Journal of Robotics and Control (JRC)
Volume 5, Issue 1, 2024
ISSN: 2715-5072, DOI: 10.18196/jrc.v5i1.20135

Leveraging a Two-Level Attention Mechanism for Deep Face Recognition with Siamese One-Shot Learning

Arkan Mahmood Albayati [1*], Wael Chtourou [2], Faouzi Zarai[3]
[1,2,3]ENET'com Sfax, University of Sfax, Sfax, Tunisia
Email: [1]albayatiarkan39@gmail.com, [2]chtourou.wael@gmail.com, [3]faouzifbz@gmail.com
*Corresponding Author

***Abstract*—Discriminative feature embedding is used for large-scale facial recognition. Many image-based facial recognition networks use CNNs like ResNets and VGG-nets. Humans prioritise different elements, but CNNs treat all facial pictures equally. NLP and computer vision use attention to learn the most important part of an input signal. The inter-channel and inter-spatial attention mechanism is used to assess face image component significance in this study. Channel scalars are calculated using Global Average Pooling in face recognition channel attention. A recent study found that GAP encodes low-frequency channel information first. We compressed channels using discrete cosine transform (DCT) instead of scalar representation to evaluate information at frequencies other than the lowest frequency for the channel attention mechanism. Later layers can acquire the feature map after spatial attention. Channel and spatial attention increase CNN facial recognition feature extraction. Channel-only, spatial-only, parallel, sequential, or channel-after-spatial attention blocks exist. Current face recognition attention approaches may be outperformed on public datasets (Labelled Faces in the Wild).**

Keywords*—One-shot; Siamese Network; Triplet Loss; Contrastive Loss; Attention.

I. Introduction

In recent past years, Convolutional Neural Networks (CNNs) have demonstrated their efficiency in various tasks by enhancing the state-of-the-art performance on such tasks [1]–[5]. More specifically, the performance of face recognition has been boosted to an unprecedented level due to the availability of novel discriminative learning methods [6]–[11] and advanced deep learning architectures [1], [4], [5], [12]. Face recognition consists of [13], [14]:

- Verification: It consists of taking two images and determining if they belong to the same person or to two persons differently, which is a binary classification problem.
- Identification: It consists of determining, among galleries, to which gallery member an image sample belongs.

Face recognition is a crucial tool in real-world scenarios. It is advantageous compared to other biometric systems (such as fingerprint, and iris) due to its non-intrusive feature and can also be performed at a distance [15]–[17]. Face recognition can be used to enforce the law. It is also helpful for the police to quickly narrow down possible suspects by automatically retrieving suspect images from police mug-shot databases.

Face recognition's main objective is to look for a discriminative feature in order to meet the following criterion: Under a suitably chosen metric space, the minimal inter-class distance is greater than the maximal intra-class distance. Several research works have been carried out in this context to obtain discriminative features. In particular, Sun et al. [18] proposed a Deep face IDentification-verification approach using contrastive loss to get efficient and effective feature representations that enlarge inter-personal variations while reducing intra-personal differences. In [19], Wen et al. proposed a center loss to improve the discriminative power of the deeply learned features. The proposed center loss learns at the same time a center for deep features of each class and penalizes the distances between the deep features and their corresponding class centers. The authors have demonstrated that, by jointly using the softmax activation function and their proposed center loss, inter-class dispensation and intra-class compactness can be achieved as much as possible. Schroff et al. [20] used a triplet loss to perform a face recognition task with the goal of enlarging inter-class distance while reducing intra-class distance. Moreover, angular margin-based loss functions have recently demonstrated their power and achieved state-of-the-art results in face recognition. For instance, Liu et al. [8], aiming to learn angularly discriminative features, proposed an angular softmax (A-Softmax) loss function for CNNs. In [6], the authors demonstrated that the softmax activation function lacks the power of discrimination in face recognition and proposed a large margin cosine loss (LMCL) to achieve minimum intra-class variance and maximum inter-class variance through virtue normalization and cosine decision margin maximization. Deng et al. [10] investigated the issues of the above angular loss functions and proposed an Additive Angular Margin Loss (ArcFace) to obtain highly discriminative features for face recognition.

Wang et al. [21] proposed a loss function called support vector guided softmax loss (SV-Softmax) for discriminative feature learning guidance.

In addition to the loss functions mentioned above aiming to improve the performance of face recognition systems, attention mechanisms have been an active research area in the past years to improve deep learning models going from NLP to computer vision, and face recognition models have been improved with such mechanisms. In face recognition, most of the deep learning models use CNN architectures like ResNet and VGG., as their backbones. For instance, Liu et al. [8] and Wang et al. [6] have adopted Residual-Style architecture with 64 layers as the feature extractor. In [10], Deng et al. adopted ResNet-50 and ResNet-101 [1] as the backbone of their architecture. Based on human cognition, while different parts of face images have different importance in face recognition, the standard CNNs consider equally different parts of face images. In addition, the feature map generated by standard CNNs has many channels, possibly leading to information redundancy among channels. This situation can also make the feature have a risk of over-fitting even if one uses some regularization methods like dropout or $l1$ and $l2$ norm. The most important parts of an input signal can be learned through the attention mechanisms which are widely used in different areas of deep learning and achieved great success in NLP [22]–[25], images classification [2], [26]–[29], images segmentation [30], [30], [31], saliency detections [32], [33], etc. Researchers have also used attention mechanisms to enhance network performance to address the above-mentioned challenges in face recognition. More specifically, Ling et al. [34] adopted channel attention and spatial attention in their face recognition model. Additionally, Rao et al. [35] proposed an attention-aware deep reinforcement learning (ADRL) method for video face recognition. Their proposal discards the misleading and confounding frames and generates a compact, refined feature for face recognition in face videos. For channel attention, the current channel attention mechanisms used in face recognition [34]–[37] employ a scalar to represent each channel through Global Average Pooling (GAP). However, GAP struggles to capture complex information for various inputs because of its simplicity. In contrast to the previous works, we consider the scalar representation of a channel as a compression problem in this paper and used discrete cosine transform (DCT) to compress the channels for the channel attention mechanism. Then, spatial attention is applied and the resulting feature map can be fed into subsequent layers. Therefore, this two-level (channel and spatial) attention module can make the standard CNNs(ResNets, VGGNets, etc.) more powerful in extracting discriminative features for face recognition tasks.

The goal of this paper is to train a deep neural network that will be able to take an image and return its identity. Therefore, it involves, as mentioned previously, face verification, where the model determines if two given images belong to the same person or two different persons, and face identification, where the model returns the identity of the person by using the gallery. It is worth noting that we suppose that the given image must contain one and only one face.

Deep learning (DL) methods often require several labeled samples per class to perform well on the task. However, in real-world scenarios, acquiring such a tremendous amount of data for training is challenging. Therefore, a model that can rely on one or a few images per person for training is preferred. One-shot learning was proposed to deal with such type of task. The training data consists of one or a few images per class in one-shot learning. The main contribution of this paper is summaries as follows.

- We exploit this advantage of one-shot with siamese-CNNs and propose a face recognition method
- In our approach, two-level attention-CNNs are used to extract rich features from the given images, and those features are fed to the classification network for one-shot learning

The rest of this paper is organized as follows. In the next section (II), we discuss the related work, while section IV provides the methodology and the proposed model. In section V, our experiment results are given, and the paper is concluded in section VI.

II. Related Work

This section provides recent works that have been carried out in the context of face recognition and approaches that use attention mechanisms in computer vision.

A. Face Recognition Methods

The accuracy of face recognition started improving in the early 2000s through engineered features, more specifically with Local binary patterns [38], [39]. However, when applied to unconstrained face databases, the performances of such feature engineering approaches are low. Since 2012 when AlexNet won ImageNet Large-Scale Visual Recognition Challenge (ILSVRC 2012) [4], deep learning approaches have been employed in several computer vision tasks such as image classification, image recognition, and so forth. Afterwards, more advanced deep learning architectures such as Inception [5], Resnet [1], VGG16,19 [40], and others [41] are employed for image classification tasks. More specifically, the architecture of these deep learning approaches consists of two main parts: The features extraction component, which extracts rich features from the raw inputs, and the classification or recognition network, which uses features extracted from the first component and performs the task on hand.

When it comes to face recognition using deep learning, Hu et al. [42] provided a good introduction to the convolutional

neural networks employed for face recognition. They especially investigated the advantages and drawbacks of CNNs and provided a useful developmental roadmap. Albeit deep Learning methods provide excellent performance compared to features engineering methods, they struggle in real-world use cases that require learning with a small amount of data, class imbalance and adjusting to a constant inflow of new class information. Therefore, one-shot learning was proposed recently to deal with this problem. Guo and Zhang [43] provided a good survey on face recognition models based on deep learning.

Sun et al. [18] demonstrated that by using both face identification and verification signals as supervision, intra-personal variations could be reduced while enlarging inter-personal distances. In their proposal, while the identification task enlarges the inter-class distances, the verification task reduces the intra-class variations by pulling together the features extracted from the same identity.

The authors of [44] derived a face representation by employing explicit 3D face modeling to apply a piecewise affine transformation, improving the face recognition model's performance.

In [45], aiming to address the class imbalance issue, Guo and Zhang proposed a one-shot face recognition model by aligning the norms of weight vectors of underrepresented classes and normal classes, thus giving the one-shot classes an equal weight.

To deal with deficient training samples in face recognition, Wang et al. [46] proposed a framework that leverages CNNs of balancing regularize and shifting center regeneration which regulates norms of weight vector into the same scale and adjusts clustering center.

Ding et al. [47] proposed a generative learning approach to synthesize meaningful data for one-shot classes by adapting the data variances from other normal classes.

Jadhav et al. [48] employed a deep attribute representation of faces to address the problem of one-shot unconstrained face recognition. They used specific attributes of human faces, like hair and gender, to fine-tune a deep CNN for face recognition.

In [49], Wu et al. proposed a hybrid method for face recognition tasks when a CNN and the nearest neighbor (NN) model are combined.

In addition, Hong et al. [50] proposed a domain adaptation network for one-shot tasks. They employed domain-adversarial training by generating synthetic images in various poses using a 3D face model.

Cheng et al. [51], aiming to produce a better representation for low-shot learning, proposed an Enforced Softmax optimization approach built upon CNNs. The proposed model leverages optimal dropout, selective attenuation, normalization, and model-level optimization.

Cao et al. [52] proposed a dataset called VGGface2 and used ResNet-50 as the network backbone structure to train the face recognition model using the proposed dataset. In the same way, [6], [8], and [53] used ResNets as the backbone structure of their network.

When it comes to the siamese network, it was initially proposed by Bromley et al. [54] for the signature verification task. Further, it was used for image recognition by [55].

Song et al. [56] proposed a mask learning technique which learns how to discard occlusion features and only uses the non-occluded facial areas for face recognition with the siamese network.

B. Attention-based Methods

The attention mechanism, often combined with the gating function and sequential approaches, was first proposed for NLP tasks to learn the most informative parts of an input signal. In recent years, the attention mechanism has been widely used for image classification and other areas of DL. More specifically, Wang et al. [28] proposed a residual attention network that is built by stacking attention modules to produce attention-aware features. The proposed attention module was incorporated into a deep residual network, which performs well on noisy input signals.

The convolutional features' inter-channel relationship was learned using a squeeze-and-excitation module through global average pooling (GAP), proposed by [2]. This module can be used in CNNs to improve the feature extraction process.

Woo et al. [57] proposed a Convolutional Block Attention Module (CBAM) for feed-forward CNNs. The proposed module sequentially infers channel and spatial attention, then these attention maps are matrices multiplied by the input feature map to adaptively refine the feature map. Our proposed model follows a similar pattern except that instead of using the GAP for channel attention, we adopt DCT, which is widely used image compression.

Wang et al. [58] proposed a non-local neural network leverage self-attention model of [59] for the video classification task.

Zhang et al. [60] proposed a self-attention-based module for better image generation which finds global dependencies within internal representations.

Even though current deep learning approaches achieved high performance in face recognition systems, the standard CNNs used as backbones consider all parts of face images equally. Secondly, because of the deepening of current architectures, the huge number of channels generated in the convolution layers introduce the information redundancy phenomenon. We believe that the model performance of face recognition can still be improved with a minor change in network architecture, even with a tiny increase in the number of learnable parameters and the computational cost. We achieve the performance improvement

of face recognition through the channel and spatial attention employed sequentially. In contrast to [34], we consider channel attention as a compression task and employ DCT instead of GAP to compress channels in a scalar.

III. Preliminaries

This section provides an overview of the main components we rely on for our deep face model.

A. Siamese Network

The siamese network is a deep neural network architecture that operates on two identical sub-networks sharing the same weights while taking as inputs two different raw data vectors and are joined after features encoded by a similarity or comparative function. Fig. 1 illustrates the Siamese network. More specifically, these two sub-networks aim to calculate the similarity between the twin raw input faces. These sub-networks are identical and share the same weights, hence the term "Siamese". In this paper, the convolutional Siamese

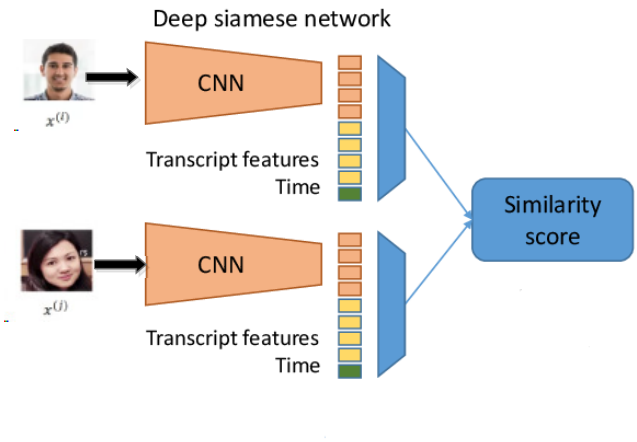

Fig. 1. Architecture of Siamese Network

network is employed to extract features given a few samples of a distribution. This model provides mainly two advantages such as (1) Fewer learnable parameters as the weights are shared (2) a Lower tendency of over-fitting. In the training phase, using a few data samples, the network takes as input a pair of faces and tries to learn how to discriminate these two images based on their labels and their features by generating probability scores. Such probability scores enable us to determine whether the pairs belong to the same person or are different.

1) Siamese Network Loss Function: This paper calculates the model error for the Siamese network during the training phase using the contrastive loss function. It is a distance-based loss as opposed to more conventional error-prediction losses. Contrastive loss is used to learn embeddings where two similar images have a low Euclidean distance, and two dissimilar images have a large Euclidean distance. This loss is expressed in Equation (1).

$$Loss(x_1^i, x_2^i) = (Y(x_1^i, x_2^i))(-log(Y_{pred}(x_1^i, x_2^i)) + (1 - Y(x_1^i, x_2^i))(-log(1 - Y_{pred}(x_1^i, x_2^i)) \quad (1)$$

Where i denotes the i^{th} index of the current batch, $Y(x_1^i, x_2^i)$ is a vector of true labels and $Y_{pre}(x_1^i, x_2^i)$ is a vector of predicted labels.

B. Discrete Cosine Transform (DCT)

Introduced by [61], 2D DCT is given in Equation (2).

$$B_{h,w}^{i,j} = cos(\frac{\pi h}{H}(i + \frac{1}{2}))cos(\frac{\pi w}{W}(j + \frac{1}{2})) \quad (2)$$

Then, it can be written as given in Equation (3).

$$f_{h,w}^{2d} = \sum_{i=0}^{H-1} \sum_{j=0}^{W-1} x_{i,j}^{2d} B_{h,w}^{i,j} \quad (3)$$

Where $i \in \{0, 1, ..., H-1\}$, $j \in \{0, 1, ..., W-1\}$ $f^{2d} \in \mathbb{R}^{H \times W}$ is the 2D DCT frequency spectrum, $x^{2d} \in \mathbb{R}^{H \times W}$ is the input signal while H and W are respectively the height and width of x^{2d}. Accordingly, the inverse is given in Equation (4).

$$x_{h,w}^{2d} = \sum_{h=0}^{H-1} \sum_{w=0}^{W-1} f_{h,w}^{2d} B_{h,w}^{i,j} \quad (4)$$

C. Neural Architecture Search

Neural Architecture Search (NAS) is the process of automating the design of neural networks' topology in order to achieve the best performance on a specific task. The goal is to design the architecture using limited resources and with minimal human intervention. It is a search algorithm with the following features:

- It operates on the search space of possible network topologies, consisting of predefined operations (e.g.convolutional layers, recurrent, pooling, fully connected, etc.) and their connections.
- A controller then chooses a list of possible candidate architectures from the search space.
- The candidate architectures are trained and ranked based on their performance on the validation test.
- The ranking is used to readjust the search and obtain new candidates.
- The process iterates until it reaches a certain condition and provides the optimal architecture.
- The optimal architecture is evaluated on the test set.

For more details about NAS, readers can refer to [62], [63] and [64]. In this paper, we use NAS to search for the best frequency component for channels in channel attention.

IV. Proposed Model

The details of our proposed architecture are presented in this section. Firstly, the general framework is discussed, then, the channel and spatial attention blocks are highlighted, and the mechanism which aggregates both blocks is detailed.

A. Architecture Overview

We proposed to employ both channel and spatial attention to learn the most important parts of face images in face recognition and mitigate the redundancy of information caused by standard CNNs because of the huge number of channels generated. When combined, These attention modules will enable our network to learn adaptively the inter-channel and inter-spatial relationships. Therefore, the global feature relationship of input face images will be learned, and a more discriminative feature will be obtained for face recognition. Fig. 2 presents both our channel and spatial attention blocks that we integrated into a ResNet-50 to obtain refined features for face recognition. More specifically, the channel attention block and the spatial attention block learn sequentially, the channel relationships matrix and spatial relationships matrix. Then, the refined feature is obtained through matrix multiplication. We will detail these attention blocks in the next subsections.

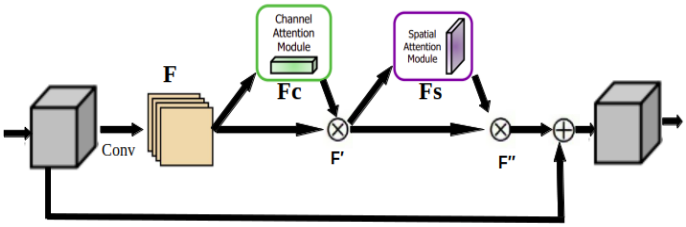

Fig. 2. Our Proposed Attention Modules with ResNet: We show in this figure, the exact position of our Channel and Spatial Attention in a ResNet Architecture

B. Channel Attention

This subsection details the channel attention block. This block produces a channel attention relationship matrix that models the inter-channel relationships of a given feature map. It focuses on "what" is meaningful given an input face image. Channel attention is widely used in computer vision tasks. More specifically, channel attention uses a scalar to represent and measure the importance of each channel of the feature map. For instance, assuming that $\mathbf{F} \in \mathbb{R}^{C\times H\times W}$ is the feature map of a convolution layer where C denotes the number of channels or feature detectors, H the height of each channel and W the width of each channel. Since the widely used scalar for channel attention represents the entire channel, it can be seen simply as a compression problem. Therefore, channel attention can be written as given in Equation (5).

$$att = sigmoid(fc(compress(\mathbf{F}))) \tag{5}$$

Where $att \in \mathbb{R}^C$ is the attention vector, $sigmoid$ and fc are, respectively, the sigmoid activation function and the mapping function like fully-connected layer or 1×1 convolution layer while $compress : \mathbb{R}^{C\times H\times W} \rightarrow \mathbb{R}^C$ is a compression technique. The current channel attention used in face recognition such as [34] uses the GAP as a compression technique. Global max pooling [57] and global standard deviation pooling [65] are other compression methods. After obtaining $\mathbb{R}^C$, each channel of $\mathbf{F}$ is scaled by the corresponding attention value as given in Equation (6).

$$\tilde{\mathbf{F}}_{:,i,:,:} = att_i \mathbf{F}_{:,i,:,:} \tag{6}$$

Where $i \in \{0, 1, 2, ..., C\}$, $\tilde{\mathbf{F}}$ denotes the output of the attention mechanism, att_i is the $i-th$ element of the attention vector and $\mathbf{F}_{:,i,:,:}$ is the $i-th$ channel of the feature map $\mathbf{F}$.

In this paper, instead of using GAP or the above-mentioned other methods as compression techniques, we use the DCT. In [66], Qin et al. have demonstrated that GAP is a weighted sum of inputs and proven that GAP is a special case of 2D DCT. Additionally, by replacing h and w by 0 in Equation (3), it becomes Equation (7):

$$\begin{aligned} f^{2d}_{\mathbf{0},\mathbf{0}} &= \sum_{i=0}^{H-1}\sum_{j=0}^{W-1} x^{2d}_{i,j} cos(\frac{\mathbf{0}}{H}(i+\frac{1}{2}))cos(\frac{\mathbf{0}}{W}(j+\frac{1}{2})) \\ &= \sum_{i=0}^{H-1}\sum_{j=0}^{W-1} x^{2d}_{i,j} \\ &= GAP(x^{2d})HW \end{aligned} \tag{7}$$

Where GAP means global average pooling. $f^{2d}_{\mathbf{0},\mathbf{0}}$ is the lowest frequency component of 2D DCT which corresponds to the GAP. It can be concluded that in the current face recognition approaches that employ the channel attention mechanism with GAP, only the lowest frequency information is preserved and other information of other frequencies is simply lost while they also encode some useful information to represent the channels and therefore need to be taken into account. Accordingly, by using 2D DCT, more information can be preserved by considering the information of other frequencies in addition to the lowest frequency component (a.k.a, GAP).

The DCT used in this paper for channel attention is described as follows: Given an input feature $\mathbf{F}$, we split, along the channel dimension C, this feature map $\mathbf{F}$ into $\{F^{(0)}, F^{(1)}, ..., F^{(n-1)}\}$ where $F^i \in \mathbb{R}^{C'\times H\times W}$, $i \in \{0, 1, ..., n-1\}$ with $C' = \frac{C}{n}$ and C is divisible by n. We assign to each F^i, its corresponding 2D DCT frequency component and the obtained results can be considered as the compression results of channel attention. More specifically, the equations of such a transformation are expressed in Equation (8).

$$\begin{aligned} Freq^i &= 2DDCT^{u_i,v_i} F^i \\ &= \sum_{h=0}^{H-1}\sum_{w=0}^{W-1} X^i_{:,h,w} B^{u_i,v_i}_{h,w} \end{aligned} \tag{8}$$

Where $Freq^i \in \mathbb{R}^{C'}$ is the resulting compression vector with dimension equal to C', $i \in \{0, 1, ..., n-1\}$ while $[u_i, v_i]$ are the frequency component 2D indices corresponding to F^i. The entire compression vector is given in Equation (9) by concatenation.

$$\begin{aligned} Freq &= compress(\mathbf{F}) \\ &= concat([Freq^{(0)}, Freq^{(1)}, ..., Freq^{(n-1)}]) \end{aligned} \tag{9}$$

Similarly to Equation (5), our channel attention is expressed in Equation (10).

$$dct_{att} = sigmoid(fc(Freq)) \tag{10}$$

From the above two last equations, it can be seen that the DCT is a general case of GAP resulting in more frequencies and, therefore, more information which is effectively enriched for representation. However, having several frequency component indices for each part brings a challenge of how to effectively choose $[u_i, v_i]$, which are the frequency component indices for each part. To solve this challenge, we used Neural Architecture Search (NAS) to search for the best frequency components. More specifically, to search for components, a set of continuous variables $\alpha = \left\{\alpha^{(u,v)}\right\}$ are assigned for each part $F^{(i)}$ which is written in Equation (11).

$$Freq^i_{nas} = \sum_{(u,v)\in O} \frac{exp(\alpha^{(u,v)})}{\sum_{(u',v')\in O} exp(\alpha^{u',v'})} 2DDCT^{u,v} F^i \tag{11}$$

Where O contains all 2D DCT frequency component indices. Once the search is performed, the frequency component of each part $F^{(i)}$ is expressed in Equation (12).

$$(u_i^*, v_i^*) = argmax_{(u,v)\in O} \left\{\alpha^{(u,v)}\right\} \tag{12}$$

The selected 2D DCT frequency components are summed and the resulting vector F_c is multiplied by the input feature map **F** and we obtain **F'** for the spatial attention block. The channel attention block is summarized in Fig. 3.

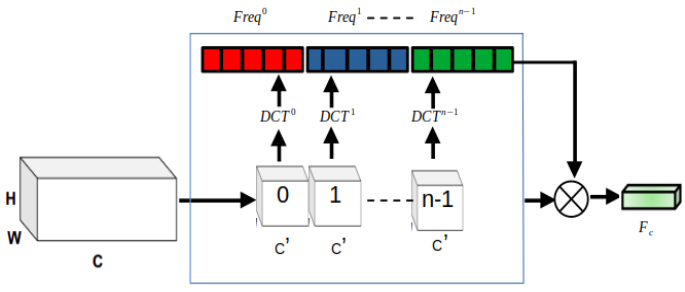

Fig. 3. Channel Attention Block

C. Spatial Attention

While channel attention focuses on "what", the spatial focuses on "where" the meaningful information is located given an input face image, which complements the channel attention. It generates a spatial attention matrix by using the inter-spatial relationship between features. In [67], Zagoruyko et al. demonstrated that the channel axis pooling operations highlight effectively informative regions. Therefore, to generate the spatial attention map, we first apply both average pooling and max pooling in parallel. Then, an efficient feature descriptor is obtained by concatenating the output of such operations. Finally, we apply a convolution layer to produce the spatial attention matrix, which encodes information on where to suppress or emphasize. This attention block is summarized in Equation (13).

$$F_s = sigmoid(f^{7\times7}([AvgPool(F'), MaxPool(F')])) \tag{13}$$

The visual detail of the spatial attention is presented in Fig. 4.

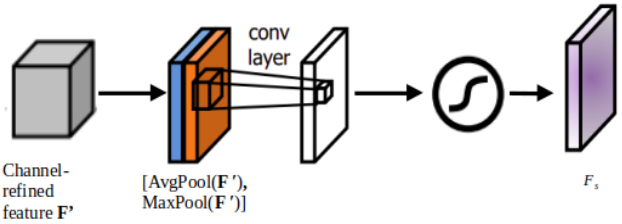

Fig. 4. Spatial Attention Block

D. Face Feature Embedding with Attention Mechanism

To understand our proposed attention blocks intuitively, we arrange them sequentially and plot them in Fig. 5. More specifically, for an intermediate feature X, a channel attention matrix is first generated through the channel attention module and matrix multiplication is applied to get the weighted feature. Then, the element-wise summation is applied to get a channel-refined feature. Sequentially, the spatial attention module generates the spatial attention matrix and matrix multiplication followed by element-wise summation, leading to a spatial-refined feature. We applied BatchNormalization to enable the network to converge fast, and the residual shortcut led to feature map X_R for the subsequent layers.

V. Experiments

In this section, we detail our experiment circumstances. More specifically, details the dataset used for training, the experiment parameters, and the obtained results. Also, we perform ablation studies to demonstrate the efficiency of our proposed attention blocks.

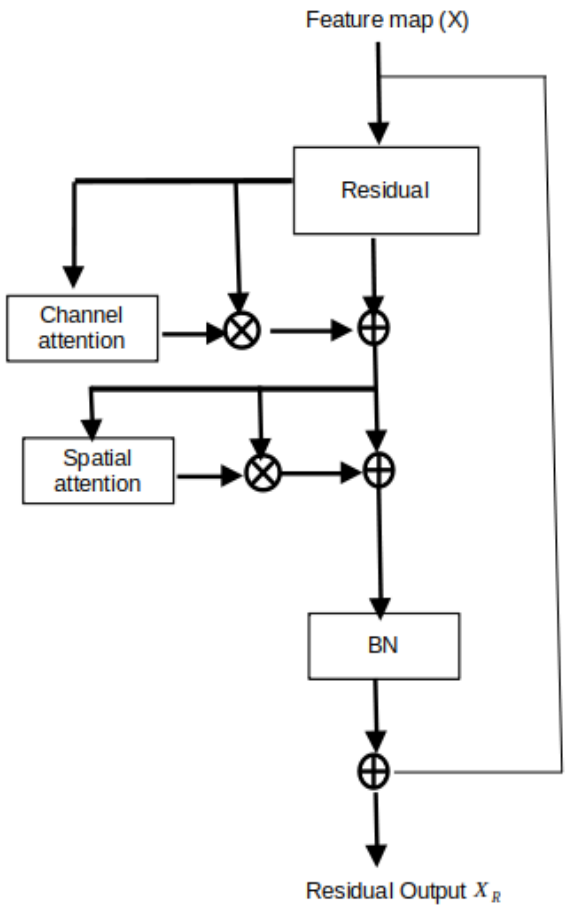

Fig. 5. Feature Embedding Module

A. Dataset

In this study, because of the limited computing resources available, we used only one public dataset called "Labeled Faces in the Wild"(LFW) [68] for training and validation. We chose the "LFW" dataset because it is a public dataset for performance benchmarking of face recognition methods widely used in the research community. It consists of 13, 000 images of faces collected from the web. It is a labeled dataset where each image is associated with a specific person. Some people have more than one image. More specifically, 1680 people have two or more distinct photos in the LFW data set. The dataset contains images with different facial positions in order to maintain consistency and ensure robustness. Slight variations images, such as facial hair, and obstructions images, such as headgear and eyewear, are also included. We discarded classes with less than 15 images and ended up with 96 classes in total. Therefore, we used 91 classes for training and 5 classes for testing. We split the training set according to an 80-20% split for training and validation, resulting in 72 for training and 19 for validation.

B. Experiment Settings

We employed a deep funneling method [69] to align face images. All images were resized to have a size of 112×112. We made small augmentations by horizontally flipping images with a probability of 0.5.

Our model was trained with a constant learning rate η equal to 0.00005 with a step-based decay method decaying at a uniform rate of 1%. We set an early stopping condition to stop the training in the case where the validation accuracy does not improve after 5 iterations. The maximum number of iterations was set to 500. We initialized the momentum for each layer to 0.5, evolving linearly until reaching a final value of 0.9. The batch size of 8 was set. We initialized the biases with the default setting of zeros in all layers, while the Glorot uniform initializer was used to initialize the weights of all layers of the network where the initializer draws samples from the uniform distribution of $[-g, g]$. g is expressed as follows:

$$g = sqrt\left(\frac{6}{fan_{in} + fan_{out}}\right) \tag{14}$$

Where fan_{in} and fan_{out} represent respectively the number of input units in the weight tensor, and the number of output units in the weight tensor.

The ResNet-50 with an attention block on top of each residual employs Triplet Loss expressed as follows:

$$L = max(D(a,p) - D(a,n) + margin, 0)) \tag{15}$$

Where a, p, and n denote the anchor, positive and negative image, respectively. Margin defines how far away the dissimilarities are (between the anchor image and the negative image and between the anchor image and the positive image.).

N-way one-shot learning is used for evaluation, where N denotes the support classes. We experiment with 3 values of N, namely, 5, 10, and 20.

C. Models Architecture

We adopt the ResNet-50 [1] as the feature extractor. We used siamese one-shot learning for face verification and KNN for face identification. The residual bottleneck's structure is $BN - ConvBN - PReLU - Conv - BN$. A kernel size of 3×3 is used in all convolution blocks. The number of channels in different blocks is respectively 64, 128, 256, and 512 as presented in Table I. Our proposed model adds an attention module, consisting of the channel and spatial attention, on top of each residual block as shown in TABLE I. The last layer of the architecture is a fully-connected layer of output dimension equal to 512, which represents the feature vector of the face image.

D. Experiment Results and Performance Comparison

A set of experiments are carried out in this part to demonstrate the performance of our proposed model. We also design a

TABLE I. MODELS ARCHITECTURE

ResNet-50	ResNet-50 with attention	Output size
conv 3×3, 64, pad 1, stride 1		112×112
$\begin{bmatrix} conv, & 3 \times 3, & 64 \\ conv, & 3 \times 3, & 64 \end{bmatrix} \times 03$	$\begin{bmatrix} conv, & 3 \times 3, & 64 \\ conv, & 3 \times 3, & 64 \\ attention, & 64 & \end{bmatrix} \times 03$	56×56
$\begin{bmatrix} conv, & 3 \times 3, & 128 \\ conv, & 3 \times 3, & 128 \end{bmatrix} \times 04$	$\begin{bmatrix} conv, & 3 \times 3, & 128 \\ conv, & 3 \times 3, & 128 \\ attention, & 128 & \end{bmatrix} \times 04$	28×28
$\begin{bmatrix} conv, & 3 \times 3, & 256 \\ conv, & 3 \times 3, & 256 \end{bmatrix} \times 14$	$\begin{bmatrix} conv, & 3 \times 3, & 256 \\ conv, & 3 \times 3, & 256 \\ attention, & 256 & \end{bmatrix} \times 14$	14×14
$\begin{bmatrix} conv, & 3 \times 3, & 512 \\ conv, & 3 \times 3, & 512 \end{bmatrix} \times 03$	$\begin{bmatrix} conv, & 3 \times 3, & 512 \\ conv, & 3 \times 3, & 512 \\ attention, & 256 & \end{bmatrix} \times 03$	7×7
Fully Connected, 512	Fully Connected, 512	

set of ablation studies to show the importance of using attention mechanisms. Additionally, we arrange the attention blocks in several ways, namely, (1) Using only the channel attention, (2) Only the spatial attention, (3) Using both in parallel, (4) Using both sequentially, especially the channel attention followed by spatial channel, (5) Using both sequentially but the channel attention after the spatial attention.

In every ablation study, the ResNet-50 is used as the feature extractor. The squeeze and excitation network is a kind of attention that was proposed in [2]. In [2], Hu et al. used Global Average Pooling for channel attention. However, as mentioned previously, GAP selects only the lowest frequency information for channel attention and other information of other frequencies is simply lost while they also encode some useful information to represent the channels. Therefore, in the performance comparison part, we implemented a ResNet-50 with a squeeze and excitation block on top of every residual.

In addition, we chose to compare our proposal with [70] because we found out that, according to our best knowledge, a recent face recognition work used a one-shot learning concept. Reference [34] is a good paper that employed both channel and spatial attention for face recognition using the "LFW" dataset even the authors used GAP in their channel attention block in contrast to our proposal where we used DCT to include in addition to the lowest frequency information, the information of other frequencies. However, the authors of [34] did not use One-Short learning.

In this paper, all models are evaluated in terms of the one-shot face recognition accuracy. The performance of different models and comparing our proposed attention module to the state-of-the-art is presented in Table II. In Table II, (1), (2), (3), (4), (5) at the bottom of the Table denote the designed ablation use cases.

Firstly, we trained our One-Shot siamese network using ResNet-50 as the backbone network without the attention concept. We achieved recognition accuracy of 97.08, 97.34, and 97.27, respectively, on 5-way, 10-way and 20-way One-Shot learning. However, including only the channel attention increase the recognition accuracy where we obtained 98.45, 98.18, and 98.04 on 5-way, 10-way and 20-way One-Shot learning, respectively. Moreover, the use of only spatial attention achieves 97.88, 98.15 and 98.02, respectively on 5-way, 10-way and 20-way demonstrating that the importance of spatial attention is lower than channel attention as the use of only channel attention obtained better results when compared to using only spatial attention. Then, we applied both channel and spatial attention in parallel and achieved recognition accuracy of 98.38, 98.42 and 98.25 on 5-way, 10-way and 20-way One-Shot learning respectively. Using channel and spatial attention increased the recognition accuracy, meaning both are important in our face recognition system. We, then, arranged the attention blocks sequentially. The channel attention block before the spatial attention achieved the highest results of **98.58**, **98.89** and **98.74** on 5-way, 10-way and 20-way One-Shot learning, among all our implemented models. Using spatial attention before channel attention is the second-best model where recognition accuracy of 98.50, 98.80 and 98.70 was achieved on 5-way, 10-way and 20-way One-Shot learning, respectively.

Finally, we compared the performance of our proposed model with two baselines, namely, [70] and [2]. In [70], no attention mechanism was used. The recognition accuracy of Chanda et al. [70] proposal on 5-way, 10-way and 20-way are 97.00, 97.50 and 95.5 respectively which is confirmed by our model without attention mechanism in Table II. However, the use of squeeze

TABLE II. EXPERIMENT RESULTS AND PERFORMANCE COMPARISON

Method	5-way (%)	10-way (%)	20-Way (%)
Chanda et al. [70]	97.00	97.50	95.50
ResNet-50 with Sqeeze and Excitation [2]	97.63	97.90	96.60
ResNet-50 (without attention)	97.08	97.34	97.27
(1)ResNet-50+Channel attention(CA)	98.45	98.18	98.04
(2)ResNet-50+Spatial attention(SA)	97.88	98.15	98.02
(3)ResNet-50+CA+SA in parallel	98.38	98.42	98.25
(4)**ResNet-50+CA+SA sequentially**	**98.58**	**98.89**	**98.74**
(5)ResNet-50+SA+CA sequentially	98.50	98.80	98.70

and excitation (attention with GAP) of the second baseline [2] achieved recognition accuracy of 97.63, 97.90 and 96.60, respectively on 5-way, 10-way and 20-way.

Based on the above-performance comparison use cases, our best model outperforms the two baselines, demonstrating the efficiency of DCT over the GAP. This was achieved because instead of using only the lowest frequency's information as the GAP does, DCT includes the meaningful information of other frequencies, boosting the recognition accuracy of our face recognition model.

VI. CONCLUSION AND FUTURE WORK

In this paper, we have investigated the use of channel and spatial attention blocks to boost deep face recognition with siamese One-Shot learning. The proposed attention module learns the global-feature relationships of aligned face images with the aim of reducing the information redundancy among channels and focusing on the most relevant parts of face feature maps. Instead of using the GAP that is used in current face recognition models, especially in the channel attention block, we consider in this paper, the scalar representation of a channel as a compression problem and used discrete cosine transform (DCT) to compress the channels for the channel attention mechanism in order to consider the information of other frequencies in addition to the information of the lowest frequency. Our attention module employs both channel and spatial attention arranged sequentially with channel attention before spatial attention. On a public dataset, we demonstrated that our attention module can achieve better results when integrated into ResNet-50 compared to the existing attention mechanism used in deep face recognition.

We believe that training our model on deeper networks like ResNet-100 with more data will help improve our model's performance, which is considered future research work.

REFERENCES

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016, doi: 10.1109/CVPR.2016.90.

[2] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation Networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132–7141, 2018, doi: 10.1109/CVPR.2018.00745.

[3] M. A. Al-Shareeda, A. A. Alsadhan, H. H. Qasim, and S. Manickam, "Software Defined Networking for Internet of Things: Review, Techniques, Challenges, and Future Directions," *Bulletin of Electrical Engineering and Informatics*, vol. 13, no. 1, pp. 638–647, 2024, doi: 10.11591/eei.v13i1.6386.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet Classification With Deep Convolutional Neural Networks," *Advances in neural information processing systems*, vol. 25, 2012, doi: 10.1145/3065386.

[5] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going Deeper With Convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015, doi: 10.1109/CVPR.2015.7298594.

[6] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "Cosface: Large Margin Cosine Loss for Deep Face Recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5265–5274, 2018, doi: 10.48550/arXiv.1801.09414.

[7] A. A. Almazroi, E. A. Aldhahri, M. A. Al-Shareeda, and S. Manickam, "Eca-Vfog: an Efficient Certificateless Authentication Scheme for 5g-Assisted Vehicular Fog Computing," *Plos one*, vol. 18, no. 6, pp. 1–20, 2023, doi: 10.1371/journal.pone.0287291

[8] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep Hypersphere Embedding for Face Recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 212–220, 2017, doi: 10.48550/arXiv.1704.08063

[9] M. A. Al-Shareeda and S. Manickam, "A Systematic Literature Review on Security of Vehicular Ad-Hoc Network (Vanet) Based on Veins Framework," *IEEE Access*, vol. 11, pp. 46218-46228, 2023, doi: 10.1109/ACCESS.2023.3274774.

[10] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive Angular Margin Loss for Deep Face Recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4685-4694, 2019, doi: 10.1109/CVPR.2019.00482.

[11] F. Wang, J. Cheng, W. Liu, and H. Liu, "Additive Margin Softmax for Face Verification," *IEEE Signal Processing Letters*, vol. 25, no. 7, pp. 926–930, 2018, doi: 10.1109/LSP.2018.2822810.

[12] Z. G. Al-Mekhlafi, M. A. Al-Shareeda, S. Manickam, B. A. Mohammed, A. Alreshidi, M. Alazmi, J. S. Alshudukhi, M. Alsaffar, and T. H. Rassem, "Efficient Authentication Scheme for 5g-Enabled Vehicular Networks Using Fog Computing," *Sensors*, vol. 23, no. 7, p. 3543, 2023, doi: 10.3390/s23073543.

[13] G. B. Huang and E. Learned-Miller, "Labeled Faces in the Wild: Updates and New Reporting Procedures," *Dept. Comput. Sci., Univ. Massachusetts Amherst, Amherst, MA, USA, Tech. Rep*, vol. 14, no. 3, 2014,

[14] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, and E. Brossard, "The Megaface Benchmark: 1 Million Faces for Recognition at Scale," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4873–4882, 2016.

[15] A. A. Almazroi, M. A. Alqarni, M. A. Al-Shareeda, and S. Manickam, "L-Cppa: Lattice-Based Conditional Privacy-Preserving Authentication Scheme for Fog Computing With 5g-Enabled Vehicular System," *Plos one*, vol. 18, no. 10, pp. 1–23, 2023, doi: 10.1371/journal.pone.0292690.

[16] G. Salomon, A. Britto, R. H. Vareto, W. R. Schwartz, and D. Menotti, "Open-Set Face Recognition for Small Galleries Using Siamese Networks," in *2020 International Conference on Systems, Signals and*

Image Processing (IWSSIP),pp. 161–166, 2020, doi: 10.1109/IWSSIP48289.2020.9145245.

[17] M. A. Al-Shareeda, A. A. Alsadhan, H. H. Qasim, and S. Manickam, "Long Range Technology for Internet of Things: Review, Challenges, and Future Directions," *Bulletin of Electrical Engineering and Informatics*, vol. 12, no. 6, pp. 3758–3767, 2023, doi: 10.11591/eei.v12i6.5214.

[18] Y. Sun, Y. Chen, X. Wang, and X. Tang, "Deep Learning Face Representation by Joint Identification-Verification," *Advances in neural information processing systems*, vol. 27, 2014, doi: 10.48550/arXiv.1406.4773.

[19] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A Discriminative Feature Learning Approach for Deep Face Recognition," in *European conference on computer vision*, vol. 9911, pp. 499–515, 2016, doi: 10.1007/978-3-319-46478-7_31.

[20] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A Unified Embedding for Face Recognition and Clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823, 2015.

[21] X. Wang, S. Wang, S. Zhang, T. Fu, H. Shi, and T. Mei, "Support Vector Guided Softmax Loss for Face Recognition," *Computer Vision and Pattern Recognition*, 2018, doi: 10.48550/arXiv.1812.11317.

[22] S. Lei, W. Yi, C. Ying, and W. Ruibin, "Review of Attention Mechanism in Natural Language Processing," *Data Analysis and Knowledge Discovery*, vol. 4, no. 5, pp. 1–14, 2020, doi: 10.11925/infotech.2096-3467.2019.1317.

[23] M. A. Al-Shareeda, S. Manickam, and M. Ali, "Ddos Attacks Detection Using Machine Learning and Deep Learning Techniques: Analysis and Comparison," *Bulletin of Electrical Engineering and Informatics*, vol. 12, no. 2, pp. 930–939, 2023.

[24] A. Galassi, M. Lippi, and P. Torroni, "Attention in Natural Language Processing," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 10, pp. 4291–4308, 2020, doi: 10.1109/TNNLS.2020.3019893.

[25] S. U. A. Laghari, S. Manickam, A. K. Al-Ani, M. A. Al-Shareeda, and S. Karuppayah, "ES-SECS/GEM: An Efficient Security Mechanism for SECS/GEM Communications," *IEEE Access*, vol. 11, pp. 31813-31828, 2023, doi: 10.1109/ACCESS.2023.3262310.

[26] C. Yu, Z. Zhang, H. Li, J. Sun, and Z. Xu, "Meta-Learning-Based Adversarial Training for Deep 3d Face Recognition on Point Clouds," *Pattern Recognition*, vol. 134, p. 109065, 2023, doi: 10.1016/j.patcog.2022.109065

[27] F. Liu, D. Chen, F. Wang, Z. Li, and F. Xu, "Deep Learning Based Single Sample Face Recognition: A Survey," *Artificial Intelligence Review*, vol. 56, no. 3, pp. 2723–2748, 2023, doi: 10.1007/s10462-022-10240-2.

[28] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual Attention Network for Image Classification," in *Computer Vision and Pattern Recognition*, pp. 3156–3164, 2017, doi: 10.48550/arXiv.1704.06904.

[29] B. A. Mohammed, M. A. Al-Shareeda, S. Manickam, Z. G. Al-Mekhlafi, A. M. Alayba, and A. A. Sallam, "Anaa-Fog: A Novel Anonymous Authentication Scheme for 5g-Enabled Vehicular Fog Computing," *Mathematics*, vol. 11, no. 6, p. 1446, 2023, doi: 10.3390/math11061446.

[30] Q. Zhao, J. Liu, Y. Li, and H. Zhang, "Semantic Segmentation With Attention Mechanism for Remote Sensing Images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2021, doi: 10.1109/TGRS.2021.3085889.

[31] B. A. Mohammed, M. A. Al-Shareeda, S. Manickam, Z. G. Al-Mekhlafi, A. Alreshidi, M. Alazmi, J. S. Alshudukhi, and M. Alsaffar, "Fc-Pa: Fog Computing-Based Pseudonym Authentication Scheme in 5g-Enabled Vehicular Networks," *IEEE Access*, vol. 11, pp. 18 571–18 581, 2023, doi: 10.1109/ACCESS.2023.3247222.

[32] M. Jian, K. -M. Lam, J. Dong and L. Shen, "Visual-Patch-Attention-Aware Saliency Detection," *in IEEE Transactions on Cybernetics*, vol. 45, no. 8, pp. 1575-1586, 2015, doi: 10.1109/TCYB.2014.2356200.

[33] M. A. Al-Shareeda, S. Manickam, B. A. Mohammed, Z. G. Al-Mekhlafi, A. Qtaish, A. J. Alzahrani, G. Alshammari, A. A. Sallam, and K. Almekhlafi, "Provably Secure With Efficient Data Sharing Scheme for Fifth-Generation (5g)-Enabled Vehicular Networks Without Road-Side Unit (RSU)," *Sustainability*, vol. 14, no. 16, p. 9961, 2022, doi: 10.3390/su14169961.

[34] H. Ling, J. Wu, J. Huang, J. Chen, and P. Li, "Attention-Based Convolutional Neural Network for Deep Face Recognition," *Multimedia Tools and Applications*, vol. 79, no. 9, pp. 5595–5616, 2020, doi: 10.1007/s11042-019-08422-2.

[35] Y. Rao, J. Lu, and J. Zhou, "Attention-Aware Deep Reinforcement Learning for Video Face Recognition," in *Proceedings of the IEEE international conference on computer vision*, pp. 3951-3960, 2017, doi: 10.1109/ICCV.2017.424.

[36] M. Sajjad, F. U. M. Ullah, M. Ullah, G. Christodoulou, F. A. Cheikh, M. Hijji, K. Muhammad, and J. J. Rodrigues, "A Comprehensive Survey on Deep Facial Expression Recognition: Challenges, Applications, and Future Guidelines," *Alexandria Engineering Journal*, vol. 68, pp. 817–840, 2023, doi: 10.1016/j.aej.2023.01.017.

[37] M. A. Al-Shareeda, S. Manickam, M. A. Saare, and N. B. Omar, "Sadetection: Security Mechanisms to Detect Slaac Attack in Ipv6 Link-Local Network," *Informatica*, vol. 46, no. 9, 2023, doi: 10.31449/inf.v46i9.4441.

[38] T. Ahonen, A. Hadid and M. Pietikainen, "Face Description with Local Binary Patterns: Application to Face Recognition," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 28, no. 12, pp. 2037-2041, 2006, doi: 10.1109/TPAMI.2006.244.

[39] M. Jabberi, A. Wali, B. B. Chaudhuri, and A. M. Alimi, "68 landmarks are efficient for 3d face alignment: what about more? 3d face alignment method applied to face recognition," *Multimedia Tools and Applications*, vol. 82 pp. 41435–41469, 2023, doi: 10.1007/s11042-023-14770-x.

[40] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *Computer Vision and Pattern Recognition*, pp. 1–14, 2014, doi: 10.48550/arXiv.1409.1556.

[41] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size," *Computer Vision and Pattern Recognition*, 2016, doi: 10.48550/arXiv.1602.07360.

[42] G. Hu et al., "When Face Recognition Meets with Deep Learning: An Evaluation of Convolutional Neural Networks for Face Recognition," *2015 IEEE International Conference on Computer Vision Workshop*, pp. 384-392, 2015, doi: 10.1109/ICCVW.2015.58.

[43] G. Guo and N. Zhang, "A survey on deep learning based face recognition," *Computer vision and image understanding*, vol. 189, p. 102805, 2019, doi: 10.1016/j.cviu.2019.102805.

[44] Y. Taigman, M. Yang, M. Ranzato and L. Wolf, "DeepFace: Closing the Gap to Human-Level Performance in Face Verification," *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1701-1708, 2014, doi: 10.1109/CVPR.2014.220.

[45] Y. Guo and L. Zhang, "One-shot Face Recognition by Promoting Underrepresented Classes," *Computer Vision and Pattern Recognition*, 2017, doi: 10.48550/arXiv.1707.05574.

[46] L. Wang, Y. Li and S. Wang, "Feature Learning for One-Shot Face Recognition," *2018 25th IEEE International Conference on Image Processing*, pp. 2386-2390, 2018, doi: 10.1109/ICIP.2018.8451464.

[47] Z. Ding, Y. Guo, L. Zhang and Y. Fu, "One-Shot Face Recognition via Generative Learning," *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition*, pp. 1-7, 2018, doi: 10.1109/FG.2018.00011.

[48] A. Jadhav, V.P. Namboodiri and K.S. Venkatesh, "Deep Attributes for One-Shot Face Recog- nition", *ECCV Workshop on 'Transfering and Adapting Source Knowledge in Computer Vision'*, 2016.

[49] Y. Wu, H. Liu, and Y. Fu, "Low-shot face recognition with hybrid classifiers," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 1933–1939, 2017.

[50] S. Hong, W. Im, J. Ryu and H. S. Yang, "SSPP-DAN: Deep domain adaptation network for face recognition with single sample per person," *2017 IEEE International Conference on Image Processing*, pp. 825-829, 2017, doi: 10.1109/ICIP.2017.8296396.

[51] Y. Cheng et al., "Know You at One Glance: A Compact Vector Representation for Low-Shot Learning," *2017 IEEE International Conference on Computer Vision Workshops*, pp. 1924-1932, 2017, doi: 10.1109/ICCVW.2017.227.

[52] Q. Cao, L. Shen, W. Xie, O. M. Parkhi and A. Zisserman, "VGGFace2: A Dataset for Recognising Faces across Pose and Age," *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition*, pp. 67-74, 2018, doi: 10.1109/FG.2018.00020.

[53] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3146–3154, 2019.

[54] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature Verification Using A" Siamese" Time Delay Neural Network," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 7, no. 4, pp. 669–688, 1993, doi: 10.1142/S0218001493000339.

[55] G. Koch, R. Zemel, R. Salakhutdinov, et al., "Siamese neural networks for one-shot image recognition," in *ICML deep learning workshop*, vol. 2, no. 1, 2015.

[56] L. Song, D. Gong, Z. Li, C. Liu, and W. Liu, "Occlusion Robust Face Recognition Based on Mask Learning With Pairwise Differential Siamese Network," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 773–782, 2019, doi: 10.1109/ICCV.2019.00086.

[57] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional Block Attention Module," in *Proceedings of the European Conference on Computer Vision*, pp. 3–19, 2018, doi: 10.1007/978-3-030-01234-2_1.

[58] X. Wang, R. Girshick, A. Gupta and K. He, "Non-local Neural Networks," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7794-7803, 2018, doi: 10.1109/CVPR.2018.00813.

[59] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," *Computation and Language*, 2017, doi: 10.48550/arXiv.1706.03762.

[60] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-Attention Generative Adversarial Networks," in *Proceedings of the 36th International Conference on Machine Learning*, vol. 97, pp. 7354–7363, 2019.

[61] N. Ahmed, T. Natarajan and K. R. Rao, "Discrete Cosine Transform," in *IEEE Transactions on Computers*, vol. C-23, no. 1, pp. 90-93, 1974, doi: 10.1109/T-C.1974.223784.

[62] T. Elsken, J. H. Metzen, and F. Hutter, "Neural Architecture Search: A Survey," *Machine Learning*, vol. 20, pp. 1–21, 2019, doi: 10.48550/arXiv.1808.05377.

[63] S. Karagiannakos, "Neural architecture search (nas): basic principles and different approaches," *https://theaisummer.com/*, 2021.

[64] M. Wistuba, A. Rawat, and T. Pedapati, "A Survey on Neural Architecture Search," *Machine Learning*, 2019, doi: 10.48550/arXiv.1905.01392.

[65] H. Lee, H. -E. Kim and H. Nam, "SRM: A Style-Based Recalibration Module for Convolutional Neural Networks," *2019 IEEE/CVF International Conference on Computer Vision*, pp. 1854-1862, 2019, doi: 10.1109/ICCV.2019.00194.

[66] Z. Qin, P. Zhang, F. Wu and X. Li, "FcaNet: Frequency Channel Attention Networks," *2021 IEEE/CVF International Conference on Computer Vision*, pp. 763-772, 2021, doi: 10.1109/ICCV48922.2021.00082.

[67] S. Zagoruyko and N. Komodakis, "Paying More Attention to Attention: Improving the Performance of Convolutional Neural Networks via Attention Transfer," *Computer Vision and Pattern Recognition*, 2016, doi: 10.48550/arXiv.1612.03928.

[68] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled Faces in the Wild: A Database Forstudying Face Recognition in Unconstrained Environments," in *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008.

[69] G. Huang, M. Mattar, H. Lee, and E. Learned-Miller, "Learning to Align From Scratch," *Advances in neural information processing systems*, vol. 25, 2012.

[70] S. Chanda, A. C. GV, A. Brun, A. Hast, U. Pal and D. Doermann, "Face Recognition - A One-Shot Learning Perspective," *2019 15th International Conference on Signal-Image Technology & Internet-Based Systems*, pp. 113-119, 2019, doi: 10.1109/SITIS.2019.00029.