IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received July 19, 2021, accepted August 18, 2021, date of publication August 24, 2021, date of current version September 1, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3107626

SiCoDeF^2 Net: Siamese Convolution Deconvolution Feature Fusion Network for One-Shot Classification

SWALPA KUMAR ROY[1], (Student Member, IEEE), PURBAYAN KAR[1], MERCEDES E. PAOLETTI[2], (Senior Member, IEEE), JUAN M. HAUT[3], (Senior Member, IEEE), RAFAEL PASTOR-VARGAS[4], (Senior Member, IEEE), AND ANTONIO ROBLES-GÓMEZ[4], (Senior Member, IEEE)

[1]Department of Computer Science and Engineering, Jalpaiguri Government Engineering College, Jalpaiguri 735102, India
[2]Department of Computer Architecture, Faculty of Computer Science, Complutense University, 28040 Madrid, Spain
[3]Department of Computers and Communications, Escuela Politecnica, University of Extremadura, 10003 Caceres, Spain
[4]Department of Communication and Control Systems, Higher School of Computer Engineering, National Distance Education University, 28015 Madrid, Spain

Corresponding author: Juan M. Haut (juanmariohaut@unex.es)

This work was supported in part by Leonardo Grants for Researchers and Cultural Creators awarded by the Fundación BBVA under Grant 2021/00203/00, in part by the eNMoLabs Research Project, in part by the CiberGID UNED Innovation Group through the CiberScratch Project, in part by the I4Labs UNED Research Group, in part by the E-Madrid-CM Network of Excellence under Grant S2018/TCS-4307, and in part by SNOLA, officially recognized Thematic Network of Excellence by the Spanish Ministry of Science, Innovation and Universities, under Grant RED2018-102725-T.

ABSTRACT Nowadays, deep convolutional neural networks (CNNs) for face recognition exhibit a performance comparable to human ability in the presence of the appropriate amount of labelled training data. However, training CNNs remains as an arduous task due to the lack of training samples. To overcome this drawback, applications demand one-shot learning to improve the obtained performances over traditional machine learning approaches by learning representative information about data categories from few training samples. In this context, Siamese convolutional network (`SiConvNet`) provides an interesting deep architecture to tackle the data limitation. In this regard, applying the convolution operation on real world images by using the trainable correlative Gaussian kernel adds correlations to the output images, which hinder the recognition process due to the blurring effects introduced by the convolution kernel application. As a result the pixel-wise and channel-wise correlations or redundancies could appear in both single and multiple feature maps obtained by a hidden layer. In this sense, convolution-based models fail to generalize the feature representation because of both the strong correlations presence in neighboring pixels and the channel-wise high redundancies between different channels of the feature maps, which hamper the effective training. *Deconvolution* operation helps to overcome the shortcomings that limit the conventional `SiConvNet` performance, learning successfully correlation-free features representation. In this paper, a simple but efficient Siamese convolution deconvolution feature fusion network (SiCoDeF^{2}Net) is proposed to learn the invariant and discriminative complementary features generated from both the (i) sub-convolution (SCoNet) and (ii) sub deconvolutional (SDeNet) networks using a concatenation operation which significantly improves the one-shot unconstrained facial recognition task. Extensive experiments performed on several widely used benchmarks, provide promising results, where the proposed SiCoDeF^{2}Net model significantly outperforms the current state-of-art in terms of classification accuracy, F1, precision and recall. The code will be available on: https://github.com/purbayankar/SiCoDeF2Net.

INDEX TERMS Convolutional neural networks (CNNs), deep learning, face recognition, one-shot learning.

I. INTRODUCTION

The face is one of the most popular biometric features for the verification and identification of a person, as it is ubiquitous for the entire human race and quite simple to obtain, as it is easily acquired in unconstrained environments through non-invasive/low-intrusiveness techniques, such as optical imaging [1], [2]. In this sense, the analysis of this data provides useful and representative features to perform accurate facial recognition tasks [3], which is critical in a wide range of information security-related systems, such as automatic access control [4], security surveillance [5], [6],

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

or smartphones applications [7], among others. As a result, face recognition [8] has acquired a significant attention during the past decades within automatic image processing. In fact, it is considered one of the main analysis problems in the range of object recognition [9], texture recognition [10] or content-based image retrieval (CBIR) [11]. However, the classification of unconstrained face images is an ill-posed problem due to the high data variability, which appear in terms of pose, illumination, expressions, age, cosmetics, artificial occlusion, degradation of image quality [12], [13], etc., and the limited number of training samples to properly cover this variability.

Many efforts have been devoted to overcome the above problems over the past decades. As a result, the current literature provides an important number of works that address the face recognition task by applying different perspectives and strategies [8], [14]–[19]. Nevertheless, they are mainly concentrated on learning invariant and discriminative feature representation from face images and videos. In this sense, it can be assumed that the learning of invariant and discriminative feature representation is the first and crucial step of any face recognition system. Broadly speaking, this step can be achieved considering two different and opposite approaches: i) the manually-designed or hand-crafted features and ii) the feature representations automatically learned from sample data [20]. Further details are provided below.

A. FROM HAND-CRAFTED TO AUTOMATIC FEATURE EXTRACTION

Focusing on the first approach, the journey of hand-crafted descriptors started at early nineties and became quite popular due to its design simplicity and its computational efficiency [21]. Indeed, hand-crafted descriptors have proven to be quite useful techniques within the computer vision research, specially when the class-specific available samples for training are limited, providing a good trade-off between accuracy and computational efficiency and extracting robust features by avoiding artifact-driven descriptors. Within the available literature, there are many interesting works that successfully employ these features, in particular those that extract local patterns across the entire image, encoding textual and gradient based information. Within face recognition task, some works stand out, for instance Ahonen *et al.* proposed a novel feature called local binary pattern (LBP) for effective face recognition system, achieving a successful classification performance [22]. Since then, many variants of LBP have been implemented, where some of them have been successfully designed and applied in face recognition tasks [23]. Inspired by LBP, Zhang *et al.* introduced local Gabor binary pattern histogram sequence (LGBPHS), which combines the magnitude part of Gabor feature with LBP operator, achieving good performance on face recognition task [24]. Also Zhang *et al.* proposed a compact and effective histogram descriptor based on Gabor phase pattern (HGPP) for robust face recognition [25]. Chen *et al.* extracted high-dimensional multi-scale LBP features from patches around the key point of facial landmarks [26]. In addition to these techniques, the local derivative is gaining attention since the most relevant and discriminative features exist in the direction of higher order derivative. In this regard, to encode the most informative eight neighbors relationship with respect to the reference pixel, Zhang *et al.* proposed the local derivative pattern (LDP), which computes the four directional derivatives of a face image for recognition and retrieval purposes [27]. Murala *et al.* modified LDP and proposed local tetra pattern (LTrP) by splitting the image along the 0° and 90° derivatives, encoding the most informative eight neighbors relationship with respect to the reference pixel [11]. Other interesting approaches have been proposed by Chen *et al.* [28], who learned dictionary-based on Fisher Vector encoding technique for face image recognition, whereas Lu *et al.* [29] introduced joint feature learning to generate sparse code dictionaries from the local patch, pooling those to produce a high dimensional feature vector.

In modern face recognition era, convolutional neural network (CNN) has shown its potential to learn compact and discriminative representation for many image processing tasks, reaching promising results in training and classification of large facial datasets. The face recognition using CNN can be generally classified into three groups: i) classification, ii) matching, and iii) identification or verification. Focusing on classification, Sun *et al.* extracted feature vectors from unconstrained face images and then predicted the label of the obtained feature vectors using the classifiers [30] or trained on different local patches with joint Bayesian ensemble model [15]. In DeepFace [31], the CNNs were used to extract deep features and to train the network on large-scale frontal face images to perform the final classification, achieving better performance in comparison with traditional methods. Wu *et al.* proposed a lightened CNN framework [32] to learn a compact and invariant embedding space for face representation under noisy face image. The main goal of the second approach is to match the pairs of face images by optimizing the verification loss directly, overcoming the problem associated with multi-classification network, which usually fails to generalize into new instances when they do not belong to training set or even not are seen during training. To increase the inter-class separability and reduce the intra-class distance, Sun *et al.* [33] combined both the verification and classification loss to design a relatively cheap network to provide further improvement. Similarly, FaceNet [34] combined triplet loss and adopted an architecture for deep object recognition to directly optimize the embedding space, training on large-scale unaligned face datasets. Parkhi *et al.* [35] trained a VGG network [36] and fine-tuned it by optimizing a triplet loss function on the top of the model. Yi *et al.* [37] also trained a deep CNN model using a relatively small face dataset to learn invariant and discriminative feature representation. Despite their results, the limitations of these methods lie, on the one hand, on the proper selection of

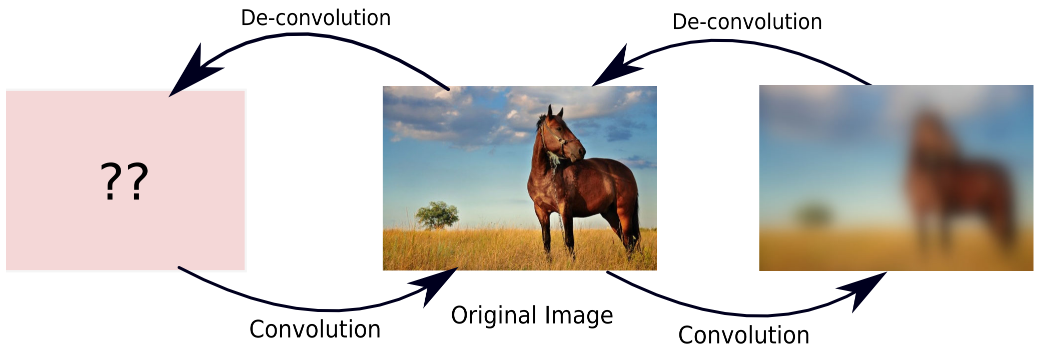


FIGURE 1. **The convolution operation applied on a real-world image (center) using a correlated Gaussian kernel adds undesired correlations to the output feature map (right), impairing object recognition due to the blurring effects of the kernels.**

negative pairs from training data and, on the other hand, on the manually determination of the threshold, which plays an important role in verification loss. To overcome these limitations, identification/verification methods propose to optimize the deep face network by combining identification and verification restrictions together. For instance, Chopra *et al.* introduced the learning of contrastive similarity energy function from training face image pairs for face verification task [38]. To minimize the intra-personal distance while maximizing the inter-personal distances, Sun *et al.* designed a deep neural model to extract effective deep IDentification-verification (DeepID2) features, where face identification and verification signals are combined and employed to supervise the model [15].

Recently, deep metric learning has been considered in the new studies about face recognition tasks to fulfill the same goal, either directly or indirectly. In metric learning, faces are learned and transformed into a low dimensional feature space, where those faces from the same instances are close to each other but stay apart for different instances, while the learning similarity function can boost the performance [39]. The current state-of-the-art regarding metric learning methods include information theoretic metric learning (ITML) [40], which learns the Mahalanobis distance based information theoretic objective function, and large margin nearest neighbor (LMNN) [41], which helps to learn the marginal constraint effects among the triplets from training samples. These models are much simpler because of both their linear nature and their shallow architectures. However, there is a lack of experimentation of these techniques on challenging, sampler, real-world human faces datasets. For instance, Koch *et al.* proposed one-shot learning framework based on convolution feature by minimizing contrastive similarity loss [38], called Siamese convolutional network [42], which has been quite successful in challenging conditions, for instance, when the number of classes during training is unknown and/or few samples are available per class.

In this paper, instead of learning simply convolutional features, the proposed Siamese convolution deconvolution feature fusion network (`SiCoDeF`^{2}`Net`) can produce complementary features through its two subnetworks: i) the Siamese convolution network (`SiConvNet`) and ii) the Siamese deconvolution network (`SiDeConvNet`), respectively. The former one is built using standard convolution layer, while the latter one comprises several deconvolution layers which completely replaces the convolution layers to efficiently remove the pixel-wise and channel-wise correlations. Experimental evaluations conducted over several widely used face benchmarks show that the proposed complementary features are able to boost the performance significantly in case of inadequate or few training samples.

The rest of the paper is organized as follows. Section II delves into the motivations behind the proposed network for face recognition, pointing out the challenges introduced by this task and the limitations of the standard CNN models. Section III provides the details of the proposed methodology, describing the architecture of our new `SiCoDeF`^{2}`Net` model. Section IV conducts several experiments over several widely used facial datasets, in particular AT&T, Yale, extended Yale-B, UFI cropped and LFW face datasets. Moreover, the proposed `SiCoDeF`^{2}`Net` model has been compared with five different deep classification models of the current state-of-the-art. Obtained results demonstrates the improvement of our proposed algorithm. Finally, Section V contains the conclusions.

II. MOTIVATION

The convolution operation is the core in convolutional neural networks (CNNs), where the receptive field (RF) plays an important role while extracting informative features by shifting it across the entire image in an overlapping fashion [43]. Due to the ability of automatic feature extraction, CNNs achieve breakthrough performance and gain enormous attention in the computer vision community [36], [44]–[46]. However, real-world images exhibit strong correlations and due to the existence of such natural correlations, receptive fields are enforce to re-learn redundant information during the convolution operation [47]–[49]. In fact, the standard CNN model wastes a significant effort in creating copies of the kernel weights by rotating, scaling or translating them, which unnecessarily increases the computational burden [50]–[52]. Besides this, as depicted in Fig. 1 applying the convolution operation on real-world images (center) using these correlated Gaussian kernels introduces undesired correlations to the output features (right), which impair recognition tasks due to the blurring effects of the kernel application during convolution. These blurring effects are intrinsically related to the receptive field of the network, which follows a Gaussian distribution instead of an uniform distribution, producing a great impact on the backpropagation, where the central pixels will have a larger gradient magnitude when compared to the border pixels. As a direct consequence, convolution-based models often fail to generalize feature representation due to the presence of strong correlations in neighboring pixels within an image or in the feature maps obtained by a specific layer. Similarly, high redundancies between different feature maps in a hidden layer (called *channel-wise* correlations) also impair the effective training. Furthermore, the real-world images can also be interpreted as the result of some unknown correlative filters, which might be difficult to find, but in terms of the deconvolution operation it is easy to estimate the deconvolution matrix in a reverse manner. Also, the re-learning of redundant information in successively convolutional layers hampers the training of deep CNNs, by wasting precious resources that do not delve into the most discriminating features of the data [53], [54]. In this sense, it is desirable to eliminate such redundancies within the convolution kernels. However, the existing CNNs fail to avoid the re-learning of such information during training, and faster convergence become a key issue for deeper network.

To overcome the above challenges, deep artificial neural network strongly demands an application which can successfully remove both the pixel-wise and channel-wise correlations before the data is processed by each layer. In this context, the *deconvolution* operation offers an interesting solution to this limitation. This operation becomes quite popular among the deep learning community and have proven to be quite effective for image classification and segmentation tasks [55]. These promising results have inspired the adoption of *network deconvolution* to design an end-to-end one-shot learning framework for face recognition, the so-called Siamese convolution deconvolution feature fusion network (`SiCoDeF`^{2}`Net`). In the proposed model, image pairs are passed through twin subnetworks, which are named *SiConvNet* and *SiDeConvNet* respectively, to perform a deep feature extraction, as we can observe in Fig. 2. Moreover, to learn the invariant and discriminative complementary feature representation the extracted features are fused using concatenation for unconstrained facial recognition.

In a nutshell, the main contributions proposed by this paper are highlighted as follows:

- To learn the invariant and discriminative complementary feature representation, the proposed `SiCoDeF`^{2}`Net` model comprises twin networks, the *SiConvNet* and *SiDeConvNet*.
- Focusing on *SiConvNet*, it is a Siamese network composed by standard convolution layers. The subnetwork of *SiConvNet* will be denoted as *SCoNet*. In addition, *SiDeConvNet* is a Siamese network composed by deconvolution layers, and its subnetworks are denoted as *SDeNet*. Both Siamese networks receive the same pair of facial images. The two data representations of each image obtained by *SiConvNet* and *SiDeConvNet* models are combined by a feature fusion strategy.
- Moreover, the subnetwork *SDeNet* in the proposed model eliminates the widely used *batch normalization* layer, achieving faster convergence towards optimization. The combination of the *SCoNet* and *SDeNet* outputs also produce state-of-the-art performance in face datasets.
- The proposed `SiDeConvF`^{2}`Net` learns complementary features by successfully avoiding the re-learning of specific redundant information. This mechanism greatly helps to encode informative features since the training set contains a small number of samples.

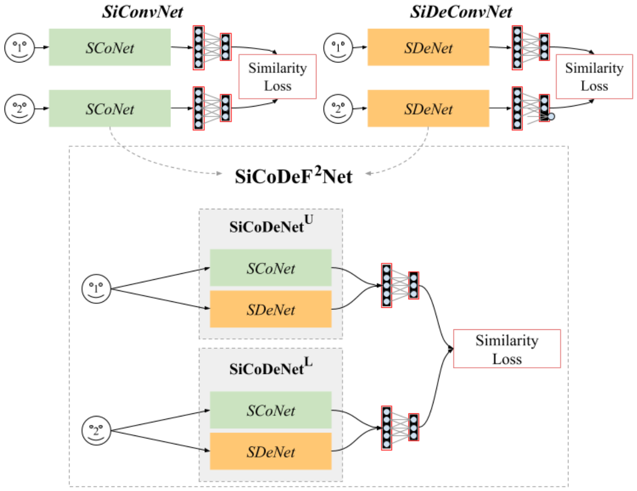


FIGURE 2. Simple graphical interpretation of the proposed network, where the *SiConvNet* and *SiDeConvNet* networks are combined to create the final `SiCoDeF`^{2}`Net` model.

In the following sections we will provide more detailed explanations about the proposed `SiDeConvF`^{2}`Net` model and analyse its performance in comparison with other methods of the current state-of-the-art.

III. PROPOSED SiCoDeF^2 NET MODEL

One-shot learning aims to learn discriminative image representation to assign the corresponding class label to those unseen examples during training with limited labelled samples. This is possible with the help of supervised distance learning and the Siamese convolution network architecture [56]. Due to its simplicity, easy design and low complexity, this model gains attention among the scientific community and has been successfully adopted for weakly supervised metric learning [38], signature verification [57], person re-identification [58] and face verification [34], where the number of labelled instances per classes in a datasets is not enough to train a traditional CNN classifier. Due to the availability of few training samples and the existence of both pixel-wise and channel-wise correlations, the network is enforced to re-learn redundant information during the back-propagation stage, hampering the effective learning of the deep model. In this section, we present the one-shot learning framework `SiCoDeF`^{2}`Net` to learn invariant and discriminative complementary feature representations by combining the two subnetworks *SCoNet* and *SDeNet*, where *SCoNet* uses standard convolution layers and the subnetwork *SDeNet* uses deconvolution layers, respectively. Fig. 2 provides a graphical scheme of the proposed network. Indeed, the combination of the complementary features can better characterise the sparse discriminative representation of embedding space. The steps of the deconvolution operation are detailed below.

A. DECONVOLUTION LAYER

The convolution operation produces correlated information through the widely used $*$ operation, which can be defined as follows:

$$\begin{aligned}\widehat{f}[i,j] &= \sum_{x=-k}^{k}\sum_{y=-k}^{k} f[x+i,y+j]h[x+i,y+j] \\ &= (f*h)[i,j] = \mathbf{H}f, \end{aligned} \tag{1}$$

where the convolution kernel h of size $(k \times k)$ is convolved over the input $f \in \mathbb{R}^{H\times W}$ (where the naturals H and W indicates the height and width dimensions) to produce the highly correlated transform output $\widehat{f}$, where $\mathbf{H}$ corresponds to convolution matrix. In this context, the aims of deconvolution operation is to eliminate the redundancies which are present in the form of correlations through the network. These redundancies can be removed by $f = \mathbf{H}^{-1}\widehat{f}$, assuming that $\mathbf{H}$ is an invertible matrix. Let kernel $(k \times k)$ overlaps the input patches extracted from $f[1:H, 1:W]$ and flatten into a vector of size $[1, k^2]$, which is stacked in column-wise into $\mathcal{X}$ and can be calculated by:

$$\mathcal{X}[,j] = f[i-r : i+r, j-r : j+r], \tag{2}$$

where $r = k/2 - 1$ and the columns of $\mathcal{X}$ shows high correlation among the overlapping patches extracted with stride of size 1. This significantly hampers the convergence of any deep network during its training stage, and even the batch normalization [59] operation becomes unsuccessful to address this limitation. Therefore deep network models require the deconvolution operation to overcome this drawback. To calculate the covariance matrix Σ, the extracted data matrix $\mathcal{X}_{S\times F}$ has to be reshaped and represented by the number of samples S and the number of features F as follows:

$$\Sigma = \frac{1}{S}(\mathcal{X}-\mu)^T(\mathcal{X}-\mu) \tag{3}$$

where μ is the mean of the data matrix and T represents the transpose. In order to avoid the correlation effects from both pixel and channel dimensions, the mean shifted centered data $(\mathcal{X}-\mu)$ is multiplied using the approximated inverse square root of Σ, resulting into $(\mathcal{X}-\mu)\cdot\mathbf{D}$ where $\mathbf{D} = \Sigma^{-\frac{1}{2}}$ is the deconvolution matrix. Algorithm 1 provides the steps to successfully compute $\mathbf{D}$. If $\mathbf{D}$ is well approximated through the widely used Newton-Schulz method [60], the covariance matrix Σ' of transformed data produces an identity matrix as below:

$$\begin{aligned}\Sigma' &= \mathbf{D}^T(\mathcal{X}-\mu)^T(\mathcal{X}-\mu)\mathbf{D} \\ &= \Sigma^{-0.5}\cdot\Sigma\cdot\Sigma^{-0.5} = \mathbf{I}\end{aligned} \tag{4}$$

Finally, the deconvolution operation is performed in terms of matrix multiplication among the deconvolved data matrix, where kernel w removes the correlations between both local neighbourhood pixels and across different channels. This can be formulated as:

$$y = (\mathcal{X}-\mu)\cdot\mathbf{D}\cdot w + b, \tag{5}$$

where b is the bias parameter. To generate the input to the next layer x_{i+1}, the same deconvolution operation in the i^{th} layer $\mathbf{D}_i$ is performed in the following way:

$$x_{i+1} = \Phi_i \circ W_i \circ \mathbf{D}_i \circ x_i \tag{6}$$

where $\circ$ is the right associated matrix multiplication operation, x_i is the input from $(i-1)^{th}$ layer, W_i is the weights in the i^{th} layer and Φ_i is the ReLU activation function [61].

Algorithm 1: Computing the Deconvolution Matrix

```
Data: C channel features $\mathcal{X} = [x_1, x_2, \dots, x_C] \in \mathcal{R}^C$
Result: Deconvolution matrix $\mathbf{D}$
while $(1 \le i \le C)$ do
    $\mathcal{X}_i = im2col(x_i)$ % according to Eq. (2);
end while
$\mathcal{X} = [\mathcal{X}_1, \mathcal{X}_2, \dots, \mathcal{X}_C]$ % column-wise concatenate;
$\widehat{\mathcal{X}} = Reshape(\mathcal{X})$ % columns are grouped on batch sizes;
$\Sigma = \frac{1}{S}\widehat{\mathcal{X}}^T\widehat{\mathcal{X}}$;
$\mathbf{D} = (\Sigma + \epsilon \cdot I)^{-\frac{1}{2}}$ % I is identity matrix & small value $\epsilon$;
```

B. PROPOSED ARCHITECTURE

Let the training set composed by n samples with C different classes denoted as $\{x_i, y_i\}_{i=1}^n$, where $y_i \in \{1, 2, \dots, C\}$ represents the corresponding label. Fig. 2 provides the graphical scheme of the proposed `SiCoDeF`^{2}`Net` architecture. In this context, the aims of the proposed model is to evaluate the similarity score between a pair of input images x_i and x_j, where the corresponding label of the image pair can be generated by the following target function:

$$SiCoDeF^2Net^{target}(x_i, x_j) = \begin{cases} 1 & \text{if } y_i = y_j, \\ 0 & \text{if } y_i \neq y_j. \end{cases} \tag{7}$$

The details of the proposed network is given step by step: As we can observe in Fig. 2, the `SiCoDeF`^{2}`Net` contains two identical networks, which are represented by $SiCoDeNet^U$

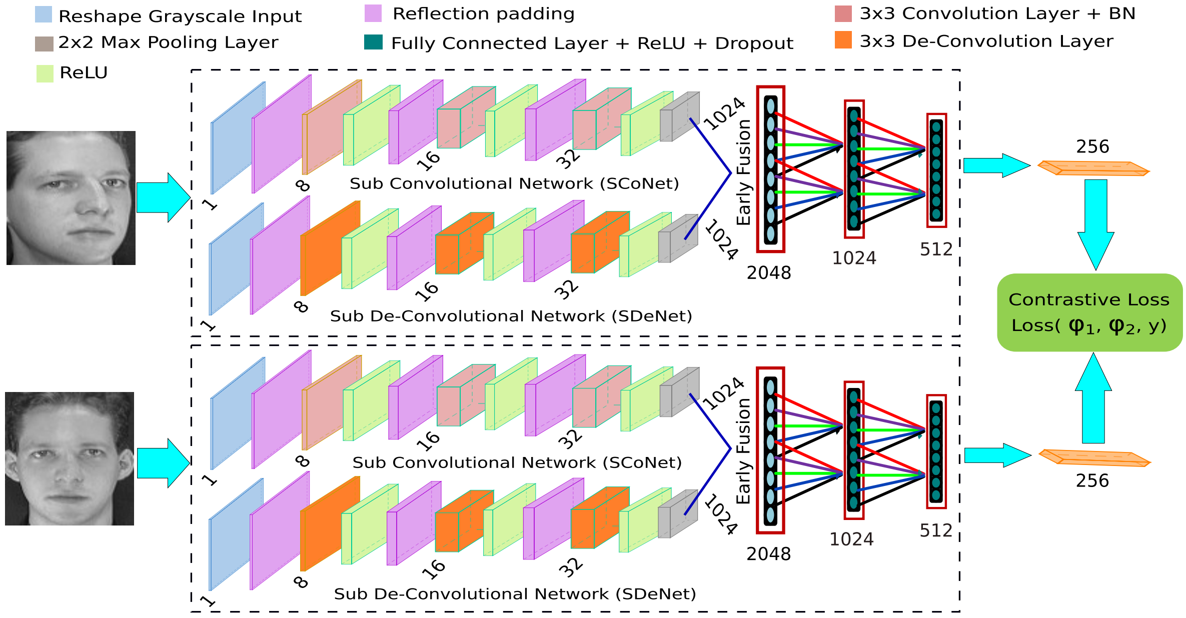


FIGURE 3. Overview of the proposed Siamese convolution deconvolution feature fusion network SiCoDeF^{2}Net.

and *SiCoDeNet*^{L}. Superscript 'U' and 'L' indicate upper and lower parts of the proposed network. To extract more robust and discriminative feature representations, both *SiCoDeNet*^{U} and *SiCoDeNet*^{L} further comprise twin networks, namely sub convolution network (*SCoNet*) and sub deconvolution network (*SDeNet*), respectively. Moreover, both of the networks *SiCoDeNet*^{U} and *SiCoDeNet*^{L} are designed to use the same parameter settings and share the same trainable weights throughout the twin networks [42]. The input is being processed through *SCoNet* to extract the 2-D convolutional features, while *SDeNet* helps to extract 2-D deconvolutional features, working as the counterpart (or complementary part) of the convolution. Furthermore, features from each path can be represented as

$$SiCoDeNet^U(x_i) = \begin{cases} H^U_{Co} = SCoNet(x_i) \\ H^U_{De} = SDeNet(x_i) \end{cases} \tag{8}$$

and

$$SiCoDeNet^L(x_j) = \begin{cases} H^L_{Co} = SCoNet(x_j) \\ H^L_{De} = SDeNet(x_j), \end{cases} \tag{9}$$

where $SCoNet(\cdot)$ and $SDeNet(\cdot)$ are the mapping function corresponding to 2-D *SCoNet* and 2-D *SDeNet*, respectively, and $H^{U/L}_{Co}$ and $H^{U/L}_{De}$ represent the output feature vectors of the twin networks, which are designed using 2-D convolution and deconvolution architectures, respectively. Both sub networks *SiCoDeNet*^{U} and *SiCoDeNet*^{L} update their weights in a mirror fashion during training stage. In particular, *SCoNet* and *SDeNet* are trained from scratch and thus both images are passed through the entire model (i.e., the two branches) in parallel, one into the *SiCoDeNet*^{U} and the another one into the *SiCoDeNet*^{L}, as Fig. 3 indicates. Furthermore, both networks contain the same number of convolution and deconvolution filters as reported in Table 1.

The discriminative latent embedded feature encoding vectors of each image are represented by $H^{U/L}_{Co}$ and $H^{U/L}_{De}$, as they are processed through the twin sub networks, *SCoNet* and *SDeNet*, respectively. Moreover, the corresponding output features Fe^U, and Fe^L of twin networks *SiCoDeNet*^{U} and *SiCoDeNet*^{L} can be derived by fusing both the output feature vectors $H^{U/L}_{Co}$ and $H^{U/L}_{De}$ as follows:

$$SiCoDeF^2Net(x_i, x_j) = \begin{cases} Fe^U = fusion(H^U_{Co}, H^U_{De}) \\ Fe^L = fusion(H^L_{Co}, H^L_{De}) \end{cases} \tag{10}$$

where $fusion(\cdot)$ represents two kinds of fusion strategies, namely 'early concatenation' and 'late concatenation', respectively. Based on that the proposed networks can be denoted by SiCoDeF^{2}Net^{early} and SiCoDeF^{2}Net^{late}, respectively. The superiority of both networks are validated through the experimental evaluations in Section IV.

At the top of twin SiCoDeF^{2}Net networks, a loss function is used to connect both models and to evaluate the similarity score between the embedded representation of both fused features Fe^U and Fe^L. This similarity score is based on the widely used Euclidean distance. Moreover, the contrastive loss [38] is such a loss function used in Siamese network and can be defined as follows:

$$\mathcal{L}oss(Fe^U, Fe^L, y) = \alpha(1-y)D_w^2 + \beta y \max(0, m - D_w)^2 \tag{11}$$

where Fe^U and Fe^L represent features of two samples, y is a binary valued function that indicates if both images are belonging to the same class or not, α and β are two constants and the margin m equal to 1 for the experiment. The Euclidean

TABLE 1. Details of layer-wise comparison between standard sub convolutional network (*SCoNet*) and the sub deconvolutional network (*SDeNet*).

SiConvNet						SiDeConvNet			
Layer	Kernel	Shape	Stride	BatchNorm	Activation	Layer	Kernel	Shape	Activation
Input		100 × 100				Input		100 × 100	
Reflection Padding	-	1 × 102 × 102	-	-	-	Reflection Padding	-	1 × 102 × 102	-
Convolution	8 × 3 × 3	8 × 100 × 100	1	Yes	ReLU	deconvolution	8 × 3 × 3	8 × 100 × 100	ReLU
Reflection Padding	-	8 × 102 × 102	-	-	-	Reflection Padding	-	8 × 102 × 102	-
Convolution	16 × 3 × 3	16 × 100 × 100	1	Yes	ReLU	deconvolution	16 × 3 × 3	16 × 100 × 100	ReLU
Reflection Padding	-	16 × 102 × 102	-	-	-	Reflection Padding	-	16 × 102 × 102	-
Convolution	32 × 3 × 3	32 × 100 × 100	1	Yes	ReLU	deconvolution	32 × 3 × 3	32 × 100 × 100	ReLU
Max Pool		32 × 50 × 50				Max Pool		32 × 50 × 50	
Fully Connected	Linear	1024			ReLU	Fully Connected	Linear	1024	ReLU
Dropout	0.5	1024			-	Dropout	0.5	1024	-
Fully Connected	Linear	512			ReLU	Fully Connected	Linear	512	ReLU
Fully Connected	Linear	128				Fully Connected	Linear	128	

distance D_w is computed based on the embedded feature space Fe^U and Fe^L, and it defined as:

$$\begin{aligned} D_w &= ||Fe^U - Fe^L|| \\ &= ||SiCoDeF^2Net(x_1, w_1) \\ &\quad - SiCoDeF^2Net(x_2, w_2)||_2, \end{aligned} \tag{12}$$

where $SiCoDeF^2Net(\cdot)$ represents the function mapped into a real valued embedded representation of a pair of sample images, x_i and x_j, when passed through the twin networks, $SiCoDeNet^U$ and $SiCoDeNet^L$, respectively, while w_1 and w_2 indicates the learned weights parameters during training through the underlying networks. In this sense, SiCoDeF^{2}Net aims to model the output embedded feature vector adjacent to each others in the low dimensional metric space when both images belong to the same class, and taken far away when both images do not belong to the same or similar class. In contrast to conventional Siamese convolutional networks, the removal of both pixel-wise and channel-wise correlation conducted by the *deconvolution* operation helps the proposed model SiCoDeF^{2}Net to learn more robust and discriminative complementary feature representation through the *fusion*. Moreover, both branches of the proposed network $SiCoDeNet^U$ and $SiCoDeNet^L$ can better approximate the images into an embedded mapping space through the $SiCoDeF^2Net(\cdot)$ model. Hence, to evaluate the contrastive loss in Eq. (11), the Euclidean distance between Fe^U and Fe^L outputs is computed. This plays an important role in bringing the embedded space close to each others, evaluating the dissimilar value close to *zero* when both instances belong from same class and obtaining the dissimilarity value larger than *one*.

In order to introduce clarity in classification, one has to determine the threshold value 1 in order to decide the pair of instances belonging to same or different classes. Fig. 4 shows the dissimilarity score evaluated on test set based on the Euclidean distance. It can be observed that scores lower than 1 are produced for similar instances, whilst scores higher than 1 are produced for images taken from AT&T and Yale face datasets. Once the whole network is trained by imposing the contrastive loss, the network computes the distance based dissimilarity score first on (x, x_i) to evaluate a test image x for all possible x_i, then it predicts the label of x on the dissimilarity score τ which is further thresholded by 1 as:

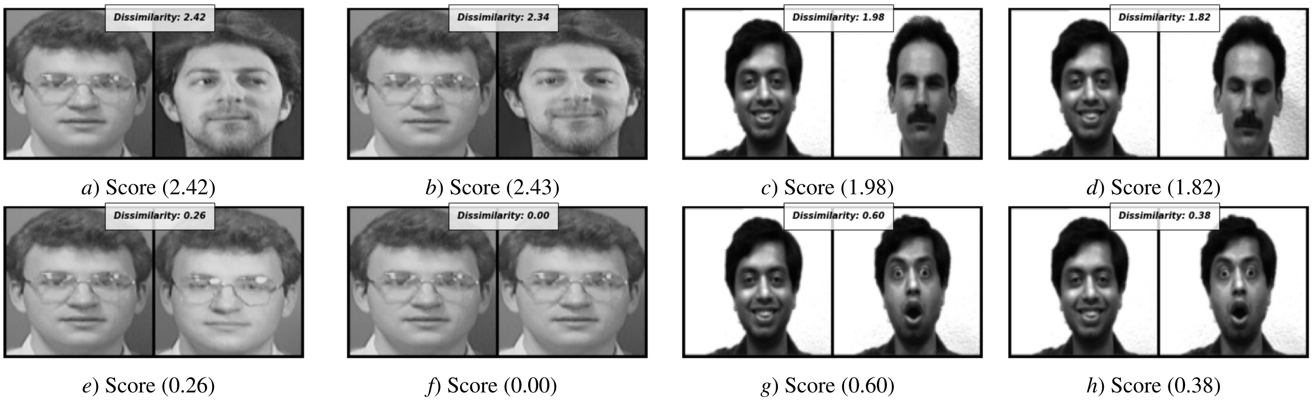
$$SiCoDeF^2Net^{target}(x, x_i) = \begin{cases} 0, & if\ \tau \geq 1 \\ 1, & else. \end{cases} \tag{13}$$

Table 1 details the layer-wise summary, providing also the size of convolution/deconvolution kernels, shape of the feature maps, the regularization technique and non-linearities used in the baseline network SiConvNet [42] and the SiDeConvNet, respectively. In addition to this, Fig. 5 provides the convergence of loss using *SiConvNet*, *SiDeConvNet*, SiCoDeF^{2}Net^{early}, and SiCoDeF^{2}Net^{late} networks for AT&T dataset, while Fig. 6 depicts the graphical visualization of the convolution and deconvolution filter banks in the second layer extracted over AT&T dataset during the training stage.

IV. EXPERIMENTAL RESULTS

This paper is mainly focused on unconstrained conditional facial recognition with the availability of few training examples. In this context, and with the aim of evaluating the performance of the proposed one-shot learning framework SiCODeF^{2}Net with the current state-of-art methods in a comprehensive manner, it has been compared with five classification models, in particular: i) standard CNN [44], ii) Bilinear-CNN [62], [63], iii) Pretrained-VGG [36], [64], iv) LightCNN [32] and v) Pretrained-VGG+KNN (VGG+KNN) [36], [64]. Moreover, two variants of the one-shot learning framework are compared, in particular the Siamese convolutional network (SiConvNet) [42], and Siamese deconvolutional network (SiDeConvNet). We have also compared the proposed model with several metric learning approaches, such as the large margin nearest Neighbor (LMNN) [41], information theoretic metric learning (ITML) [40], least squares metric learning (LSML) [65], and Mahalanobis metric for clustering (MMC) [66] based on the extracted VGGFace2 Pre-trained features [67].

Experiments have been conducted using five different and widely used benchmark datasets, which includes AT&T, Yale, extended Yale-B, LFW and UFI cropped face datasets. Furthermore, all the experiments have been performed with Ubuntu 18.04LTS operating system and NVIDIA Titan V 12-GB graphics processing unit. The training is conducted



a) Score (2.42) *b*) Score (2.43) *c*) Score (1.98) *d*) Score (1.82)

e) Score (0.26) *f*) Score (0.00) *g*) Score (0.60) *h*) Score (0.38)

FIGURE 4. Dissimilarity values among the instances of different subjects and the instances of similar subjects randomly taken from (a)-(b) and (e)-(f) AT&T dataset and (c)-(d) and (g)-(h) Yale dataset.

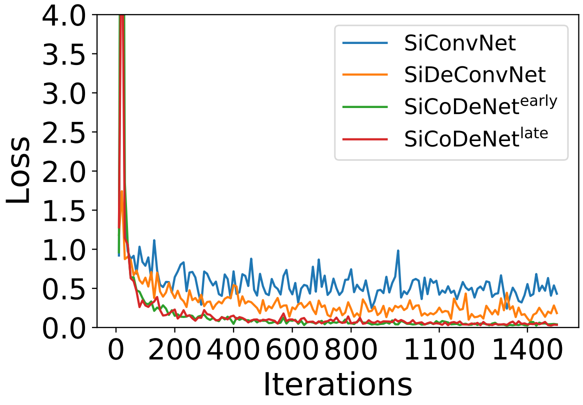
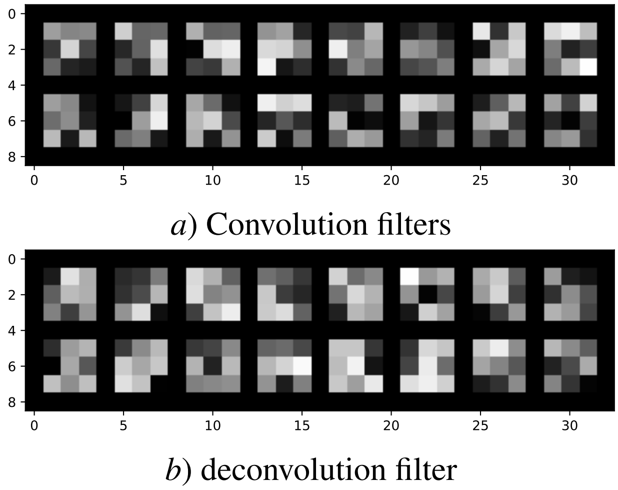


FIGURE 5. Graphical evaluation of loss using *SiConvNet*, *SiDeConvNet*, SiCoDeF^{2}Net^{early}, and SiCoDeF^{2}Net^{late} for AT&T dataset.



a) Convolution filters

b) deconvolution filter

FIGURE 6. Visualizing the convolutional and deconvolutional filters from the second layers of SCoNet and SDeNet network which are trained on AT&T dataset.

5 times, where each one conducts 200 epochs, with batch of size 64. To evaluate the performance, the mean accuracy of the model is reported. The cosine annealing scheduler is used to update the learning rate, which has initially set to 0.0005, with a momentum value of 0.9. The network parameters have also been optimized through Adam optimizer [68] during training stage. The details about the datasets are described below.

A. FACE RECOGNITION DATASETS

The **AT&T** dataset [69] contains 40 different subjects, where each one comprises 10 examples of 92 × 112 pixels, with 256 grey levels per pixel. Moreover, the images were captured at different times, varying the lighting, facial expressions (open and closed eyes, smiling and not smiling) and facial details (with glasses and no glasses). All the images were taken against a dark homogeneous background with the subjects in an upright, frontal position (with tolerance for some side movement).

The **Yale** face dataset [14] contains 165 grayscale images of 15 different subjects. There are 11 images per subject which were captured from different settings such as center-light, with glasses, happy, left-light, with no glasses, normal, right-light, sad, sleepy, surprised, and wink, respectively. The **extended Yale-B** contains images of 38 different subjects, comprising a total of 2432 images under 64 different illumination condition [70], [71]. The dataset is further divided into 5 groups based on the illumination angles where Group-1 includes 7 images per subject from 0° to 12°, Group-2 includes 12 images per subject from 13° to 25°, Group-3 includes 12 images per subject from 26° to 50°, Group-4 includes 14 images per subject from 51° to 77° and Group-5 includes 19 images in each subject on and above 78°. It is observed that those images of Group-4 and Group-5 are the most challenging and difficult to classify.

The **UFI-Cropped** dataset [72] contains images of 605 subjects with an average of 7.1 images per person in the training set and one in the test set, where images are cropped into a size of 128 × 128 pixels.

Finally, the **LFW** face dataset [73] contains 13233 images from 5749 different people where 1680 among them have two or more different photos. These images were collected from the web and were processed (i.e., detected and centered) by the Viola Jones face detector.

TABLE 2. Classification results and training time in ms per sample using all the models on (a) AT&T, (b) YALE, (c) UFI-Cropped and (d) LFW face datasets.

(a) AT&T Dataset

Metrics	CNN	Bilinear-CNN	Pretrained-VGG	LightCNN	VGG+KNN	SiConvNet	SiDeConvNet	SiCoDeNet^{early}	SiCoDeNet^{late}
Accuracy	93±1.5%	94±1.5%	93±1%	94±1.5%	94±2%	93.33±1.05%	96.66±1.23%	97.82±0.35%	**98.54**±0.37%
F1	0.91	0.92	93	0.93	93	0.923	0.9696	0.9798	**0.9831**
Precision	0.91	0.92	92	0.92	0.93	0.9	0.9411	0.9524	**0.9682**
Recall	0.91	0.92	92	0.92	0.92	0.9523	1	1	1
Time	8	14	17	19	24	**4.5**	4.7	6.3	6.1

(b) Yale Dataset

Metrics	CNN	Bilinear-CNN	Pretrained-VGG	LightCNN	VGG+KNN	SiConvNet	SiDeConvNet	SiCoDeNet^{early}	SiCoDeNet^{late}
Accuracy	66±3%	70±2%	78±3%	80±1.5%	82±2%	90.9±1.29%	93.45±1.11%	95.28±0.47%	**96.47**±0.26%
F1	0.72	0.74	79	0.82	0.84	0.909	0.93	0.9512	**0.9623**
Precision	0.65	0.69	0.73	0.75	0.77	1	1	1	1
Recall	0.65	0.69	0.73	0.75	0.77	0.9	0.923	0.9423	**0.9513**
Time	12	18	21	24	28	**5.3**	5.6	7.4	7.2

(c) Extended Yale-B Dataset

Metrics	CNN	Bilinear-CNN	Pretrained-VGG	LightCNN	VGG+KNN	SiConvNet	SiDeConvNet	SiCoDeNet^{early}	SiCoDeNet^{late}
Accuracy	63±1.5%	67±2%	75±1.5%	77±3%	79±1.5%	83.41±2.67%	86.52±2.54%	87.58±3.25%	**88.08**±3.12%
F1	0.66	0.70	0.79	0.80	0.83	0.8632	0.8907	0.8976	**0.9029**
Precision	0.61	0.66	0.74	0.75	0.77	0.7829	0.8013	0.8079	**0.823**
Recall	0.61	0.66	0.74	0.75	0.77	0.9024	0.938	0.9689	**0.9769**
Time	19	26	28	32	36	**9.2**	9.5	12.1	12

(d) UFI-Cropped Dataset

Metrics	CNN	Bilinear-CNN	Pretrained-VGG	LightCNN	VGG+KNN	SiConvNet	SiDeConvNet	SiCoDeNet^{early}	SiCoDeNet^{late}
Accuracy	48±3%	53±2%	61±3%	62±1.5%	64±1%	75.77±2.3%	77.67±1.95%	79.54±1.87%	**79.92**±2.04%
F1	0.53	0.55	0.65	0.67	0.70	0.771	0.8131	0.8376	**0.8456**
Precision	0.48	0.53	0.57	0.59	0.61	0.7015	0.7323	0.7512	**0.7578**
Recall	0.48	0.53	0.57	0.59	0.61	0.8869	0.9112	0.9345	**0.9467**
Time	42	58	78	87	96	**20.1**	21	25.6	25.2

(d) LFW Dataset

Metrics	CNN	Bilinear-CNN	Pretrained-VGG	LightCNN	VGG+KNN	SiConvNet	SiDeConvNet	SiCoDeNet^{early}	SiCoDeNet^{late}
Accuracy	65±2%	69±2.5%	85±2%	86±1.5%	89±1%	92.93±2.34%	94.24±1.55%	97.50±0.38%	**97.81**±0.25%
F1	0.67	0.70	0.86	0.88	0.90	0.9232	0.9474	0.9751	**0.9779**
Precision	0.64	0.68	0.84	0.85	0.88	0.9223	0.9417	0.9717	**0.9791**
Recall	0.64	0.68	0.84	0.85	0.88	0.9288	0.9475	0.9751	**0.9781**
Time	54	76	101	132	151	**31**	32.2	42	41.3
Parameters	53M	106M	138M	**6M**	138M	165M	165M	333M	330M

TABLE 3. Classification results obtained by metric learning approaches using VGGFace2 pre-trained features on AT&T, YALE, extended YALE-B, UFI cropped, and LFW face datasets.

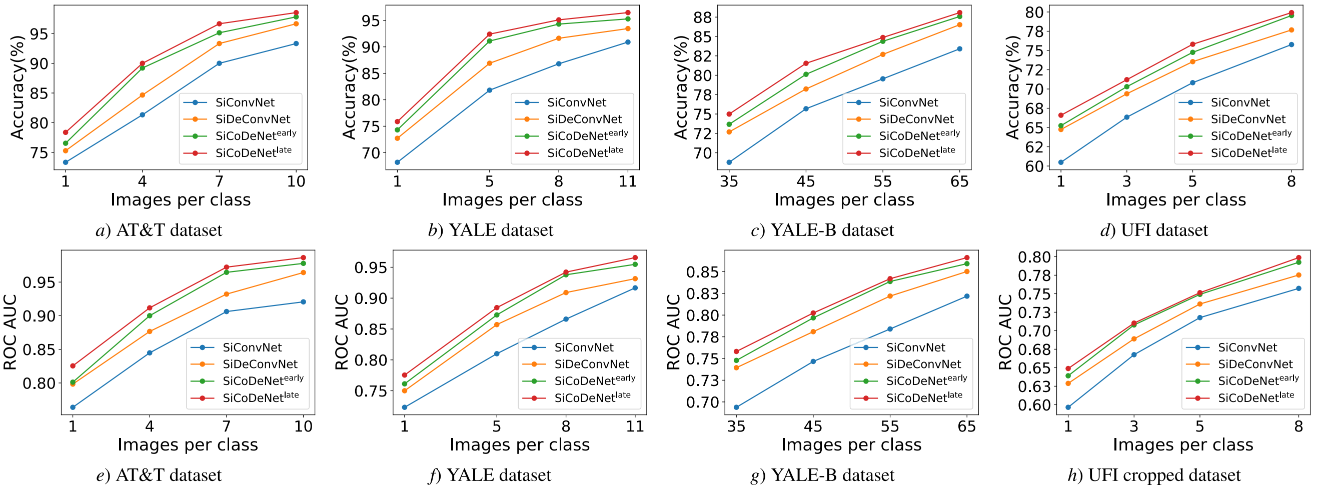
Datasets	KNN	LMNN	ITML	LSML	MMC	SiCoDeNet^{early}	SiCoDeNet^{late}
AT&T	94.78±0.54%	95.46±0.47%	95.21±0.61%	95.14±0.23%	93.84±0.38%	97.28±0.35%	**98.54**±0.37%
Yale	89.12±0.41%	89.79±0.69%	89.71±0.78%	89.34±0.57%	88.67±0.51%	95.28±0.47%	**96.47**±0.26%
Ex. Yale-B	80.35±3.62%	81.54±3.21%	81.43±3.14%	81.35±3.78%	79.77±3.29%	87.58±3.25%	**88.02**±3.12%
UFI	71.83±1.69%	72.23±2.14%	72.19±2.03%	72.18±1.89%	71.54±1.97%	79.54±1.87%	**79.92**±2.04%
LFW	92.24±2.04 %	92.86±1.74%	92.64±1.45%	92.37±2.14%	91.77±1.67%	97.50±0.38%	**97.81**±0.25%

TABLE 4. Impact of activations on the proposed `SiDeConvF`^{2}`Net` using AT&T, Yale, Ex. Yale-B, UFI cropped, and LFW face datasets.

datasets	ReLU	PReLU	Mish	LiSHT
AT&T	98.54±0.37%	98.54±0.37%	98.52±0.42%	**98.50**±0.45%
Yale	96.47±0.26%	96.47±0.26%	96.46±0.3%	**96.46**±0.3%
Ex. Yale-B	88.02±3.12%	88.02±3.12%	88.02±3.02%	**88.01**±3.1%
UFI	79.92±2.04%	79.92±2.04%	79.88±2.05%	**79.89**±2.08%
LFW	97.81±0.25 %	97.81±0.25%	97.78±0.23%	**97.77**±0.33%

B. EXPERIMENTAL SETTINGS

In order to perform unbiased experiments, we have split the entire AT&T dataset into training and testing sets. Additionally, we have randomly chosen 37 classes from the 40 available classes to create the training set, while the remaining 3 classes are used for evaluating the performance of `SiCoDeF`^{2}`Net` model. Similarly, we have divided the Yale face dataset by randomly selecting 13 classes for the training set and the remaining 2 classes were used to create the testing set. There are a total of 38 classes in the extended Yale-B dataset, of which 35 have been randomly chosen for the training set, while the remaining 3 classes are used for the testing set. The UFI-Cropped dataset comprises 605 classes, of which 560 classes have been randomly selected and used in the training set and the remaining 45 classes have been considered into the testing set. Finally, regarding the LFW dataset, 5000 classes have been randomly selected from the 5749 available classes to create the training set, and the remaining 749 classes have been used to evaluate the proposed model during testing. All the images have been reshaped into 100 × 100 to feed the network. For this purpose, the disjoint training-testing strategy has been considered, i.e., all the samples of the training classes have been considered for training the networks, while the rest of the classes are used for testing. In this sense, there is no test information during the training, which makes the evaluation even more challenging and interesting in order to check the generalization of the model.



a) AT&T dataset

b) YALE dataset

c) YALE-B dataset

d) UFI dataset

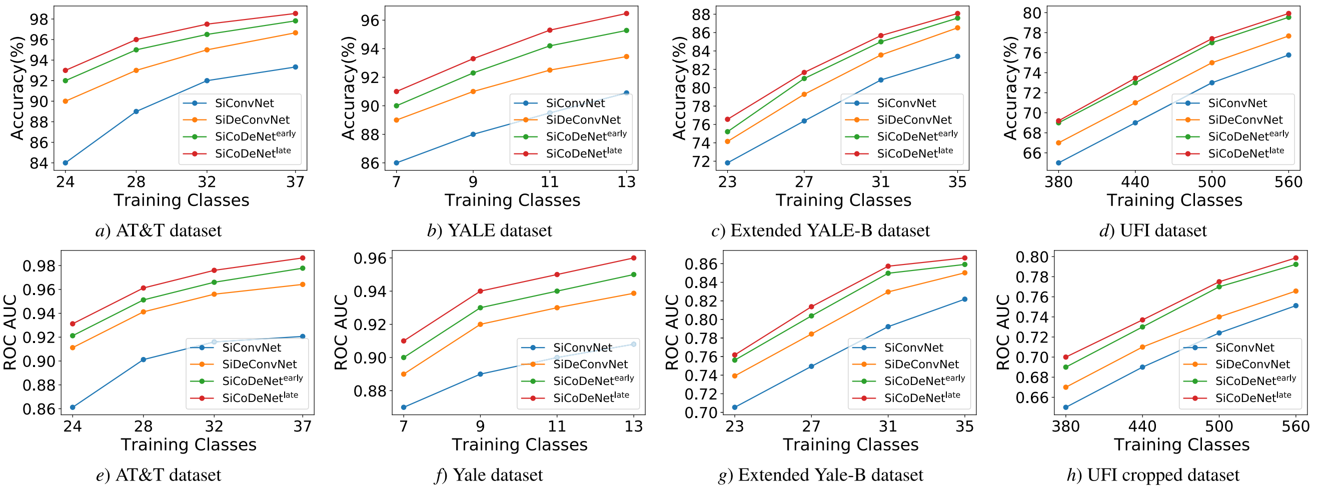
e) AT&T dataset

f) YALE dataset

g) YALE-B dataset

h) UFI cropped dataset

FIGURE 7. Evolution of Accuracy vs. Images-per-class shown for (a) AT&T (b) YALE (c) UFI-Cropped datasets and ROC vs. Images-per-class shown for (d) AT&T (e) YALE (f) UFI-Cropped datasets.



a) AT&T dataset

b) YALE dataset

c) Extended YALE-B dataset

d) UFI dataset

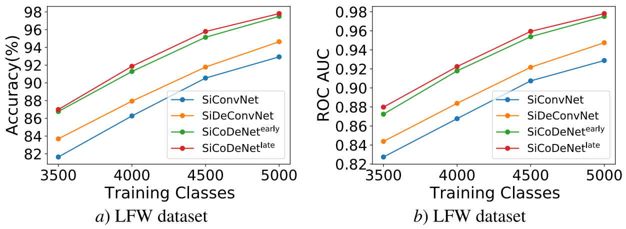
e) AT&T dataset

f) Yale dataset

g) Extended Yale-B dataset

h) UFI cropped dataset

FIGURE 8. Evolution of accuracy vs. Images-per-class shown for (a) AT&T (b) YALE (c) UFI-Cropped datasets and ROC vs. Images-per-class shown for (d) AT&T (e) YALE (f) UFI-Cropped datasets.



a) LFW dataset

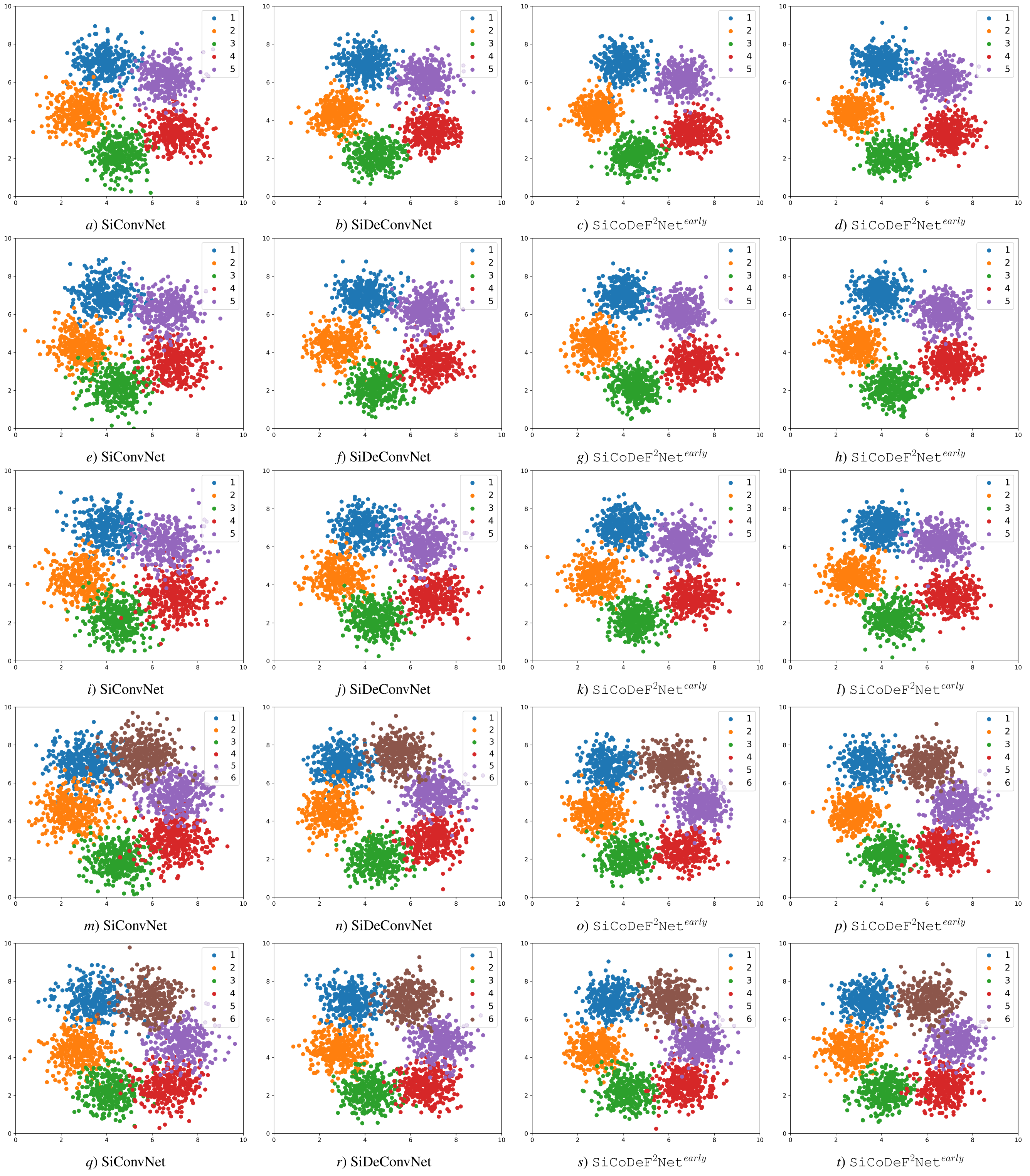
b) LFW dataset

FIGURE 9. Evolution of accuracy vs. Images-per-class (a) and ROC vs. Images-per-class (b) shown for LFW dataset.

C. PERFORMANCE ANALYSIS

1) COMPARISON WITH OTHER DEEP LEARNING CLASSIFIERS

One shot learning classification is domain specific and quite effective when there are few training samples available for training and the number of classes are not known during training. The conventional *SiConvNet* performs well in the described scenario. However, due to the pixel and channel wise correlations into the output feature maps, the *SiConvNet* model fails to generalize the feature representation and hence, scarifies the recognition performance up to some extends. To learn invariant and discriminative complementary feature representation, the proposed network `SiCoDeF`^{2}`Net` combines both the convolutional and deconvolutional features at the top of the network, boosting the recognition performance. Table 2 reports the obtained classification results in terms of accuracy, F1, Precision and achieved Recall using `SiCoDeF`^{2}`Net` and comparing it with CNN, Bilinear-CNN, Pretrained-VGG, LightCNN, VGG+KNN, *SiConvNet*, and *SiDeConvNet*, respectively. It can be seen that the proposed model significantly outperforms all the classification models, achieving state-of-the-art results for every dataset. In addition, the number of trainable parameters is also reported for every classification model, including the baseline networks and the proposed networks. As we can observe, our proposed network contains more parameters as it includes two subnetworks. In addition to



a) SiConvNet

b) SiDeConvNet

c) SiCoDeF^{2}Net^{early}

d) SiCoDeF^{2}Net^{early}

e) SiConvNet

f) SiDeConvNet

g) SiCoDeF^{2}Net^{early}

h) SiCoDeF^{2}Net^{early}

i) SiConvNet

j) SiDeConvNet

k) SiCoDeF^{2}Net^{early}

l) SiCoDeF^{2}Net^{early}

m) SiConvNet

n) SiDeConvNet

o) SiCoDeF^{2}Net^{early}

p) SiCoDeF^{2}Net^{early}

q) SiConvNet

r) SiDeConvNet

s) SiCoDeF^{2}Net^{early}

t) SiCoDeF^{2}Net^{early}

FIGURE 10. 2-D feature visualization via t-SNE, samples are represented in points and different samples are shown with different colored generated through *SiConvNet*, *SiDeConvNet*, SiCoDeF^{2}Net^{early}, and SiCoDeF^{2}Net^{early} on (a)-(d) AT&T (e)-(h) Yale (i)-(l) Extended Yale-B and (m)-(p) UFI-Cropped (q)-(t) LFW face datasets, respectively.

this, Table 2 reports the training times (in terms of ms per sample) for all datasets using each model. It can be observed that the time taken by the proposed network is comparable or even better as compared to the CNN, Bilinear-CNN, Pretrained-VGG, VGG+KNN and LightCNN, respectively.

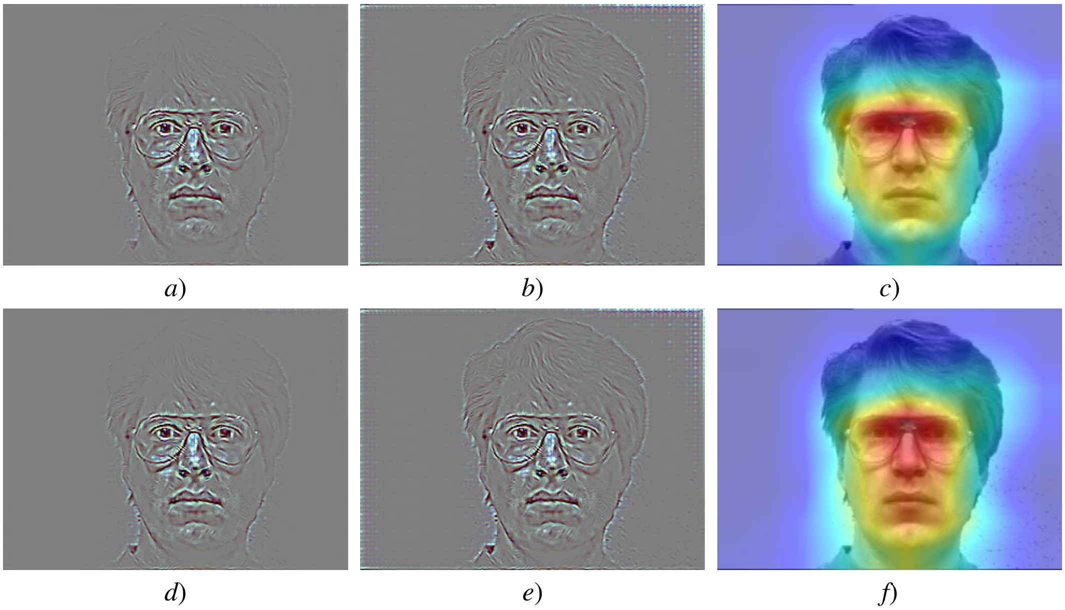


FIGURE 11. Gradcam visualization: (a)-(c) provides the visualization of convolutional features and (d)-(f) illustrate the visualization of deconvolution features.

TABLE 5. KL-divergences between the twin network SiCoDeNet^U(P) and SiCoDeNet^L(Q) on AT&T dataset.

Image pairs	Network	Distribution	KL-Divergence (P||Q)			
			Minimun fusion	Maximum fusion	Early concatenate	Late concatenate
Identity	SiCoDeNet^U	SCoNet (P) SDeNet (Q)	0.0976	0.0823	0.0684	0.0423
	SiCoDeNet^L	SCoNet (P) SDeNet (Q)	0.0954	0.0874	0.0677	0.0511
Different	SiCoDeNet^U	SCoNet (P) SDeNet (Q)	0.1124	0.1132	0.1268	0.1384
	SiCoDeNet^L	SCoNet (P) SDeNet (Q)	0.1145	0.1157	0.1250	0.1368

2) COMPARISON WITH OTHER DEEP METRIC LEARNING MODELS

Moreover, to show the effectiveness of the proposed SiCoDeF^{2}Net, Table 3 reports the achieved recognition rates considering different metric learning approaches, in particular KNN, LMNN, ITML, LSML, and MMC where VGGFace2 pre-trained features are extracted to train the models. It can be observed that the performance of metric learning approaches are better than the classification models on Table 2, and sometimes they are comparable with the conventional *SiConvNet*. However, our proposed networks *SiCoDeNet^{early}* and *SiCoDeNet^{late}* achieve the best classification performances, in particular the *SiCoDeNet^{late}* model reaches the best results, obtaining between 1 and 3 percentage points more than traditional methods.

3) ROBUSTNESS EVALUATION BY CHANGING THE NUMBER OF IMAGES PER CLASS AND THE NUMBER OF CLASSES

In order to assess the robustness of the proposed model, some experiments have been performed by varying the number of images per class, where *SiConvNet*, and *SiDeConvNet* have been compared with other two versions of the proposed model, i.e. SiCoDeF^{2}Net^{early} and SiCoDeF^{2}Net^{late}. Moreover, these experiments have been conducted over the described datasets considering the modified experimental setting discussed in SubSection IV-B. In this sense, Figs. 7(a)-(d) illustrate the accuracy curve while Figs. 7(e)-(h) provide the Receiver Operating Characteristic (ROC) curve considering: 1, 4, 7 and 10 images per class from AT&T dataset; 1, 5, 8, and 11 images per class from Yale dataset; 35, 45, 55 and 65 images per class from extended Yale-B faces collection, and 1, 3, 5 and 8 images from each class of UFI-cropped dataset, respectively.

In addition, to explore the capability of the one-shot learning classification provided by the proposed networks, similar experiments have been conducted by changing the number of classes during the training stage, while the classification results are evaluated considering the remaining classes. In this sense, Figs. 8(a)-(d) illustrate the accuracy curve while Figs. 8(e)-(h) illustrate the ROC curve considering: 24, 28, 32 and 37 classes from AT&T dataset; 7, 9, 11 and 13 classes from Yale collection; 23, 27, 31, 35 classes from extended Yale-B face dataset, and 380, 440, 500, 560 classes from UFI-cropped dataset, respectively. The accuracy and ROC curves for LFW dataset are shown in Figs. 9(a)-(b) considering different numbers of classes for the training of the network, particularly 3500, 4000, 4500, and 5000 different classes. As pointed before, the remaining classes are used to evaluate the models performance during testing.

The results of both experiments show for all the cases that, even with very small amount of training samples and big number of unseen classes during training, the proposed models achieve good classification results in terms of both accuracy rates and ROC measurement, respectively. Due to the pixel and channel wise redundancies, the *SiConvNet* performs slightly worse than *SiDeConvNet*. However, both

TABLE 6. Triplet loss impacts using the baseline Siamese networks and both the proposed SiCoDeF^{2}Net^{early} and SiCoDeF^{2}Net^{late} on AT&T, Yale, Extended Yale-B and UFI Cropped face datasets.

Metrics	AT&T dataset				Yale dataset			
	SiConvNet	SiDeConvNet	SiCoDeNet^{early}	SiCoDeNet^{late}	SiConvNet	SiDeConvNet	SiCoDeNet^{early}	SiCoDeNet^{late}
Accuracy	73.64±2.95%	78.23±3.15%	80.14±2.87%	**80.79**±3.12%	71.4±3.27%	74.21±3.54%	75.67±2.5%	**75.94**±2.34%
F1	0.7182	0.7633	0.7924	**0.8011**	0.684	0.7238	0.7441	**0.7587**
Precision	0.7056	0.7545	0.7735	**0.7829**	0.7124	0.7434	0.7513	**0.7624**
Recall	0.7535	0.8024	0.8265	**0.8358**	0.7012	0.7387	0.7434	**0.7513**
ROC AUC	0.7488	0.7913	0.8126	**0.8194**	0.7113	0.7454	0.7544	**0.7614**

Metrics	Extended Yale-B dataset				UFI dataset			
	SiConvNet	SiDeConvNet	SiCoDeNet^{early}	SiCoDeNet^{late}	SiConvNet	SiDeConvNet	SiCoDeNet^{early}	SiCoDeNet^{late}
Accuracy	74.21±1.89%	76.35±1.67%	78.77±2.5%	**79.14**±2.31%	70.23±3.83%	72.74±2.14%	73.52±3.85%	**74.91**±4.12%
F1	0.7538	0.7744	0.7834	**0.7946**	0.7214	0.7424	0.7554	**0.7618**
Precision	0.7121	0.7374	0.7518	**0.7623**	0.6522	0.6719	0.6827	**0.6914**
Recall	0.8077	0.8428	0.8777	**0.8818**	0.8214	0.8429	0.8614	**0.8729**
ROC AUC	0.7344	0.7537	0.7749	**0.7928**	0.7077	0.7267	0.7344	**0.7457**

networks SiCoDeF^{2}Net^{early} and SiCoDeF^{2}Net^{late} significantly outperform the baseline represented by conventional *SiConvNet* and *SiDeConvNet*, respectively.

4) ROBUSTNESS EVALUATION CONSIDERING DIFFERENT ACTIVATION FUNCTIONS

In order to evaluate the robustness of the proposed SiCoDeF^{2}Net model under different activation functions, an experiment has been conducted considering *ReLU* [61], *PReLU*, [74] *Mish* [75], and *LiSHT* [76] activation functions, employing the same network architecture. Table 4 reports the achieved results using AT&T, Yale, Extended Yale-B, UFI Cropped, and LFW face datasets. It can be observed that the proposed SiCoDeF^{2}Net model achieves similar performance for every activation function, so we can infer that the proposed network is quite independent of the activation function considered within the architecture.

5) FEATURE REPRESENTATION EVALUATION

To intuitively illustrate the advantages of the extracted deconvolutional feature map over the standard convolutional feature maps, we applied the Grad-CAM [77] to provide the visualization of the obtained features. It can also be readily observed from Fig. 11 that the deconvolution has more discriminative feature representations than the convolution. Similarly, in order to visualize and evaluate the discriminative power of the feature representation obtained by the proposed SiCoDeF^{2}Net model, Figs. 10(a)-(d) provide the graphical representation using t-SNE visualization [78]. As we can observe, SiConvNet, SiDeConvNet, SiCoDeF^{2}Net^{early}, and SiCoDeF^{2}Net^{late} have been tested over AT&T dataset. In a similar way, Figs. 10(e)-(h) depict the obtained feature representations for Yale collection, Figs. 10(i)-(l) provide the graphical visualization for Extended Yale-B dataset, Figs. 10(m)-(p) depict the obtained representation for UFI face dataset, and finally, Figs. 10(q)-(t) provide the obtained representation for LFW face dataset. It can be clearly observed that the obtained features are compact, invariant and more separable in comparison with those obtained by the baseline networks *SiConvNet* and *SiDeConvNet*. In this sense, the great separability of test feature representation is one of the paramount reasons for the success of our proposed model, which reaches better performance than the current state-of-art methods.

TABLE 7. Triplet loss impacts using the baseline Siamese networks and both the proposed SiCoDeF^{2}Net^{early} and SiCoDeF^{2}Net^{late} on LFW face dataset.

Metrics	LFW face Dataset			
	SiConvNet	SiDeConvNet	SiCoDeNet^{early}	SiCoDeNet^{late}
Accuracy	90.21±0.89%	91.28±1.06%	93.02±0.67%	**93.64**±0.44%
F1	0.9044	0.9157	0.9322	**0.9384**
Precision	0.9019	0.9166	0.9285	**0.9312**
Recall	0.9114	0.9152	0.9354	**0.9404**
ROC AUC	0.9033	0.9185	0.9319	**0.9377**

D. EFFECTIVENESS OF COMPLEMENTARY FEATURES

To measure the degree of dissimilarity between the generated features from the *SCoNet* and *SDeNet* subnetworks of the twin networks *SiCoDeNet*^{U} and *SiCoDeNet*^{L}, we have obtained the relative entropy using an asymmetrical Kullback-Leibler (KL) divergence measurement. In particular, the KL divergence between the distribution *SCoNet* (P) and *SDeNet* (Q) on the same probability space X can be defined by the following asymmetrical function:

$$D_{KL}(P||Q) = \sum_{x \in X} P(x) \log \frac{P(x)}{Q(x)}, \tag{14}$$

which is based on the constraint $D_{KL}(P||Q) \neq D_{KL}(Q||P)$.

Table 5 provides the dissimilarity values calculated from the fused (minimum, maximum, early and late concatenations) feature distribution of the *SCoNet* and *SDeNet* subnetworks, using both *SiCoDeNet*^{U} and *SiCoDeNet*^{L} networks on a similar and dissimilar face pairs extracted from AT&T face dataset. Among the various fusion techniques, early and late concatenations achieve the score closest to zero when compared with similar pair and furthest away from zero for the dissimilar face pair. These results support the previous experiments, where the late concatenation performs better than others for all the face datasets, providing an interesting information theoretic reason behinds the obtained results.

E. CONTRASTIVE LOSS VS. TRIPLET LOSS

In order to study the impact of loss function in the proposed SiCoDeF^{2}Net model, some experiments have been conducted considering the contrastive and the triplet loss. Indeed, these functions have been evaluated using the baseline networks, i.e. *SiConvNet*, and *SiDeConvNet*, on AT&T, Yale, Extended Yale-B, UFI and LFW face datasets. Table 6 shows the obtained performance of these models, comparing the obtained results with those provided by the proposed networks $\texttt{SiCoDeF}^2\texttt{Net}^{early}$, and $\texttt{SiCoDeF}^2\texttt{Net}^{late}$ using triplet losses on AT&T, Yale, Extended Yale-B and UFI face datasets. It can be seen from Table 6 that the contrastive loss performs significantly better than the triplet loss for all the face datasets. This is mainly due to the important lack of training samples that are available during training. In addition, Table 7 shows the obtained performance of these models on LFW face dataset, the behavior of which is quite similar to that observed earlier in AT&T, Yale, Extended Yale-B and UFI face datasets. This inspires us to evaluate both the proposed $\texttt{SiCoDeF}^2\texttt{Net}^{early}$ and $\texttt{SiCoDeF}^2\texttt{Net}^{late}$ networks using contrastive loss as shown in Eq. (11).

V. CONCLUSION

In this paper, we have proposed a simple but efficient Siamese convolution-deconvolution feature fusion network ($\texttt{SiCoDeF}^2\texttt{Net}$) to learn invariant and discriminative complementary features from two subnetworks, i.e. *SCoNet* and *SDeNet*, following a feature fusion strategy at the top of the network for one-shot face classification. Through a comprehensive experimentation over different face-datasets, evaluating also different classification measurements, it can be observed that, although both networks share the same architecture, the *deconvolution* operation in *SDeNet* can successfully replace the widely used convolution and batch normalization operations of the conventional *SCoNet*, reaching an outstanding performance during classification. Moreover, the proposed networks $\texttt{SiCoDeF}^2\texttt{Net}^{early}$ and $\texttt{SiCoDeF}^2\texttt{Net}^{late}$ can successfully learn the convolution and deconvolution features, whilst significantly outperforming the results of widely used current state-of-art classifiers for the considered face datasets.

REFERENCES

[1] D. H. Salvadeo, N. D. Mascarenhas, J. Moreira, A. L. Levada, and D. C. Corrêa, "RBPCA MaxLike: A novel statistic classifier for face recognition based on block-based PCA and covariance matrix regularization," in *Proc. 17th Int. Conf. Syst., Signals Image Process.*, 2010, pp. 69–72.

[2] I. Masi, Y. Wu, T. Hassner, and P. Natarajan, "Deep face recognition: A survey," in *Proc. 31st SIBGRAPI Conf. Graph., Patterns Images (SIBGRAPI)*, Oct. 2018, pp. 471–478.

[3] S. Lawrence, C. L. Giles, A. C. Tsoi, and A. D. Back, "Face recognition: A convolutional neural-network approach," *IEEE Trans. Neural Netw.*, vol. 8, no. 1, pp. 98–113, Jan. 1997.

[4] R. Ibrahim and Z. M. Zin, "Study of automated face recognition system for office door access control application," in *Proc. IEEE 3rd Int. Conf. Commun. Softw. Netw.*, May 2011, pp. 132–136.

[5] B. Kamgar-Parsi, W. Lawson, and B. Kamgar-Parsi, "Toward development of a face recognition system for watchlist surveillance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 10, pp. 1925–1937, Oct. 2011.

[6] C. Conde, I. M. de Diego, and E. Cabello, "Face recognition in uncontrolled environments, experiments in an airport," in *Proc. Int. Conf. E-Bus. Telecommun.* Berlin, Germany: Springer, Jul. 2011, pp. 20–32.

[7] Y. Shen, M. Yang, B. Wei, C. T. Chou, and W. Hu, "Learn to recognise: Exploring priors of sparse face recognition on smartphones," *IEEE Trans. Mobile Comput.*, vol. 16, no. 6, pp. 1705–1717, Jun. 2017.

[8] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *ACM Comput. Surv.*, vol. 35, no. 4, pp. 399–458, 2003.

[9] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," *Int. J. Comput. Vis.*, vol. 104, no. 2, pp. 154–171, 2013.

[10] S. Lazebnik, C. Schmid, and J. Ponce, "A sparse texture representation using affine-invariant regions," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2, Jun. 2003, p. II.

[11] S. Murala, R. P. Maheshwari, and R. Balasubramanian, "Local tetra patterns: A new feature descriptor for content-based image retrieval," *IEEE Trans. Image Process.*, vol. 21, no. 5, pp. 2874–2886, May 2012.

[12] R. Chellappa, C. L. Wilson, and S. Sirohey, "Human and machine recognition of faces: A survey," *Proc. IEEE*, vol. 83, no. 5, pp. 705–741, May 1995.

[13] A. Bansal, A. Nanduri, C. D. Castillo, R. Ranjan, and R. Chellappa, "UMDFaces: An annotated face dataset for training deep networks," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, Oct. 2017, pp. 464–473.

[14] P. N. Belhumeur, J. P. Hespanha, and D. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 711–720, Jul. 1997.

[15] Y. Sun, Y. Chen, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 1988–1996.

[16] S. Chakraborty, S. K. Singh, and P. Chakraborty, "Local gradient hexa pattern: A descriptor for face recognition and retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 1, pp. 171–180, Jan. 2018.

[17] J.-C. Chen, V. M. Patel, and R. Chellappa, "Unconstrained face verification using deep CNN features," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2016, pp. 1–9.

[18] X. Tan, S. Chen, Z.-H. Zhou, and F. Zhang, "Face recognition from a single image per person: A survey," *Pattern Recognit.*, vol. 39, no. 9, pp. 1725–1745, 2006.

[19] R. Jafri and H. R. Arabnia, "A survey of face recognition techniques," *J. Inf. Process. Syst.*, vol. 5, no. 2, pp. 41–68, 2009.

[20] A. Bansal, C. Castillo, R. Ranjan, and R. Chellappa, "The do's and don'ts for CNN-based face verification," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, Oct. 2017, pp. 2545–2554.

[21] L. Nanni, S. Ghidoni, and S. Brahnam, "Handcrafted vs. non-handcrafted features for computer vision classification," *Pattern Recognit.*, vol. 71, pp. 158–172, Nov. 2017.

[22] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 12, pp. 2037–2041, Dec. 2006.

[23] M. Pietikäinen, A. Hadid, G. Zhao, and T. Ahonen, "Computer vision using local binary patterns," in *Computer Vision Using Local Binary Patterns*. London, U.K.: Springer, 2011, pp. E1–E2.

[24] W. Zhang, S. Shan, W. Gao, X. Chen, and H. Zhang, "Local Gabor binary pattern histogram sequence (LGBPHS): A novel non-statistical model for face representation and recognition," in *Proc. 10th IEEE Int. Conf. Comput. Vis. (ICCV)*, vol. 1, Oct. 2005, pp. 786–791.

[25] B. Zhang, S. Shan, X. Chen, and W. Gao, "Histogram of Gabor phase patterns (HGPP): A novel object representation approach for face recognition," *IEEE Trans. Image Process.*, vol. 16, no. 1, pp. 57–68, Jan. 2007.

[26] D. Chen, X. Cao, F. Wen, and J. Sun, "Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3025–3032.

[27] B. Zhang, Y. Gao, S. Zhao, and J. Liu, "Local derivative pattern versus local binary pattern: Face recognition with high-order local pattern descriptor," *IEEE Trans. Image Process.*, vol. 19, no. 2, pp. 533–544, Feb. 2010.

[28] J.-C. Chen, S. Sankaranarayanan, V. M. Patel, and R. Chellappa, "Unconstrained face verification using Fisher vectors computed from frontalized faces," in *Proc. IEEE 7th Int. Conf. Biometrics Theory, Appl. Syst. (BTAS)*, Sep. 2015, pp. 1–8.

[29] J. Lu, V. E. Liong, G. Wang, and P. Moulin, "Joint feature learning for face recognition," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 7, pp. 1371–1383, Jun. 2015.

[30] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1891–1898.

[31] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "DeepFace: Closing the gap to human-level performance in face verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1701–1708.

[32] X. Wu, R. He, Z. Sun, and T. Tan, "A light CNN for deep face representation with noisy labels," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 11, pp. 2884–2896, Nov. 2018.

[33] Y. Sun, X. Wang, and X. Tang, "Deeply learned face representations are sparse, selective, and robust," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 2892–2900.

[34] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 815–823.

[35] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, X. Xie, M. W. Jones, and G. K. L. Tam, Eds. BMVA Press, Sep. 2015, pp. 41.1–41.12.

[36] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[37] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," 2014, *arXiv:1411.7923*. [Online]. Available: http://arxiv.org/abs/1411.7923

[38] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2005, pp. 539–546.

[39] M. Guillaumin, J. Verbeek, and C. Schmid, "Is that you? Metric learning approaches for face identification," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep. 2009, pp. 498–505.

[40] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, "Information-theoretic metric learning," in *Proc. 24th Int. Conf. Mach. Learn.*, 2007, pp. 209–216.

[41] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *J. Mach. Learn. Res.*, vol. 10, no. 2, pp. 207–244, 2009.

[42] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," in *Proc. ICML Deep Learn. Workshop*, vol. 2, Lille, France, 2015, pp. 1–8.

[43] J. L. Long, N. Zhang, and T. Darrell, "Do convnets learn correspondence?" in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 1601–1609.

[44] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[45] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, "Deep learning for computer vision: A brief review," *Comput. Intell. Neurosci.*, vol. 2018, pp. 1–13, Feb. 2018.

[46] A. Brunetti, D. Buongiorno, G. F. Trotta, and V. Bevilacqua, "Computer vision and deep learning techniques for pedestrian detection and tracking: A survey," *Neurocomputing*, vol. 300, pp. 17–33, Jul. 2018.

[47] S. Han, J. Pool, S. Narang, H. Mao, E. Gong, S. Tang, E. Elsen, P. Vajda, M. Paluri, J. Tran, B. Catanzaro, and W. J. Dally, "DSD: Dense-sparse-dense training for deep neural networks," 2016, *arXiv:1607.04381*. [Online]. Available: http://arxiv.org/abs/1607.04381

[48] Z. Wei, Y. Sun, J. Wang, H. Lai, and S. Liu, "Learning adaptive receptive fields for deep image parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 2434–2442.

[49] Y. Chen, H. Fan, B. Xu, Z. Yan, Y. Kalantidis, M. Rohrbach, Y. Shuicheng, and J. Feng, "Drop an octave: Reducing spatial redundancy in convolutional neural networks with octave convolution," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2019, pp. 3435–3444.

[50] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, Sep. 2014, pp. 818–833.

[51] D. E. Worrall, S. J. Garbin, D. Turmukhambetov, and G. J. Brostow, "Harmonic networks: Deep translation and rotation equivariance," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 5028–5037.

[52] M. Weiler, F. A. Hamprecht, and M. Storath, "Learning steerable filters for rotation equivariant CNNs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 849–858.

[53] B. Wu, A. Wan, X. Yue, P. Jin, S. Zhao, N. Golmant, A. Gholaminejad, J. Gonzalez, and K. Keutzer, "Shift: A zero FLOP, zero parameter alternative to spatial convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9127–9135.

[54] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, and C. Xu, "GhostNet: More features from cheap operations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 1580–1589.

[55] C. Ye, M. Evanusa, H. He, A. Mitrokhin, T. Goldstein, J. A. Yorke, C. Fermuller, and Y. Aloimonos, "Network deconvolution," in *Proc. Int. Conf. Learn. Represent.*, 2020, pp. 1–20.

[56] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1724–1734.

[57] S. Dey, A. Dutta, J. I. Toledo, S. K. Ghosh, J. Lladós, and U. Pal, "SigNet: Convolutional Siamese network for writer independent offline signature verification," 2017, *arXiv:1707.02131*. [Online]. Available: http://arxiv.org/abs/1707.02131

[58] A. Sikdar and A. S. Chowdhury, "Scale-invariant batch-adaptive residual learning for person re-identification," *Pattern Recognit. Lett.*, vol. 129, pp. 279–286, Jan. 2020.

[59] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015, *arXiv:1502.03167*. [Online]. Available: http://arxiv.org/abs/1502.03167

[60] C. Lomont, "Fast inverse square root," Purdue Univ., Lafayette, IN, USA, Tech. Rep., 2003. Accessed: Aug. 25, 2021. [Online]. Available: http://www.lomont.org/Math/Papers/2003/InvSqrt.pdf

[61] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. 27th Int. Conf. Mach. Learn. (ICML)*, 2010, pp. 807–814.

[62] T.-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear CNN models for fine-grained visual recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1449–1457.

[63] A. R. Chowdhury, T.-Y. Lin, S. Maji, and E. Learned-Miller, "One-to-many face recognition with bilinear CNNs," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2016, pp. 1–9.

[64] M. M. Ghazi and H. K. Ekenel, "A comprehensive analysis of deep learning based representation for face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2016, pp. 34–41.

[65] E. Y. Liu, Z. Guo, X. Zhang, V. Jojic, and W. Wang, "Metric learning from relative comparisons by minimizing squared residual," in *Proc. IEEE 12th Int. Conf. Data Mining*, Dec. 2012, pp. 978–983.

[66] E. P. Xing, M. I. Jordan, S. J. Russell, and A. Y. Ng, "Distance metric learning with application to clustering with side-information," in *Proc. Adv. Neural Inf. Process. Syst.*, 2003, pp. 521–528.

[67] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "VGGFace2: A dataset for recognising faces across pose and age," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2018, pp. 67–74.

[68] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980

[69] F. S. Samaria and A. C. Harter, "Parameterisation of a stochastic model for human face identification," in *Proc. IEEE Workshop Appl. Comput. Vis.*, Dec. 1994, pp. 138–142.

[70] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman, "From few to many: Generative models for recognition under variable pose and illumination," in *Proc. 4th IEEE Int. Conf. Autom. Face Gesture Recognit.*, Mar. 2000, pp. 277–284.

[71] K.-C. Lee, J. Ho, and D. J. Kriegman, "Acquiring linear subspaces for face recognition under variable lighting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 5, pp. 684–698, May 2005.

[72] L. Lenc and P. Král, "Unconstrained facial images: Database for face recognition under real-world conditions," in *Proc. Mex. Int. Conf. Artif. Intell.* Cham, Switzerland: Springer, Oct. 2015, pp. 349–361.

[73] 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 *Proc. Workshop Faces Real-Life Images, Detection, Alignment, Recognit.*, 2008, pp. 1–14.

[74] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1026–1034.

[75] D. Misra, "Mish: A self regularized non-monotonic activation function," 2019, *arXiv:1908.08681*. [Online]. Available: http://arxiv.org/abs/1908.08681

[76] S. K. Roy, S. Manna, S. R. Dubey, and B. B. Chaudhuri, "LiSHT: Non-parametric linearly scaled hyperbolic tangent activation function for neural networks," 2019, *arXiv:1901.05894*. [Online]. Available: http://arxiv.org/abs/1901.05894

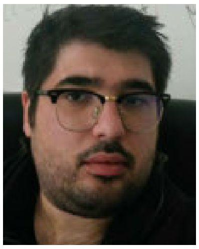
[77] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 618–626.

[78] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.

SWALPA KUMAR ROY (Student Member, IEEE) received the bachelor's degree in computer science and engineering from West Bengal University of Technology, Kolkata, India, in 2012, and the master's degree in computer science and engineering from the Indian Institute of Engineering Science and Technology, Shibpur, Howrah, India, (IIEST Shibpur), in 2015. He is currently pursuing the Ph.D. degree jointly with the Computer Vision and Pattern Recognition Unit, Indian Statistical Institute, Kolkata, and the Department of Computer Science and Engineering, University of Calcutta, Kolkata. He was a Project Linked Person with the Optical Character Recognition (OCR) Laboratory, Computer Vision and Pattern Recognition Unit, Indian Statistical Institute, from July 2015 to March 2016. He is an Assistant Professor with the Department of Computer Science and Engineering, Jalpaiguri Government Engineering College, Jalpaiguri, West Bengal, India. His research interests include computer vision, deep learning, and remote sensing. He has served as a Reviewer for the IEEE Transactions on Geoscience and Remote Sensing and IEEE Geoscience and Remote Sensing Letters.

PURBAYAN KAR is currently pursuing the B.Tech. degree with the Department of Computer Science and Engineering, Jalpaiguri Government Engineering College. His research interests include deep learning and remote sensing.

MERCEDES E. PAOLETTI (Senior Member, IEEE) received the B.Sc. and M.Sc. degrees in computer engineering from the University of Extremadura, Cáceres, Spain, in 2014 and 2016, respectively, and the Ph.D. degree from the Hyperspectral Computing Laboratory (HyperComp), Department of Technology of Computers and Communications, University of Extremadura, in 2020. Her Ph.D. degree was supported by the University Teacher Training Programme from the Spanish Ministry of Education, as a member of HyperComp, Department of Technology of Computers and Communications, University of Extremadura. She is currently a Researcher at the Department of Computer Architecture and Automation, Complutense University, Madrid. Her research interests include remote sensing and analysis of very high spectral resolution with the current focus on DL and high performance computing. She was a recipient of the 2019 Outstanding Paper Award recognition in the IEEE WHISPERS 2019 Conference and the Outstanding Ph.D. Award at the University of Extremadura, in 2020. She has served as a Reviewer for the IEEE Transactions on Geoscience and Remote Sensing and IEEE Geoscience and Remote Sensing Letters, in which she was recognized as the Best Reviewer, in 2019 and 2020.



JUAN M. HAUT (Senior Member, IEEE) received the B.Sc. and M.Sc. degrees in computer engineering and the Ph.D. degree in information technology from the University of Extremadura, Caceres, Spain, in 2011, 2014, and 2019, respectively. He is currently a Professor with the Department of Computers and Communications, University of Extremadura, where he is also a member of the Hyperspectral Computing Laboratory (HyperComp), Department of Technology of Computers and Communications. His Ph.D. degree was supported by the University Teacher Training Programme from the Spanish Ministry of Education. His research interests include remote sensing data processing and high-dimensional data analysis, applying machine (deep) learning, and cloud computing approaches. In this sense, he has authored/coauthored more than 40 JCR journal articles (more than 30 in IEEE journals) and more than 30 peer-reviewed conference proceeding papers. Some of his contributions have been recognized as hot-topic publications for their impact on the scientific community. He was a recipient of the Outstanding Ph.D. Award at the University of Extremadura, in 2019, and the Outstanding Paper Award in the 2019 and 2021 IEEE WHISPERS Conferences. From his experience as a Reviewer, it is worth mentioning his active collaboration in more than ten scientific journals, such as the IEEE Transactions on Geoscience and Remote Sensing, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, and IEEE Geoscience and Remote Sensing Letters. He has been awarded with the Best Reviewer recognition of the IEEE Geoscience and Remote Sensing Letters and IEEE Transactions on Geoscience and Remote Sensing, in 2018 and 2020, respectively. Furthermore, he has guest-edited three special issues on hyperspectral remote sensing for different journals. He is also an Associate Editor of the IEEE Transactions on Geoscience and Remote Sensing, IEEE Gesocience and Remote Sensing Letters, and IEEE Journal on Miniaturization for Air and Space Systems.

RAFAEL PASTOR-VARGAS (Senior Member, IEEE) received the M.Sc. degree in physics from Complutense University, Madrid, Spain, in 1994, and the Ph.D. degree in computer science from UNED, Madrid, in 2006. He was an Innovation Manager with the Innovation and Development Centre, UNED, from 2004 to 2009, and the General Manager for incorporating innovative services to UNED's learning model, from 2009 to 2011. He is currently an Associate Professor at the Control and Communication Systems Department, UNED. He is the Vice Dean of technology at the Computer Science Engineering Faculty and the Coordinator of the official master in engineering and data science. He has participated in a big number of projects financed in public calls, some of them with special relevance for companies and administrations with an international scope, international journals, and international conferences. He teaches graduate and post-graduate courses related to the network interconnection and security domains, among others. His research interests include quality of service support in distributed systems, the development of infrastructure and algorithms for e-learning, and cybersecurity.

ANTONIO ROBLES-GÓMEZ (Senior Member, IEEE) was born in Albacete, Spain, in 1980. He received the M.Sc. and Ph.D. degrees in computer science engineering from the University of Castilla-La Mancha, in 2004 and 2008, respectively. He is currently an Associate Professor at the Control and Communication Systems Department, UNED, and the Coordinator of the official master in computer science engineering. He has participated in a big number of projects financed in public calls, some of them with special relevance for companies and administrations with an international scope, international journals, and international conferences. He teaches graduate and post-graduate courses related to the network interconnection and security domains, among others. His research interests include quality of service support in distributed systems, the development of infrastructure and algorithms for e-learning, and cybersecurity. He is the Secretary of the IEEE Spain Section Board.

• • •