**20.00**	**47.25**	84.23
	ResNet101	62.85	83.92	95.77	47.85	67.50	89.64	**33.93**	56.51	94.15	16.25	40.00	80.81
	ResNet152	63.92	87.14	96.50	54.28	**79.28**	94.36	31.12	53.70	94.14	18.00	45.25	83.87
Feature Extraction + SVM	ResNet18	96.78	99.64	98.93	94.28	98.57	98.77	71.30	78.74	98.65	55.50	75.50	94.24
	ResNet34	97.14	**100**	98.97	95.00	98.57	98.83	73.60	88.78	98.56	54.75	79.75	94.81
	ResNet50	97.50	99.64	98.96	**96.42**	98.92	98.82	75.89	90.31	98.89	56.75	80.75	**95.54**
	ResNet101	98.21	**100**	**98.98**	96.07	**99.64**	**98.91**	75.89	90.69	99.00	55.50	79.00	94.98
	ResNet152	**98.57**	99.64	98.97	96.07	98.92	98.87	**77.55**	**91.58**	**99.21**	**61.50**	**81.25**	94.76
fine-tuning	ResNet18	96.42	99.28	98.93	92.14	97.85	98.68	71.05	87.12	98.56	53.50	74.25	93.83
	ResNet34	97.50	**100**	98.97	94.64	99.28	98.89	72.96	88.65	98.52	55.25	79.50	94.86
	ResNet50	98.21	99.64	98.96	95.70	98.21	98.75	76.28	89.80	98.83	57.00	80.50	**95, 60**
	ResNet101	98.21	99.64	98.85	95.70	98.92	98.87	77.17	90.43	98.86	59.00	81.00	95.26
	ResNet152	**98.92**	**100**	**98.98**	**95.71**	**100**	**98.95**	**78.83**	**91.07**	**99.06**	**62.00**	**81.25**	94.97
Deep Ensembles	ResNet18+ResNet50	98.57	99.64	98.96	96.79	98.93	98.87	78.06	89.67	98.83	62.50	80.00	95.40
	ResNet18+ResNet101	97.85	**100**	98.98	97.14	99.29	98.89	78.70	90.82	98.93	60.00	80.25	95.07
	ResNet18+ResNet152	98.57	**100**	98.97	96.79	99.29	98.95	79.59	91.07	98.98	63.50	80.50	95.11
	ResNet34+ResNet50	97.85	99.64	98.96	96.07	99.29	98.91	78.57	90.56	98.79	61.25	81.75	95.68
	ResNet34+ResNet101	98.57	**100**	**98.99**	97.14	99.64	98.92	79.08	90.69	98.86	62.00	82.00	95.52
	ResNet34+ResNet152	97.85	**100**	98.98	97.14	99.64	98.94	79.97	91.33	98.99	62.00	82.75	95.73
	ResNet50+ResNet101	98.92	**100**	**98.99**	96.43	99.29	98.93	79.97	91.20	98.98	63.50	82.00	95.79
	ResNet50+ResNet152	98.57	**100**	98.97	97.50	99.64	98.93	80.74	91.33	99.03	64.50	82.50	95.83
	ResNet101+ResNet152	**99.64**	**100**	**98.99**	97.85	99.64	98.92	81.38	91.96	**99.14**	65.50	83.00	95.76
	ResNet (50+101+152)	**99.64**	**100**	98.98	**98.57**	**100**	**98.98**	81.12	91.58	99.10	65.75	**85.50**	**96.06**
	ResNet (34+50+101+152)	98.57	**100**	**98.99**	97.85	99.64	98.96	**81.89**	**92.22**	99.03	**67.25**	84.00	96.03
Previous work	Raghavendra et al. [59]	86.36	-	-	-	-	-	-	-	-	-	-	-
	Alshazly et al. [60]	70.20	-	-	-	-	-	-	-	-	-	-	-
	Chowdhury et al. [61]	67.26	-	-	-	-	-	67.01	-	-	-	-	-
	Hassaballah et al. [62]	73.71	-	-	-	-	-	38.76	-	-	49.60	-	-
	Emersic et al. [5]	-	-	-	-	-	-	-	-	-	49.60	-	-
	Dodge et al. [26]	-	-	-	-	-	-	-	-	-	56.35	74.80	-
	Dodge et al. [26]	-	-	-	-	-	-	-	-	-	68.50	83.00	-
	Zhang et al. [63]	-	-	-	-	-	-	-	-	-	50.00	70.00	-
	Sultana et al. [64]	-	-	-	-	-	-	73.00	86.00	-	-	-	
	Emersic et al. [65]	-	-	-	-	-	-	-	-	-	62.00	80.35	95.51
	Alshazly et al. [66]	94.50	99.40	98.90	89.80	97.70	98.58	-	-	-	-	-	-
	Alshazly et al. [57]	97.50	99.64	98.41	93.21	96.78	98.63	79.08	90.43	98.92	-	-	-
	Omara et al. [67]	97.84	-	-	-	-	-	68.06	-	-	-	-	-
	Zhang et al. [68]	93.96	-	-	-	-	-	-	-	-	-	-	-
	Omara et al. [69]	96.82	-	-	-	-	-	-	-	-	-	-	-
	Khaldi and Benzaoui [70]	96.00	99.00	94.47	-	-	-	-	-	-	50.53	76.35	80.97
	Hassaballah et al. [71]	72.29	-	-	-	-	-	-	-	-	54.10	-	-
	Priyadharshini et al. [72]	96.99	-	-	-	-	-	-	-	-	-	-	-
	Khaldi et al. [73]	-	-	-	-	-	-	-	-	-	48.48	-	-
	Khaldi et al. [74]	98.33	-	-	-	-	-	-	-	-	51.25	-	-

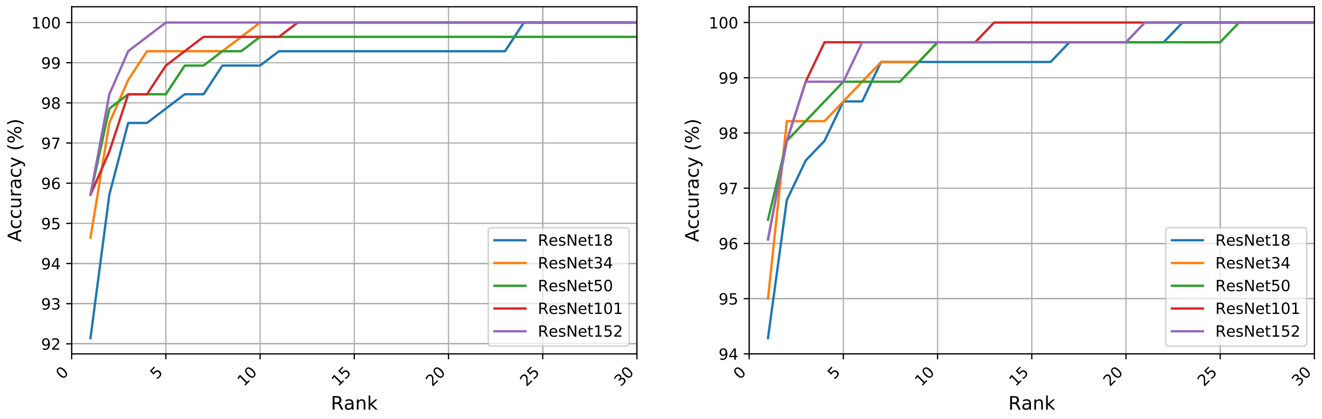


(a) Feature Extraction + SVM

(b) Fine-tuning

FIGURE 7. CMC curves generated on the test set from the AMI dataset for two different learning strategies and various ResNet models.

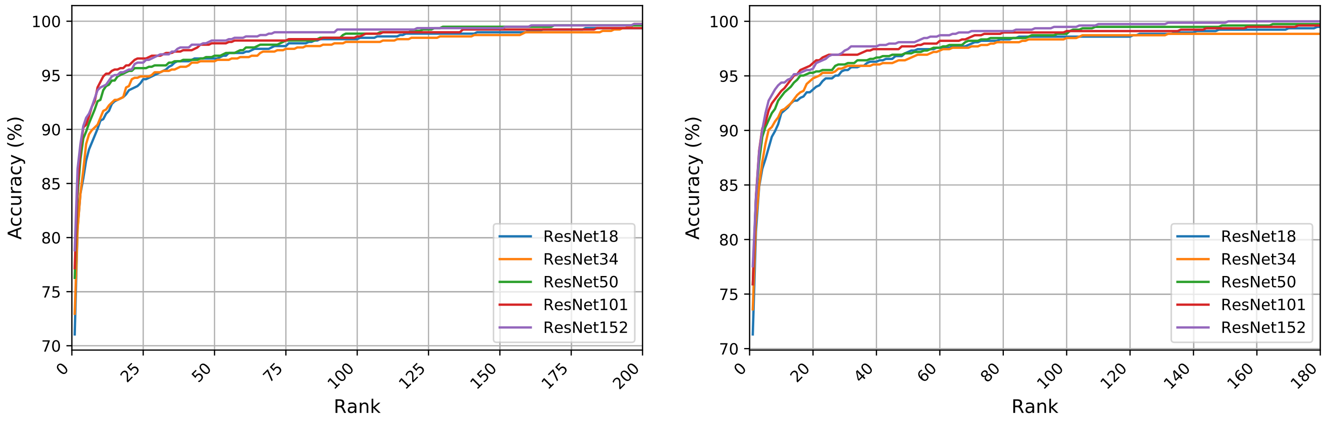
dataset. For the ResNet architectures we observe a performance drop of up to 4% because of the removal of auxiliary parts of skin and hair. Another observation from the results in Table 3 is the consistent phenomenon that recognition performance increases with network depth for the fine-tuning strategy. This gives a strong hint that the ability to represent



(a) Feature Extraction + SVM

(b) Fine-tuning

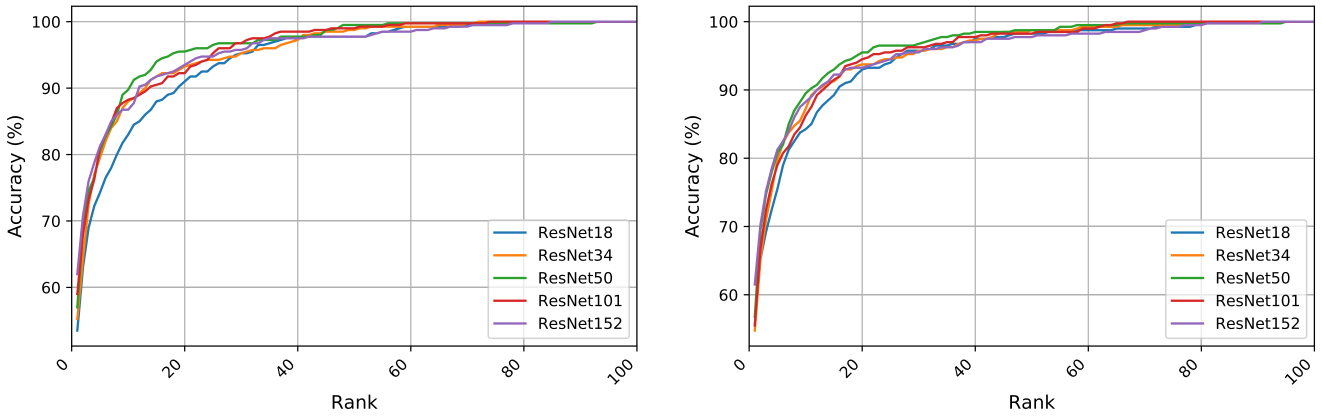
FIGURE 8. CMC curves generated on the test set from the AMIC dataset for two different learning strategies and various ResNet models.



(a) Feature Extraction + SVM

(b) Fine-tuning

FIGURE 9. CMC curves generated on the test set from the WPUT dataset for two different learning strategies and various ResNet models.



(a) Feature Extraction + SVM

(b) Fine-tuning

FIGURE 10. CMC curves generated on the test set from the AWE dataset for two different learning strategies and various ResNet models.

more discriminative features is higher when the network is deeper. However, this observation only holds if all layers are adjusted to the ear recognition problem.

C. FEATURE EXTRACTION + SVM

Hybrid models consisting of a fine-tuned feature extractor and an SVM classifier are employed to improve the overall performance. The dependence of model depth and increased performance is also found for this learning strategy. Using ResNet-152 as a backbone and training SVMs on its features improves performance by about 1% on the AMIC dataset. In these two cases the SVMs are able to generalize better on the rather small datasets because of their maximizing margin principle. Nevertheless, using SVMs does not lead to any significant improvement on the AMI, WPUT, and the AWE datasets. In order to attain these slightly improved results, no data augmentation was applied during the training of the SVMs at all. However, the same improvements can be obtained by switching data augmentation on during training the SVMs. Therefore, it seems that the CNNs were able to effectively learn the augmentation steps and that the extracted features might be somewhat invariant to the augmentation transformations.

D. DEEP ENSEMBLES

In order to build robust recognition systems and to improve performance even further, multiple models of varying depth are combined to form an ensemble. The top ensembles in Table 3 enhance the R1 recognition rate by 1% on the AMI dataset, about 3% on the AMIC dataset, about 3% on the WPUT dataset, and above 5% on the AWE dataset. Generally, including all five networks into an ensemble is not the best decision as the worse performing models do not add efficacious contributions. Thus, smaller ensembles made out of the deeper models turn out to perform better. When considering pair-wise ensembles only, the combination of ResNet-101 and ResNet-152 delivers the best R1 results for all datasets. The fact that the pair of the very deepest models perform best supports the hypothesis that deeper models perform better using the fine-tuning strategy. Across the datasets, the combination of the deepest models ResNet-50, ResNet-101 and ResNet-152 consistently shows very good results which makes this ensemble the new top state-of-the-art classifier for the AMI, AMIC and WPUT datasets.

VII. VISUAL EXPLANATIONS

The visualization technique applied in this work aims at providing class-discriminative (i.e. localizing discriminative regions in ear images) and high-resolution (i.e. capturing fine-grained details) visualizations to assist in understanding the predictions made by our ResNet models. The technique represents a combination of two powerful approaches called the Gradient-weighted Class Activation Mapping (Grad-CAM) [10] and Guided Backpropagation [75], which is named Guided Grad-CAM. The algorithm uses the former technique for providing class-discriminative explanations and the latter for creating high resolution visualizations. The Grad-CAM approach utilizes the class-specific gradient information to localize the important image regions and the localizations are then combined with the pixel-space visualizations to generate high-resolution and class-discriminative visualizations.

To explain the prediction of each model, we first select various ear images from the test set of each dataset where the models correctly identified the subjects. We also provide some cases of misclassified ear images in order to get useful insights and a better understanding of the false predictions. Then, we apply the Grad-CAM approach to produce low-resolution localization maps. Following the Grad-CAM algorithm, the localization maps are obtained by a weighted sum of the feature maps in the last convolutional layer. Then, the Guided Grad-CAM algorithm is applied to enhance the visualization with fine-grained details. Tables 4, 5 and 6 illustrate several examples of correct and misclassified ear images from each dataset along with the localization maps and the Guided Grad-CAM visualizations.

Table 4 presents localization maps for the AMI and AMIC datasets. We can observe that for correct decisions the model consistently focuses on the geometrical structure of the ear as the most discriminative region for making correct predictions. In the case of correct decisions the Guided Grad-CAM visualizations unveil that the model focuses on the geometrical shape of the entire ear, strips of the ear and that it ignores all background and hair information. In one of the correctly classified images a person wears glasses, which are ignored by the network. Usually, the contour of the ear helix is captured. Distinctive ear parts, that are frequently highlighted by the Guided Grad-CAM visualization, are the intertragal notch and the triangular fossa. Often, the lower part of the lobule is highlighted, too. These findings justify the superior performance of the ear recognition models and that they work as intended by considering the ear structure as a robust and distinguishable region for recognizing different subjects. In contrary, Table 4 also shows misclassified ear images from both datasets on the right hand side. The provided visualizations show that when the model focuses on textures at the ear boundary, it tends to overstate auxiliary information such as haircut or skin texture. There are some cases where the network ignores the shape of the entire ear and rather bases its decision on local ear regions or single edges that originate from the contrast where dark hair occludes the skin. There is also one example where a small skin is irregularity recognized by the network, which leads to a misclassification.

Table 5 shows example ear images that were correctly classified and misclassified from the WPUT dataset. Some of the images do not show the entire ear such that the geometrical structure of some ear parts stay unknown. As can be seen from the localization maps the model is able to correctly identify subjects even when substantial parts of the ear are occluded by hair or accessories. Indeed, in some cases where substantial parts of the ear are invisible, the models utilize auxiliary information from hair style and accessories

TABLE 4. Guided Grad-CAM visualizations from the AMI and AMIC ear datasets.

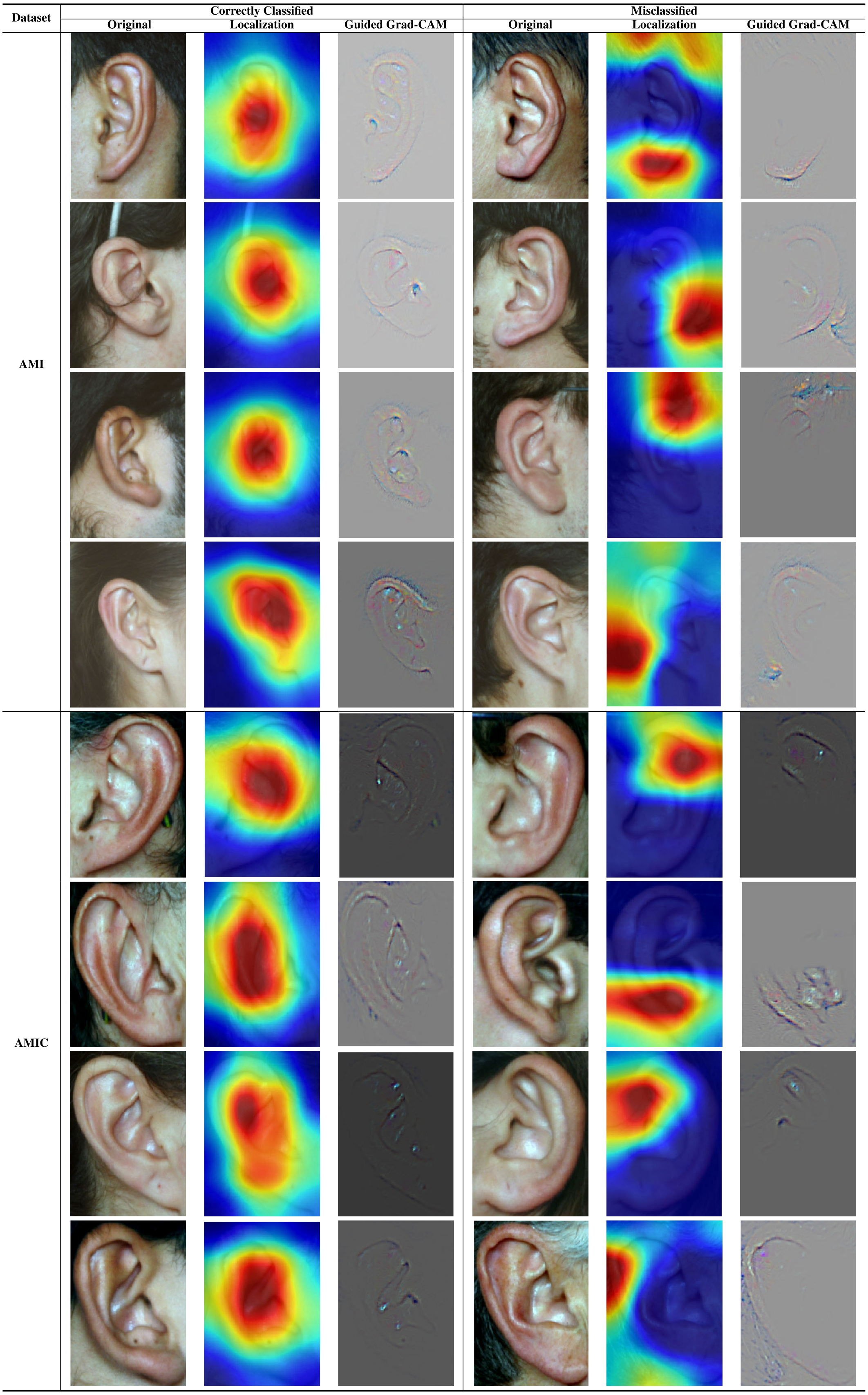


TABLE 5. Guided Grad-CAM visualizations from the WPUT ear dataset.

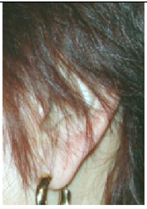
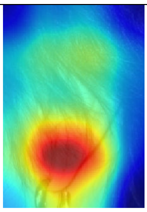
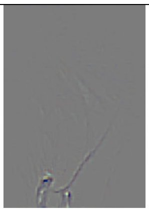
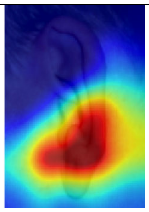
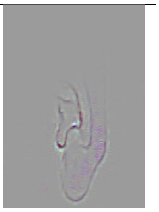
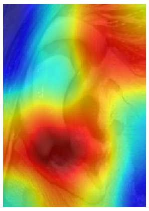
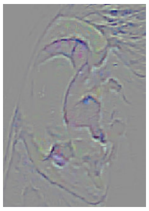
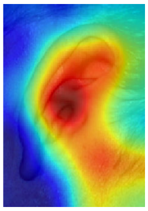
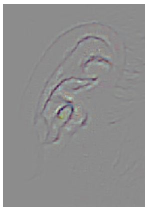
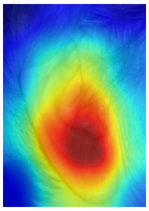
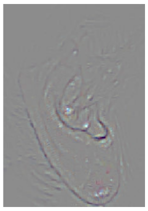
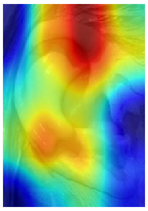
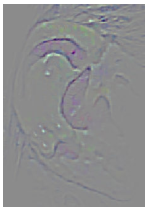
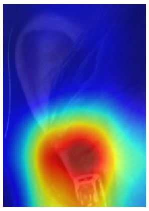
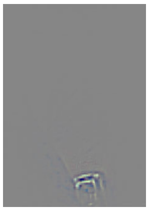
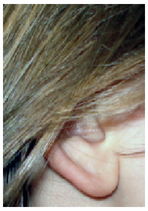
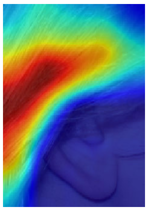
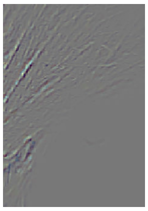
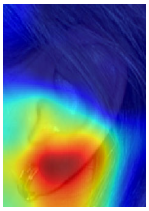
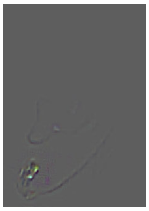
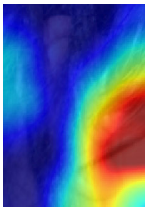
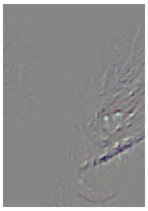
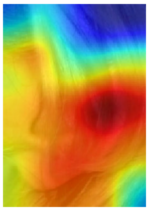
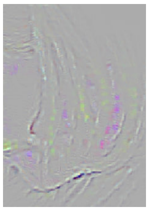
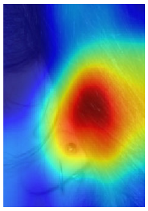
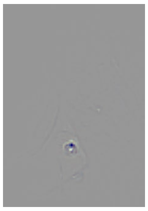
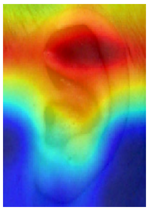
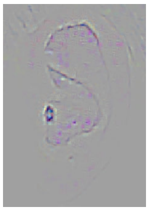
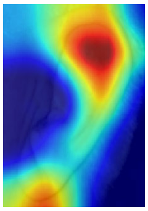
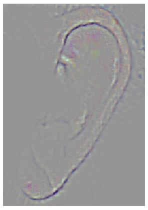
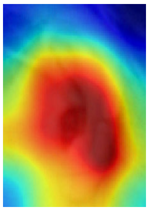
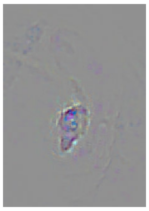
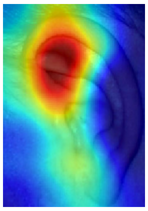
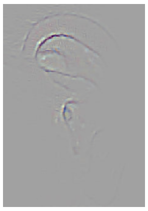
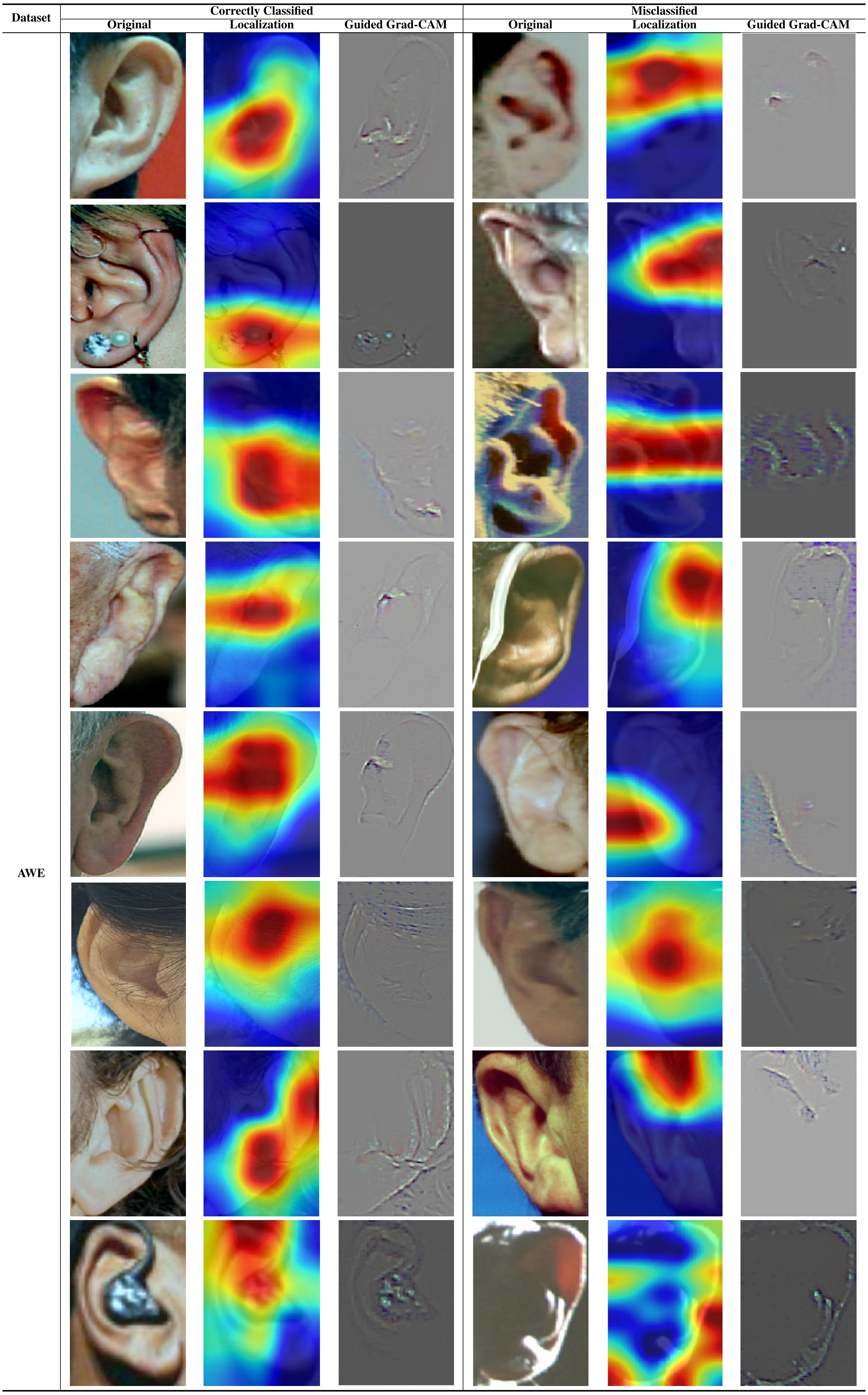
Dataset	Correctly Classified			Misclassified		
	Original	Localization	Guided Grad-CAM	Original	Localization	Guided Grad-CAM
WPUT						
						
						
						
						
						
						
						

TABLE 6. Guided Grad-CAM visualizations from the AWE ear dataset.



to make their predictions. We attribute the reason for this to the consistent occlusions throughout ear images for a specific subject. In the Grad-CAM localization and the Guided Grad-CAM visualizations one can observe how accessories like ear rings largely influence the decision of our networks. They can add some distinctive shapes and textures to identify the subjects, even when the ear structure is partly occluded. The impact of occlusions on performance is little for images showing mild degrees of head rotations or viewing angles. Contrarily, severe head rotations and hair occlusions have a detrimental effect on recognition performance as well as on localizing discriminative ear regions as can be seen from the misclassified images in Table 5. For example, the first two example images on the right hand side exhibit extreme viewing angles. Here, distinctive ear features like the shape of the helix, the intertragal notch and the triangular fossa are not visible or strongly deformed by an extreme perspective. This is in contrast to the mild degrees of head rotation on the left side of the table with correctly classified images, where information about the shape of the ear is limited, but still present. This issue could be mitigated through extra alignment steps in the recognition pipeline to recompense for the severe head rotations.

Table 6 illustrates examples of correctly classified and misclassified ear images from the AWE dataset. The correct model predictions tend to consider the geometrical shape of the ears even in the presence of accessories and ear plugs as can be seen in Table 6. Nevertheless, one can see that sometimes the networks focus on details in the geometrical structure of the ear or accessories like ear rings. In addition, some images suffer from distortions due to extreme viewing angles, bad lighting conditions or poor resolution. We think that the difficulty to access the entire ear structure in some training images creates an incentive to identify the subjects based on local details or hair texture. Nevertheless, low contrast images and severe head rotations can impact the recognition performance negatively. On the right hand side of Table 6 one can study some of the misclassified ear images. The wide range of ear image variations in the AWE dataset makes it difficult for the models to learn consistent and robust features, which can be justified from the provided visualizations. There is not much consistency in the image regions that are used for the wrong predictions. Here, we show cases where the misleading image region is the upper tip of the ear, the lower end of the scapha, some part of the helix, a strong contrast between the ear and the dark background or a single patch of hair.

VIII. CONCLUSION

This paper introduces ear recognition models based on five different variants of deep ResNet architectures. We proposed three methods of transfer learning, which can be used with other deep CNN architectures to learn discriminative ear features and to improve the overall recognition performance. Extensive experiments were conducted on four challenging publicly available ear image datasets, which consist of images collected under constrained and unconstrained conditions. In order to address the wide variability in ear images such as geometric transformations, occlusions, different image sizes and varying aspect ratios for each of the considered datasets, we proposed to embed each image into a fixed-size canvas to preserve the aspect ratios. Moreover, when training the models we introduced two different data augmentation pipelines to suit the type of variations in both, the constrained and unconstrained ear datasets. Our experimental results show considerable improvements in the recognition rates on all datasets and our proposed models achieve state-of-the-art recognition performance.

In order to make our models more transparent and to uncover the black-box nature of the deep models we applied a visualization technique that highlighted the important image regions responsible for the model predictions. The provided visualizations indicate that consistently focusing on the geometrical structure of the ear shape is the most discriminative region for getting correct predictions, whereas relying on auxiliary information such as the haircut or skin texture can result in wrong decisions. The visualizations also show the limited impact of partial ear occlusions and mild degrees of head rotations on performance, whereas, severe occlusion by hair and severe head rotations have a detrimental impact on recognition performance.

REFERENCES

[1] K. Chang, K. W. Bowyer, S. Sarkar, and B. Victor, "Comparison and combination of ear and face images in appearance-based biometrics," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 8, pp. 1160–1165, Sep. 2003.

[2] H. Nejati, L. Zhang, T. Sim, E. Martinez-Marroquin, and G. Dong, "Wonder ears: Identification of identical twins from ear images," in *Proc. 21st Int. Conf. Pattern Recognit. (ICPR)*, Nov. 2012, pp. 1201–1204.

[3] D. Meng, M. S. Nixon, and S. Mahmoodi, "On distinctiveness and symmetry in ear biometrics," *IEEE Trans. Biometrics, Behav., Identity Sci.*, vol. 3, no. 2, pp. 155–165, Apr. 2021.

[4] L. Olanrewaju, O. Oyebiyi, S. Misra, R. Maskeliunas, and R. Damasevicius, "Secure ear biometrics using circular kernel principal component analysis, Chebyshev transform hashing and Bose–Chaudhuri–Hocquenghem error-correcting codes," *Signal, Image Video Process.*, vol. 14, no. 5, pp. 847–855, 2020.

[5] Ž. Emeršič, V. Štruc, and P. Peer, "Ear recognition: More than a survey," *Neurocomputing*, vol. 255, pp. 26–39, Sep. 2017.

[6] B. Zhang, Z. Mu, J. Chen, and J. Dong, "A robust algorithm for ear recognition under partial occlusion," in *Proc. 32nd Chin. Control Conf.*, Jul. 2013, pp. 3800–3804.

[7] S. El-Naggar and T. Bourlai, "Evaluation of deep learning models for ear recognition against image distortions," in *Proc. Eur. Intell. Secur. Informat. Conf. (EISIC)*, Nov. 2019, pp. 85–93.

[8] U. Kacar and M. Kirci, "ScoreNet: Deep cascade score level fusion for unconstrained ear recognition," *IET Biometrics*, vol. 8, no. 2, pp. 109–120, Mar. 2019.

[9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[10] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

[11] Z. Wang, J. Yang, and Y. Zhu, "Review of ear biometrics," *Arch. Comput. Methods Eng.*, vol. 28, no. 1, pp. 149–180, 2021.

[12] A. Kamboj, R. Rani, and A. Nigam, "A comprehensive survey and deep learning-based approach for human recognition using ear biometric," *Vis. Comput.*, to be published, doi: 10.1007/s00371-021-02119-0.

[13] A. Benzaoui, A. Hadid, and A. Boukrouche, "Ear biometric recognition using local texture descriptors," *J. Electron. Imag.*, vol. 23, no. 5, Sep. 2014, Art. no. 053008.

[14] A. Kumar and C. Wu, "Automated human identification using ear imaging," *Pattern Recognit.*, vol. 45, no. 3, pp. 956–968, 2012.

[15] Z. Mu. (2009). *USTB Ear Database*. [Online]. Available: http://www1.ustb.edu.cn/resb/en/index.htm

[16] F. N. Sibai, A. Nuaimi, A. Maamari, and R. Kuwair, "Ear recognition with feed-forward artificial neural networks," *Neural Comput. Appl.*, vol. 23, no. 5, pp. 1265–1273, 2013.

[17] S. Sajadi and A. Fathi, "Genetic algorithm based local and global spectral features extraction for ear recognition," *Expert Syst. Appl.*, vol. 159, Nov. 2020, Art. no. 113639.

[18] A. Fathi, P. Alirezazadeh, and F. Abdali-Mohammadi, "A new global-Gabor-Zernike feature descriptor and its application to face recognition," *J. Vis. Commun. Image Represent.*, vol. 38, pp. 65–72, Jul. 2016.

[19] V. Ojansivu and J. Heikkilä, "Blur insensitive texture classification using local phase quantization," in *Proc. Int. Conf. Image Signal Process.*, 2008, pp. 236–243.

[20] J. D. Bustard and M. S. Nixon, "Toward unconstrained ear recognition from two-dimensional images," *IEEE Trans. Syst., Man, Cybern., A, Syst. Humans*, vol. 40, no. 3, pp. 486–494, May 2010.

[21] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[22] Ž. Emeršič, D. Stepec, V. Struc, P. Peer, A. George, A. Ahmad, E. Omar, T. E. Boult, R. Safdaii, Y. Zhou, S. Zafeiriou, D. Yaman, F. I. Eyiokur, and H. K. Ekenel, "The unconstrained ear recognition challenge," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, Oct. 2017, pp. 715–724.

[23] Ž. Emeršič, A. K. SV, B. S. Harish, W. Gutfeter, J. N. Khiarak, A. Pacut, and E. Hansley, "The unconstrained ear recognition challenge 2019," in *Proc. Int. Conf. Biometrics (ICB)*, Jun. 2019, pp. 1–15.

[24] F. I. Eyiokur, D. Yaman, and H. K. Ekenel, "Domain adaptation for ear recognition using deep convolutional neural networks," *IET Biometrics*, vol. 7, no. 3, pp. 199–206, May 2018.

[25] E. E. Hansley, M. P. Segundo, and S. Sarkar, "Employing fusion of learned and handcrafted features for unconstrained ear recognition," *IET Biometrics*, vol. 7, no. 3, pp. 215–223, May 2018.

[26] S. Dodge, J. Mounsef, and L. Karam, "Unconstrained ear recognition using deep neural networks," *IET Biometrics*, vol. 7, no. 3, pp. 207–214, May 2018.

[27] Ž. Emeršič, B. Meden, P. Peer, and V. Štruc, "Evaluation and analysis of ear recognition models: Performance, complexity and resource requirements," *Neural Comput. Appl.*, vol. 32, no. 20, pp. 15785–15800, Oct. 2020.

[28] H. Alshazly, C. Linse, E. Barth, and T. Martinetz, "Deep convolutional neural networks for unconstrained ear recognition," *IEEE Access*, vol. 8, pp. 170295–170310, 2020.

[29] V. T. Hoang, "EarVN1.0: A new large-scale ear images dataset in the wild," *Data Brief*, vol. 27, Dec. 2019, Art. no. 104630.

[30] P. P. Sarangi, D. R. Nayak, M. Panda, and B. Majhi, "A feature-level fusion based improved multimodal biometric recognition system using ear and profile face," *J. Ambient Intell. Humanized Comput.*, to be published, doi: 10.1007/s12652-021-02952-0.

[31] T. Jabid, M. H. Kabir, and O. Chae, "Robust facial expression recognition based on local directional pattern," *ETRI J.*, vol. 32, no. 5, pp. 784–794, 2010.

[32] P. Yan and K. W. Bowyer, "Biometric recognition using 3D ear shape," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 8, pp. 1297–1308, Aug. 2007.

[33] S. Xie, R. Girshick, P. Dollar, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1492–1500.

[34] 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. 4700–4708.

[35] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Training very deep networks," in *Int. Conf. Neural Inf. Process. Syst.*, 2015, pp. 2377–2385.

[36] K. He and J. Sun, "Convolutional neural networks at constrained time cost," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 5353–5360.

[37] 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.

[38] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *2015*, pp. 448–456.

[39] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. Int. Conf. Mach. Learn.*, 2010, pp. 807–814.

[40] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A comprehensive survey on transfer learning," *Proc. IEEE*, vol. 109, no. 1, pp. 43–76, Jan. 2021.

[41] H. Alshazly, C. Linse, M. Abdalla, E. Barth, and T. Martinetz, "COVID-nets: Deep CNN architectures for detecting COVID-19 using chest CT scans," *PeerJ Comput. Sci.*, vol. 7, p. e655, Jul. 2021.

[42] S. Kornblith, J. Shlens, and Q. V. Le, "Do better ImageNet models transfer better?" in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2661–2671.

[43] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 3320–3328.

[44] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: An astounding baseline for recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2014, pp. 806–813.

[45] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[46] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge, U.K.: Cambridge Univ. Press, 2000.

[47] H. Alshazly, C. Linse, E. Barth, and T. Martinetz, "Explainable COVID-19 detection using chest CT scans and deep learning," *Sensors*, vol. 21, no. 2, p. 455, Jan. 2021.

[48] H. Kaushik, D. Singh, M. Kaur, H. Alshazly, A. Zaguia, and H. Hamam, "Diabetic retinopathy diagnosis from fundus images using stacked generalization of deep models," *IEEE Access*, vol. 9, pp. 108276–108292, 2021.

[49] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 6405–6416.

[50] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[51] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.

[52] E. Lobacheva, N. Chirkova, M. Kodryan, and D. P. Vetrov, "On power laws in deep ensembles," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 2375–2385.

[53] J. Zhao, X. Xie, X. Xu, and S. Sun, "Multi-view learning overview: Recent progress and new challenges," *Inf. Fusion*, vol. 38, pp. 43–54, Nov. 2017.

[54] P. Hu, X. Peng, H. Zhu, J. Lin, L. Zhen, and D. Peng, "Joint versus independent multiview hashing for cross-view retrieval," *IEEE Trans. Cybern.*, early access, Oct. 29, 2020, doi: 10.1109/TCYB.2020.3027614.

[55] X. Peng, H. Zhu, J. Feng, C. Shen, H. Zhang, and J. T. Zhou, "Deep clustering with sample-assignment invariance prior," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 11, pp. 4857–4868, Nov. 2020.

[56] E. González-Sánchez, "Biometria de la oreja," Ph.D. dissertation, Dept. Comput. Sci., Universidad de Las Palmas de Gran Canaria, Madrid, Spain, 2008.

[57] H. Alshazly, C. Linse, E. Barth, and T. Martinetz, "Ensembles of deep learning models and transfer learning for ear recognition," *Sensors*, vol. 19, no. 19, p. 4139, Sep. 2019.

[58] D. Frejlichowski and N. Tyszkiewicz, "The West Pomeranian University of technology ear database–A tool for testing biometric algorithms," in *Proc. Int. Conf. Image Anal. Recognit.*, 2010, pp. 227–234.

[59] R. Raghavendra, K. B. Raja, and C. Busch, "Ear recognition after ear lobe surgery: A preliminary study," in *Proc. IEEE Int. Conf. Identity, Secur. Behav. Anal. (ISBA)*, Feb. 2016, pp. 1–6.

[60] H. A. Alshazly, M. Hassaballah, M. Ahmed, and A. A. Ali, "Ear biometric recognition using gradient-based feature descriptors," in *Proc. Int. Conf. Adv. Intell. Syst. Inform.* Springer, 2018, pp. 435–445.

[61] D. P. Chowdhury, S. Bakshi, G. Guo, and P. K. Sa, "On applicability of tunable filter bank based feature for ear biometrics: A study from constrained to unconstrained," *J. Med. Syst.*, vol. 42, no. 1, p. 11, Jan. 2018.

[62] M. Hassaballah, H. A. Alshazly, and A. A. Ali, "Ear recognition using local binary patterns: A comparative experimental study," *Expert Syst. Appl.*, vol. 118, pp. 182–200, Mar. 2019.

[63] Y. Zhang, Z. Mu, L. Yuan, and C. Yu, "Ear verification under uncontrolled conditions with convolutional neural networks," *IET Biometrics*, vol. 7, no. 3, pp. 185–198, May 2018.

[64] M. Sultana, P. Polash Paul, and M. Gavrilova, "Occlusion detection and index-based ear recognition," *J. WSCG*, vol. 23, no. 2, pp. 121–129, 2015.

[65] Z. Emersic, D. Stepec, V. Struc, and P. Peer, "Training convolutional neural networks with limited training data for ear recognition in the wild," in *Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2017, pp. 988–994.

[66] H. Alshazly, C. Linse, E. Barth, and T. Martinetz, "Handcrafted versus CNN features for ear recognition," *Symmetry*, vol. 11, no. 12, p. 1493, Dec. 2019.

[67] I. Omara, A. Hagag, G. Ma, F. E. Abd El-Samie, and E. Song, "A novel approach for ear recognition: Learning Mahalanobis distance features from deep CNNs," *Mach. Vis. Appl.*, vol. 32, no. 1, pp. 1–14, Jan. 2021.

[68] J. Zhang, W. Yu, X. Yang, and F. Deng, "Few-shot learning for ear recognition," in *Proc. Int. Conf. Image, Video Signal Process.*, 2019, pp. 50–54.

[69] I. Omara, A. Hagag, S. Chaib, G. Ma, F. E. Abd El-Samie, and E. Song, "A hybrid model combining learning distance metric and DAG support vector machine for multimodal biometric recognition," *IEEE Access*, vol. 9, pp. 4784–4796, 2021.

[70] Y. Khaldi and A. Benzaoui, "A new framework for grayscale ear images recognition using generative adversarial networks under unconstrained conditions," *Evolving Syst.*, to be published, doi: 10.1007/s12530-020-09346-1.

[71] M. Hassaballah, H. A. Alshazly, and A. A. Ali, "Robust local oriented patterns for ear recognition," *Multimedia Tools Appl.*, vol. 79, nos. 41–42, pp. 31183–31204, Nov. 2020.

[72] R. A. Priyadharshini, S. Arivazhagan, and M. Arun, "A deep learning approach for person identification using ear biometrics," *Appl. Intell.*, vol. 51, no. 4, pp. 2161–2172, 2020.

[73] Y. Khaldi and A. Benzaoui, "Region of interest synthesis using image-to-image translation for ear recognition," in *Proc. Int. Conf. Adv. Aspects Softw. Eng. (ICAASE)*, Nov. 2020, pp. 1–6.

[74] Y. Khaldi, A. Benzaoui, A. Ouahabi, S. Jacques, and A. Taleb-Ahmed, "Ear recognition based on deep unsupervised active learning," *IEEE Sensors J.*, early access, Jul. 26, 2021, doi: 10.1109/JSEN.2021.3100151.

[75] J. Tobias Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," 2014, *arXiv:1412.6806*. [Online]. Available: http://arxiv.org/abs/1412.6806

HAMMAM ALSHAZLY received the B.Sc. degree in computer science from South Valley University, Egypt, in 2006, the M.Sc. degree in computer science from the University of Mumbai, India, through a scholarship from the Indian Council for Cultural Relations (ICCR), in 2014, and the Ph.D. degree in computer science from South Valley University, in 2018. From February 2019 to January 2021, he was a Postdoctoral Researcher with the Institute for Neuro- and Bioinformatics, University of Lübeck, Germany. He is currently working as an Assistant Professor with the Department of Computer Science, Faculty of Computers and Information, South Valley University. He has published articles in conferences and peer-reviewed journals. His research interests include deep learning, biometrics, computer vision, machine learning, and artificial intelligence. He was awarded the Partnership and Ownership (ParOwn) Initiative in 2010 for a period of six months from Monash University, Australia. During his Ph.D. degree, he was awarded a Fulbright Scholarship for ten months to complete part of his research work at the University of Kansas, Lawrence, KS, USA. He serves as a reviewer for several journals.

CHRISTOPH LINSE received the B.Sc. degree in physics from the Rheinische Friedrich–Wilhelms-Universität Bonn, Germany, in 2014, the M.Sc. degree in physics from the Norwegian University of Science and Technology, Norway, through a scholarship from Cusanuswerk, in 2016, and the second M.Sc. degree in computational science from the Institute for Neuro- and Bioinformatics, University of Lübeck, Germany, where he is currently pursuing the Ph.D. degree in computational neurosciences. He is currently enrolled at the county Land Schleswig Holstein to qualify small and medium-sized businesses in Schleswig Holstein to digitalize their value added chain introducing state-of-the-art techniques from machine learning and deep learning. His research interests include computational neurosciences, machine learning, and deep learning.

ERHARDT BARTH (Member, IEEE) received the Ph.D. degree in electrical engineering from the Technical University of Munich, Munich, in 1994. He was a Research Associate with the Department of Communications Engineering, Technical University of Munich, and a Visiting Fellow with the Department of Computer Science, Melbourne University, Australia, where he was supported by the Gottlieb–Daimler and Karl–Benz Foundation. Then, he was a Researcher with the Department of Medical Psychology, University of Munich, and a Klaus-Piltz Fellow at the Institute for Advanced Study, Berlin. From 1997 to 1998, he was a member of the NASA Vision Science and Technology Group, NASA Ames, Moffet Field, Mountain View, CA, USA. He is currently a Professor of computer science and the Deputy Director of the Institute for Neuro- and Bioinformatics, University of Lübeck, Germany. He also leads the research on human and machine vision at the Institute for Neuro- and Bioinformatics. His research interests include computer vision and machine learning. In May 2000, he received a Schloessmann Award from the Max-Planck Gesellschaft.

SAHAR AHMED IDRIS received the B.Sc. and M.Sc. degrees in mathematical sciences from the Faculty of Mathematical Sciences, University of Khartoum, and the Ph.D. degree in mathematics from the Faculty of Science, Sudan University of Science and Technology, in 2016. She is currently working as an Assistant Professor with the College of Industrial Engineering, King Khalid University, Abha, Saudi Arabia. Her research interests include mathematics, digital image processing, and meta-heuristic algorithms.

THOMAS MARTINETZ (Senior Member, IEEE) received the degree in physics from the Technical University of Munich, Germany, and the Ph.D. degree in theoretical biophysics from the Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana–Champaign, Champaign, IL, USA. From 1991 to 1996, he led the project Neural Networks for automation control at Corporate Research Laboratories, Siemens AG, Munich. From 1996 to 1999, he was a Professor of neural computation with Ruhr-University of Bochum and the Head of the Center for Neuroinformatics. From 2006 to 2008, he was the Vice-Rector of the University of Lübeck, Germany. From 2008 to 2011, he was the Vice-President of Research and Technology Transfer. Since 2013, he has been a Chairman of the Senate, University of Lübeck. He was co-founded the software companies: Consideo, the Pattern Recognition Company, and Gestigon. He is currently a Professor of computer science and the Director of the Institute for Neuro- and Bioinformatics, University of Lübeck. He has published more than 300 research articles in peer-reviewed journals and international conferences. His research interests include neural networks, machine learning, and artificial intelligence.