IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 17, 2021, accepted May 9, 2021, date of publication May 17, 2021, date of current version May 27, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3080712

A Study on the Performance of Unconstrained Very Low Resolution Face Recognition: Analyzing Current Trends and New Research Directions

LUIS S. LUEVANO[1], (Member, IEEE), LEONARDO CHANG[1], HEYDI MÉNDEZ-VÁZQUEZ[2], YOANNA MARTÍNEZ-DÍAZ[2], AND MIGUEL GONZÁLEZ-MENDOZA[1]

[1]Tecnologico de Monterrey, School of Engineering and Sciences, Monterrey 64849, Mexico
[2]Advanced Technologies Application Center (CENATAV), Havana 12200, Cuba

Corresponding author: Luis S. Luevano (luis.s.luevano@tec.mx)

The work of Luis S. Luevano was supported in part by the National Council of Science and Technology of Mexico under the CONACYT Scholarship Grant 768608.

ABSTRACT In the past decade, research in the face recognition area has advanced tremendously, particularly in uncontrolled scenarios (face recognition in the wild). This advancement has been achieved partly due to the massive popularity and effectiveness of deep convolutional neural networks and the availability of larger unconstrained datasets. However, several face recognition challenges remain in the context of very low resolution homogeneous (same domain) and heterogeneous (different domain) face recognition. In this survey, we study the seminal and novel methods to tackle the very low resolution face recognition problem and provide an in-depth analysis of their design, effectiveness, and efficiency for a real-time surveillance application. Furthermore, we analyze the advantage of employing deep learning convolutional neural networks, while presenting future research directions for effective deep learning network design in this context.

INDEX TERMS Low resolution face recognition, unconstrained face recognition, coupled mappings, super resolution, efficient face recognition models, lightweight convolutional neural networks.

I. INTRODUCTION

Pattern recognition algorithms have evolved very swiftly in the past decades. Computing power and storage nowadays allow us to process large datasets even in a single computer, allowing us to propose and implement a more robust and accurate pattern recognition models. These models have applications in diverse areas such as cybersecurity [9], critical industrial systems [60], social networking [59], [61], among others. In pattern recognition for computer vision, automated face recognition is a very relevant area of research. Applications for automated face recognition systems include aiding in law enforcement, forensics, surveillance tracking, and biometrics authentication [55], among others. Automated face recognition is the task of identifying or verifying a person's identity using a computer system, where this person's face image serves as a reference for posterior recognition. Today, face recognition algorithms can run in real-time on a smartphone. However, challenges still exist in this area, including face recognition in uncontrolled environments with low image resolution and image artifacts. Face recognition algorithms still need to improve in their robustness, reliability, and inference time performance.

The face recognition task is comprised of four steps: face detection, alignment, feature extraction, and identity matching. A classic algorithm for the face detection step is the Boosting approach proposed by Viola and Jones [97]. The viola-Jones face detector uses classifiers in a cascade fashion, making it robust and efficient at inference run-time. Face alignment consists of using the detected face landmarks to estimate a frontal position for that face image. This process is also known as face frontalization. Some face alignment methods map the new face position in a 2D space or a 3D space [8]. More recent approaches based on convolutional neural networks include Multi-task Cascaded Convolutional Networks [119] and the Retinaface detector [20]. Both approaches couple the face detection and alignment steps and show an impressive benefit of training them together. For the feature extraction step, Traditional approaches include Eigenfaces [95], Fisherfaces [3], the extraction of Local Binary Pattern Histograms [24], Gabor [17], SIFT [57], and SURF [2] features; and at the start of this decade, works with learned

The associate editor coordinating the review of this manuscript and approving it for publication was Ali Shariq Imran.

descriptors such as [11] started to emerge. However, these methods struggle with capturing the non-linearity (deformations, pose, and lighting conditions) of face appearances in unconstrained scenarios. The face recognition evaluation task has two variants: identification and verification. Identification refers to a one-to-many probe and gallery matching, while verification refers to a one-to-one probe and gallery verification. Simultaneously, the evaluation can be open-set, where probes can either appear or not in the gallery, or closed-set, where the probes always appear in the gallery.

Face recognition in uncontrolled scenarios, also called face recognition in the wild, refers to identifying or verifying a person from face images taken from scenes with variations in illumination, scale, viewpoint, aging, partial occlusion, and image quality. Automated face recognition in uncontrolled scenarios has made remarkable progress in the last decade, mainly due to the popularization of methods based on neural networks. Popular benchmark databases such as Labeled Faces in the Wild (LFW) [41], YouTube Faces (YTF) [105], and CelebA [56] have gathered most of the attention from researchers in the face recognition field. Today it is possible to obtain a recognition accuracy performance of more than 96%, even in real-time constraints (more than 30 frames per second) [66] in these datasets. In turn, these models sacrifice model explainability [58] in favor of achieving better recognition performance.

Additional approaches for bringing face recognition and biometrics authentication methods to a real-world application include protection against computer vision-based authentication system adversarial attacks such as CNN-based Anti-spoofing [71] and multi-factor authentication systems [81]. Surveillance systems can also benefit from matching Near Infrared and Visible Spectrum images (NIR-VIS) as infrared sensors become more widely available. As noted by [73], these approaches focus on modeling the change of illumination from the visible spectrum domain to the infrared one. Modern approaches in this area include [35], [124], and [75].

Due to the recent increased focus on real-world video surveillance applications, there is also an increasing need for face recognition algorithms robust to very low resolution scenarios. This newer very low resolution face recognition challenge translates to poor face recognition accuracy performance for traditional face recognition algorithms. Recognition accuracy performance rapidly declines once face regions with an area of 32×32 pixels or below are present [7]. We refer to these pixel area sizes to *very low resolution (VLR)* in this paper. Furthermore, at the start of this decade, datasets accurately representing this VLR video surveillance scenario were unavailable. Recent efforts for more accurate real-world surveillance application datasets for very low resolution face recognition have emerged in the past few years. These databases include the SCface dataset [29], UCCS [84], IJB-S [48], TinyFace [14], and Survface [16].

In contrast to previous studies for very low resolution face recognition [53], [54], and [100], we explore and analyze the efficacy and efficiency of the state-of-the-art methods on challenging datasets for this problem. We focus on the aspects each method uses to solve the very low resolution problem and the complexity elements impacting efficiency performance, while also pointing out its limitations. Additionally, in comparison with [72], this survey also covers more recent methods from the state of the art relevant to specific native VLR face recognition datasets, described in Subsection I-A. Furthermore, we discuss other face recognition approaches focused primarily on efficient run-time computation. Lastly, we provide new research directions stemming from advancements on lightweight convolutional neural networks, capsule networks, and knowledge distillation.

In this study, we make the following contributions:

- Provide an efficiency and efficacy analysis of current state-of-the-art methods for very low resolution face recognition.
- Provide an in-depth study of the advantages and limitations of current approaches for the very low resolution face recognition problems from an application perspective.
- Provide an overview of alternative modern lightweight architectures and their performance on very low resolution face datasets.
- Provide insights for future research directions.

This paper is organized as follows. Section I presents the unconstrained very low resolution face recognition problem. Section II analyzes the state-of-the-art methods for solving the VLR FR Problem, divided into Subsection II-A for heterogeneous approaches and Subsection II-B for efficient homogeneous approaches. Section III presents and discusses the performance results of state-of-the-art approaches in relevant unconstrained VLR FR datasets (Subsections III-A and III-B), discusses the research challenges affecting unconstrained VLR FR performance (Subsection III-C), and discusses future research directions for the VLR FR area (Subsection III-D. Finally, Section IV gives the final remarks of our study.

A. DATASETS FOR VERY LOW RESOLUTION FACE RECOGNITION

Currently, very low resolution face recognition under surveillance scenarios is a very niche research area with limited datasets available. Recent efforts for expanding studies in this area include the datasets described in Table 1. The SCface [29], Point and Shoot [4], IJB-S [48], UCCSface [84], QMUL-Survface [16], and QMUL-Tinyface [14] are the available benchmark datasets for unconstrained very low resolution face recognition. Extensive studies across all the previously mentioned benchmark datasets are not present in the current state of the art.

Fig. 1 shows samples of all the aforementioned datasets. We can appreciate that the datasets suited for heterogeneous face recognition benchmarking are the SCface, Point and Shoot, UCCSface, and IJB-S, because they contain both high

TABLE 1. Summary of unconstrained datasets for very low resolution face recognition. Taken and complemented from [53]. Most of the available datasets come from real-world surveillance imagery and contain high resolution and low resolution pairs, except QMUL-surface and QMUL-Tinyface. However, these two datasets contain the largest number of images and identities, making them suitable training deep learning methods.

Database name	Source	Quality	Static image/video	# subjects	# images
Point and Shoot [4]	Manually Collected	HR + blur	static + video	558	12,178
SCface [29]	Surveillance	HR + LR	static	130	4,160
QMUL-Survface [16]	Surveillance	LR	static + video	15,573	463,507
QMUL-TinyFace [14]	Web	LR	static	5,139	169,403
UCCSface [84]	Surveillance	HR + blur	static	308	6,337
IJB-S [48]	Surveillance	HR + LR	static + video	202	3 million+

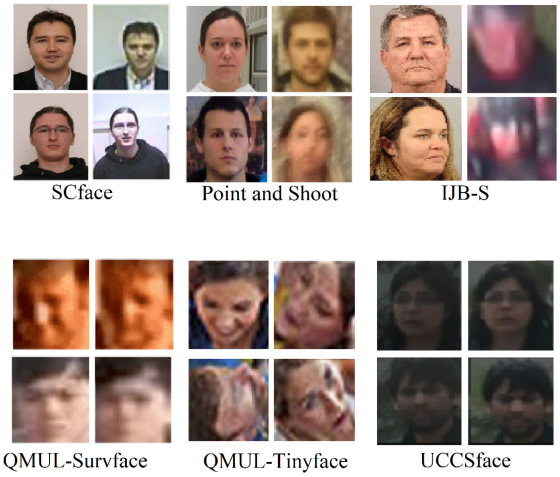

FIGURE 1. Example of subjects in the different datasets available for very low resolution face recognition, taken from their respective dataset papers [4], [16], [29], [48], [14], and [84]. The SCface [29], Point and Shoot [4], IJB-S [48], and UCCSface [84] are the most suitable for heterogeneous face recognition, because they supply the HR and VLR image pairs for each identity. The QMUL-Survface [16] and QMUL-Tinyface [14] are suitable for the homogenous face recognition problem, where only VLR images are available.

resolution and native low resolution imagery. On the other hand, the QMUL-Survface and QMUL-TinyFace datasets are suitable for deep learning homogeneous face recognition due to their large number of images and identities. These are very different problems on their own. However, in a multi-network architecture solution, subnetworks could be pre-trained using the VLR homogeneous datasets only, for instance.

Table 1 summarizes the datasets available for the very low resolution face recognition specific task. The source column indicates the scenario where the data was obtained, the quality column indicates the type of the available images where "HR" corresponds to High Resolution imagery, "LR" to Very Low Resolution imagery and "blur" to blurred low resolution images. The static image/video column indicates whether the dataset has static images and also contains video data or not. The next columns indicate the total number of identities and number of images.

II. STATE-OF-THE-ART METHODS FOR SOLVING THE VERY LOW RESOLUTION FACE RECOGNITION PROBLEM

The face recognition problem at very low resolution has two variants: homogeneous and heterogeneous. In homogeneous face recognition, we match images that come from the same source domain. In this case, both the probe images and the rest of the reference images come from the unconstrained VLR domain. In heterogeneous face recognition, we match images from different domains: the probe VLR images with the high resolution gallery images. As such, a domain gap exists between the VLR probe image taken from the surveillance camera and the high resolution reference gallery image taken in a controlled environment. The VLR probe images have a 32×32 resolution or less, and the reference gallery HR images have a 100×100 resolution or more. The heterogeneous variant of the problem is the hardest one due to the domain disparity between the probe and camera resolutions in a varying range of conditions. Fig. 2 summarizes the taxonomy for the state-of-the-art solutions for the variants of the very low resolution face recognition problem. This taxonomy is divided into methods for the Heterogeneous face recognition problem, which are: Projection methods (Coupled Mappings) and Synthesis methods (Super Resolution), and the Homogeneous feature extraction and matching for the homogeneous face recognition problem. This distinction is common in the heterogeneous face recognition literature [73]. Projection methods, called Coupled Mapping methods in LR face recognition literature, aim to project both domains into the same unified space. Synthesis methods, commonly referred to as Super Resolution methods, aim to project domain into the other. In this case, the VLR images get projected into the HR domain to perform face recognition. The Homogeneous feature extraction and matching methods are the traditional face recognition methods that perform a direct one-to-one comparison with all the face images. We propose to further divide each approach into traditional and deep learning methods, since they have an important performance gap in terms of accuracy and efficiency.

A. HETEROGENEOUS APPROACHES FOR VERY LOW RESOLUTION FACE RECOGNITION

This subsection describes the state-of-the-art methods for solving the heterogeneous very low resolution face recognition problem. These approaches are divided into Projection and Synthesis methods, also commonly known as Coupled Mappings and Super Resolution methods respectively.

1) PROJECTION METHODS: COUPLED MAPPINGS

Coupled mapping methods fall into the category of projection methods. This type of methods aim to find an adequate

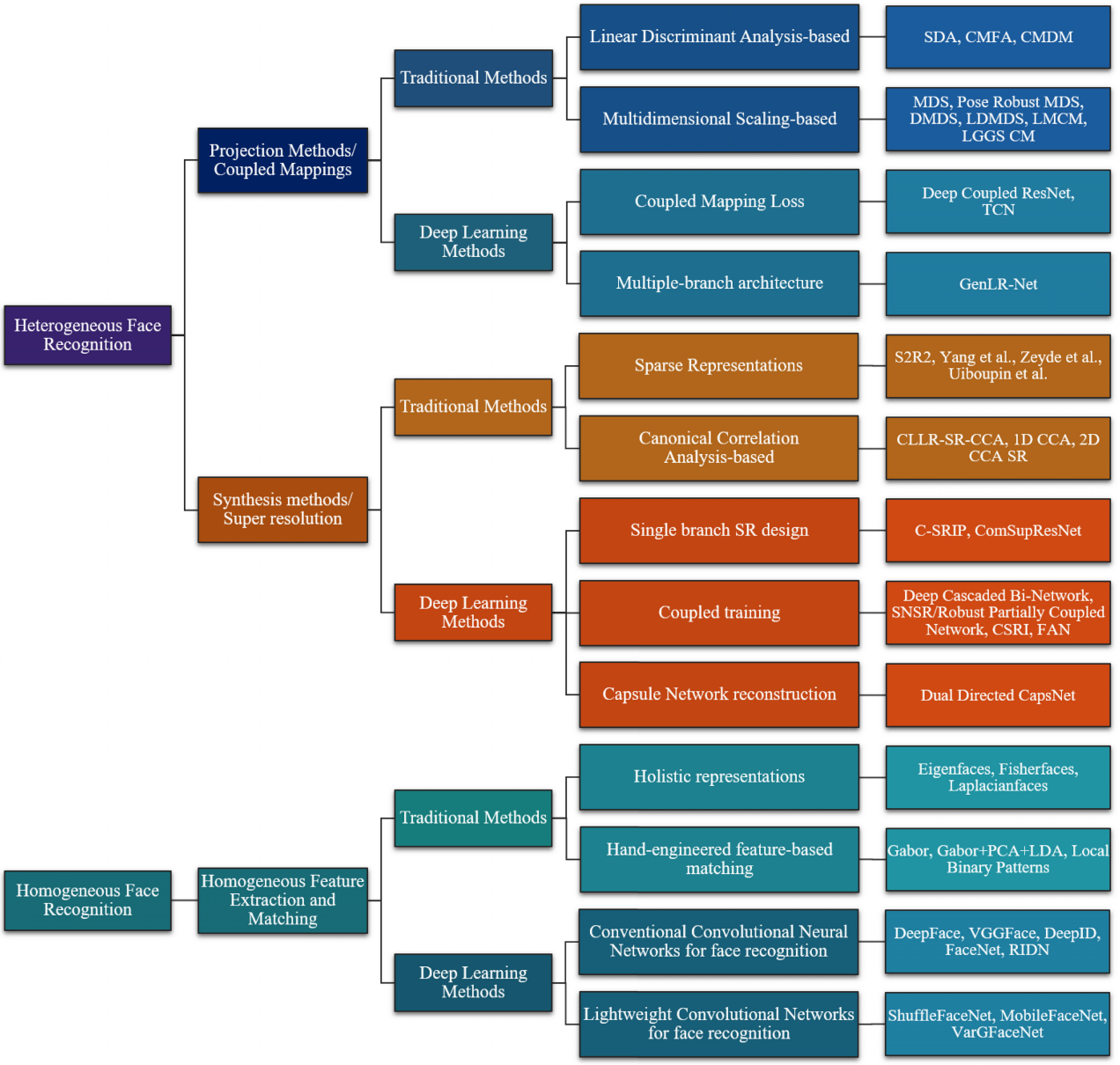

FIGURE 2. **Taxonomy of the very low resolution face recognition state of the art approaches. We propose to divide every approach into both traditional and deep learning methods due to the increasing effectiveness and popularity of deep learning methods in the past decade. Deep learning methods tend to have a better generalization and take advantage of graphical acceleration (GPU) hardware to close the gap for a real-time application. Approaches for bringing the real-world application to reasonable inference times on CPU include the Traditional Multidimensional Scaling Coupled Mappings and the Lightweight Convolutional Neural Networks for homogeneous feature extraction and matching methods.**

representation of data from different domains by projecting the HR and VLR images to a single unified space. Afterwards, similarity metrics are be computed for classification and posterior optimization operations, as illustrated in Fig. 3. Table 2 shows a summary of all the projection methods discussed in this subsection.

a: CLASSICAL APPROACHES FOR COUPLED MAPPINGS

Classical methods for Coupled Mappings include **Coupled Marginal Discriminant Mappings** [120], **Multidimensional Scaling (MDS)** [6] and **Pose-Robust MDS** [5] by Biswas *et al.*, **Simultaneous Discriminant Analysis** [12], and **Coupled Marginal Fisher Analysis** [88]. These methods are based on feature extraction, projection, and using a variant of Linear Discriminant Analysis for classification and matching. The optimization method of choice for MDS problems is the iterative majorization algorithm [103].

The **Multidimensional Scaling (MDS)** approach proposed by Biswas *et al.* [6] proposed to optimize the distance between the transformed feature vectors using a combined transformation matrix with three different regularizing terms. The goal of the optimization is to approximate the distance

TABLE 2. Summary of coupled mapping techniques. Most of the traditional approaches in the last years are based on Multidimensional Scaling [6] and improved upon using locality constraints and enforcing inter-class margins. The rest of the methods are Deep Learning-based methods based on ResNet/VGG architectures enforcing cross-resolution learning at an architecture branch level or the loss function level.

Method	Approach	Reported metrics
Simultaneous Discriminant Analysis (SDA) (2011) [12]	Learn projection matrices using LDA-based scatter matrices for images from LR and HR domains and their matching combinations.	Multi-PIE [36] mean accuracy: 96.46% on LR probe. No efficiency metrics reported.
Coupled Marginal Fisher Analysis (2012) [88]	Marginal Fisher Discriminant Analysis-based optimization. Objective function is to minimize the inter-class and intra-class ratio of the sum of the distances between the projected features to the unified space.	Multi-PIE [36] mean accuracy: 96.80% on LR probe. No efficiency metrics reported.
Coupled Marginal Discriminant Mappings (CMDM) (2015) [120]	Model and optimize the ratio of similarity (scatter) matrices between and inter-class for solving as an eigen-decomposition problem. Similar to CMFA.	FERET mean accuracy: 88.5%. No efficiency metrics reported.
Multidimensional Scaling (MDS) (2012) [6]	Optimize projected LR and HR feature distance and approximate them to HR features from the source domain.	SCface mean accuracy: 60%. No efficiency metrics reported.
Pose-Robust MDS (2013) [5]	Based on MDS [6]. Model and estimate mode and median matrices from viewpoint, illumination and eigenimage info from training set.	Multi-PIE: over 80% recognition rate. SCface: outperforms SIFT+PCA, SIFT+LDA, SURF+PCA, LBP on rank-1 accuracy CMC curve by more than 10% margin. No efficiency metrics reported.
Discriminative Multidimensional Scaling (DMDS) (2018) [108]	Inspired by MDS [6]. Adds inter-class and intra-class constraints to better project the features pertaining to each class in the latent subspace.	SCface mean accuracy: 79.92%. No efficiency metrics reported.
Local-Consistency Preserved DMDS (LDMDS) (2018) [108]	Complementary to DMDS. Only optimizes the sample distance from the same domains, not accross.	SCface mean accuracy: 81.54%. No efficiency metrics reported.
Large Margin Copuled Mappings (LMCM) (2016) [116]	Based on LDA. Maximizes class margins using weights from constructed class graphs. Class is centroid-based.	SCface mean accuracy: 60.4%. Inference: takes 8.5 microseconds per image (117.65 face images per second) on i5-4200U CPU.
Local Geometry to Global Structure CM [86] (2015)	Minimizes the mappings that minimize the distance of LR and HR neighbors from the same class. Subsequently, combine these mappings to generate the global projection matrix.	SCface: 43.2% mean accuracy. No efficiency metrics reported.
Deep Coupled ResNet (2018) [62]	Deep learning-based method. Uses trunk-branch structure. Feature extraction using a ResNet style subnetwork and branch FC subnetworks for LR and HR images. Rescales images for data augmentation strategy. Optimization using softmax and centerloss functions.	SCface: 88.2% mean accuracy. No efficiency metrics reported.
GenLR-Net (2018) [69]	Deep learning-based method. Uses VGG face architecture as base. Uses multiple classification losses at different points in the network to model the LR-HR relationship and a final contrastive loss function to learn the LR projection closer to the HR projected features.	LFW: 90.00% mean ver. rate. CFP: 77.28% ver. rate. No efficiency metrics reported.
Transferable Coupled Network (TCN) (2019) [115]	Dual branch architecture forHR and VLR input with ResNet/VGGFace backbone. Only the HR subnetwork is pretrained. Optimized using the triplet loss by setting HR and LR anchors with positive and negative reinforcements from the opposite domain.	SCface: 89.37% mean accuracy with ResNet backbone. No efficiency metrics reported.

of the transformed feature vectors in the projected space, to the distance of the samples in the default high resolution space. The regularizing terms use an independent parameter to control the approximation rate to the sample's distance in the high resolution space and the class separability.

Later, Biswas *et al.* proposed **Pose-Robust MDS** [5], which introduced a pose estimation after the MDS computation. The pose estimation process firstly projects the probe image on the tensor basis extracted from the training low resolution set, then estimates pose using the fiducial locations

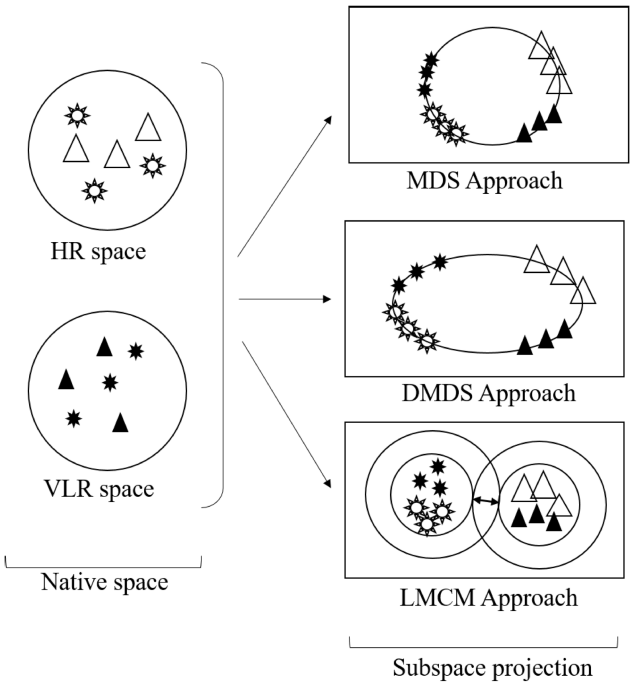

FIGURE 3. Visual comparison of classical Coupled Mappings methods. In MDS [6], both domains get projected into a single common subspace with the HR and the VLR features are closer but separated in the intra-class space. In contrast, DMDS [108] introduces inter-class and intra-class constraints that promote a larger, more discriminative, latent space. In LMCM [116], the improved inter-class and intra-class margins in the common subspace lead to a better recognition performance in unconstrained scenarios, as opposed to prior techniques.

in relation to the median fiducial locations from the training set. It involves modeling mode matrices for the subjects in the training set, spaces of viewpoint, illumination, and eigenimage vectors [95]. Then, after computing a TensorFace component, the coefficient vectors for the face normalization are computed. The authors also introduced robustness by slightly perturbating the fiducial locations. The MDS and pose components are independently trained, which helps at not heavily taxing the total training time by the pose estimation component and its increased number of fiducial locations due to augmentations.

The approach proposed by Zhang *et al.* [116] called **Large Margin Coupled Mappings** constructs inter-class and intra-class graphs, then learns the projection matrices by maximizing the class margins using the weights from the graph distances of the inter-class graph and vice-versa for the weights in the intra-class graph. Then, it introduces the intra-class scatter as a regularizing term. The method uses a scatter matrix approach for computing the inter-class and intra-class distance, where a class centroid is used as a reference, the same as in LDA approaches. Fig. 3 illustrates the margin enforced by the class graphs in the common subspace. The reported inference run-time is of 8.5 microseconds on an Intel Core i5-4200U Laptop CPU, which is equivalent to processing 117.65 face images per second.

The **Discriminative Multidimensional Scaling (DMDS)** approach proposed by Yang *et al.* [108] aims to find the projection matrices to minimize the distance between intra-class samples in the common latent subspace. This work is directly inspired by the previously mentioned MDS approach [6], and it features two additional inter-class and intra-class constraints to add discriminative potential in the latent subspace. These constraints contribute to the projection matrix optimization by using the projected sample distance in the latent subspace according to its class. Fig. 3 graphically illustrates the discriminative space encouraged by the method's constraints. The authors also proposed a variant of the method called **Local-Consistency Preserved DMDS (LDMDS)**, where they changed the optimization of the sample distance to only optimize the distance of the samples from the same domain space additional to the MDS optimization function.

The **Local Geometry to Global Structure CM** approach proposed by Shi and Qi [86] aims to minimize the distance between projected features as well. However, this approach uses a k-neighbor approach to influence the distance optimization in both intra-class and inter-class projected groups. In the inter-class constraint, they included an independent term that heavily penalizes the nearest neighbors and maximizes the margin across classes. After the projection, the global structure concatenating the resulting LR and HR feature vectors is built and utilized to optimize the projection.

b: DEEP LEARNING APPROACHES FOR COUPLED MAPPINGS

Few methods using Deep Learning with a coupled mapping strategy have been proposed for the very low resolution face recognition problem, such as [54] and [37]. However, they do not aim to model the unified space for cross-resolution face recognition, rather to extract robust features from LR and HR faces. In contrast, the **Deep Coupled ResNet** method in [62] consists of one residual trunk network which specializes in feature extraction across different resolutions and two branch networks that minimize the distance between intra-class samples. Both of the branch networks utilize the same loss function for optimization. This architecture is illustrated in Fig. 4.

The strength of this method relies on the robustness of the extracted features using the trunk network. It also employs modern face recognition heuristics such as using a PReLU [94] activation function.

This method achieves the best recognition rate for the SCface dataset. For the camera positioned at 4.2 meters, it yields an accuracy of 73.3%, corresponding to more than a 10% advantage over the previously described methods. Even though this method is more effective than the previous ones, it makes assumptions regarding the resolution factor by using bilinear interpolation in various steps, mirroring a data augmentation strategy where we identify that there is room for improvement. A more intelligent method such as super resolution for face hallucination could be used to improve the image reconstruction process or a different network design for synthesizing images of different resolutions.

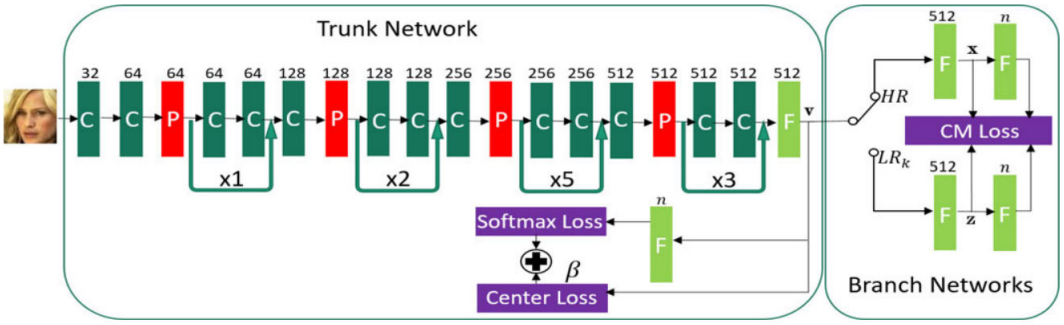

FIGURE 4. Overview of the Deep Coupled Resnet architecture, taken from [62]. This architecture features a single ResNet-style network for feature extraction where the Coupled Mapping loss fits the generated features using the images from both HR and VLR domains. The network weights are updated by the CenterLoss [104] function for more accurate face recognition performance.

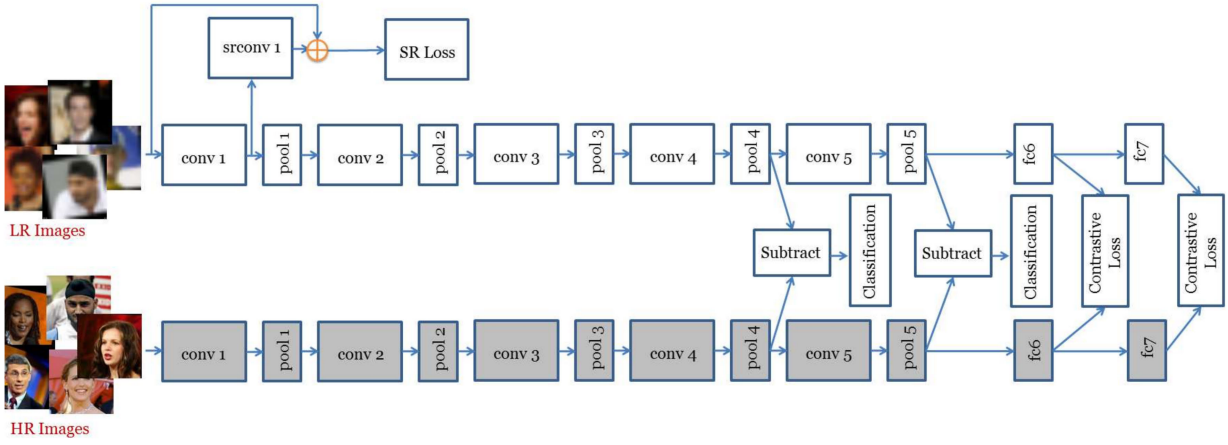

FIGURE 5. GenLR-Net structure, taken from [69]. The weights from the VGG [74] backbone for the high resolution sub-network are pre-trained while the low resolution sub-network is fully trained.

Another method using a projection approach is the **GenLR-Net** [69], detailed in Fig. 5, which uses the VGG face network as a base. This projection method also includes a small Super Resolution component, which marginally improves performance. This network features two subnetworks: one for low resolution images and the other one for high resolution images. This method uses two loss functions: an inter-class and intra-class loss before the final convolution and pooling layers with the contrastive loss for the fully connected layers. The contrastive loss gradient gets propagated only to the low resolution network and not to the high resolution network. The intuition is to project the features closer to their high resolution counterparts, a similar intuition to [90]. This method was not tested on the SCface dataset.

The **Transferable Coupled Network (TCN)** [115] bridges the domain gap by learning the LR subnetwork parameters from the fixed and pre-trained HR subnetwork and its optimization process. The process is shown in Fig. 6. In the same vein as other domain adaptation methods, it uses the triplet loss to learn HR and LR anchors and updating them with cross-resolution examples from the same subject. The transferable triplet loss also enforces inter-class margins for the identity anchors. The authors propose using VGGFace or ResNet architectures for the HR and LR subnetworks. Using ResNet for feature extraction outperforms the VGGFace alternative in both the LFW and SCface datasets. However, using the ResNet backbone is significantly more computationally expensive. The authors reported a mean accuracy performance of 89.37% on the SCface dataset and 95.05% in the VLR probe sizes on the LFW dataset.

c: DISCUSSION ON PROJECTION METHODS

As we have showed, most of the classical methods are based on MDS and Fisher Analysis/LDA. While we can see a clear recognition performance boost on Pose-Robust MDS versus the original MDS by Biswas, other methods have shown that enforcing inter and intra-class margins yield better robustness than estimating frontal poses for the methods. The DMDS and LDMDS methods use this margins to improve performance. However, none of these classical methods project the VLR samples closer to the projected HR samples in the unified space, a notion which has aided in the performance of newer deep learning-based Super Resolution methods. In terms of efficiency, these methods allow servicing many cameras even

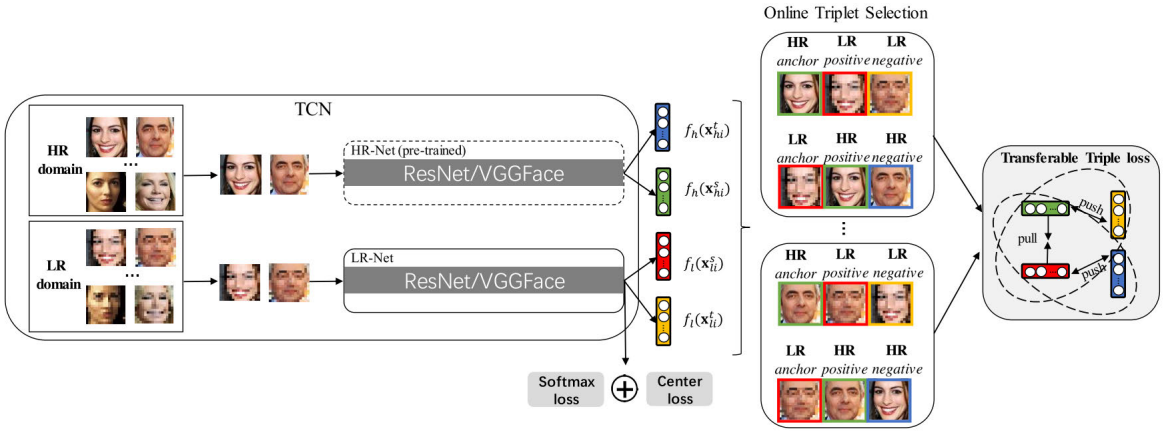

FIGURE 6. Overview of the Transferable Coupled Network, taken from [115]. This method features multiple independent subnetworks for processing HR and VLR imagery, based on seminal ResNet or VGGFace for feature extraction. The triplet loss optimizes the networks with resolution-specific anchors.

on affordable hardware, such as Laptop CPUs as evidenced by LMCM, assuming ten frames per second as real-time performance.

In the case of deep learning-based methods, their most glaring limitation is that they are not optimized for real-time performance on CPU or affordable GPUs. In this area, efficient deep learning approaches have not been proposed. TCN and DCR both feature ResNet backbone architectures while GenLR-Net features a VGGFace backbone. From both of these architecture types, ResNet architectures yield a better accuracy performance. However, both of these backbones have been outperformed by lightweight CNN architectures [67] even when fine-tuning to VLR datasets. As such, ResNet is also a shortcoming of these deep learning approaches.

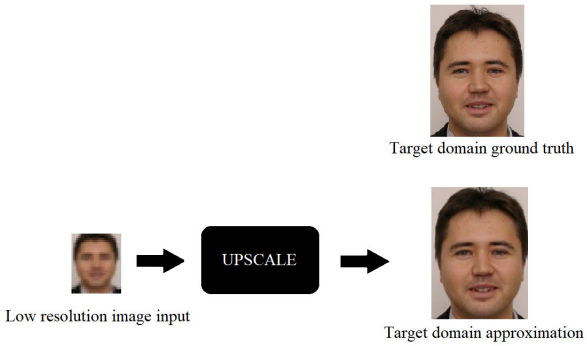

FIGURE 7. Basic idea of super resolution methods for face recognition. In this type of approach, the synthesis is made from the low resolution probe image space to the gallery high resolution space only. The high resolution image remains the same. Then, a similarity measurement is employed to score the upscaled probe against the gallery images.

2) SYNTHESIS METHODS: SUPER RESOLUTION-BASED APPROACHES

Super Resolution methods aim to upscale a low resolution image by a factor, usually of $4\times$. For the context of very low resolution face recognition, they first upscale the very low resolution image to the high resolution space and later perform face recognition. Some of these methods have a face recognition constraint in the upscaling network. Fig. 7 illustrates the basic idea of super resolution methods.

Having a recognition constraint in the same super resolution process yields better accuracy than performance super resolution and recognition separately. Running the two tasks separately can hinder performance in some datasets as stated in [14]. An example of this phenomenon is the approach described on [118], which uses a face recognition loss to train the overall super resolution network as well. This method outperforms state-of-the-art methods not tailored for face recognition purposes, such as the Laplacian Super Resolution Network [49]. Table 3 shows the most successful super resolution methods specifically made for and tested on very low resolution face datasets.

Super Resolution is an active research area, however, most works are focused only on increasing the most popular metrics for this task. This metrics are the Structural Similarity Index (SSIM) and Peak Signal-to-Noise Ratio (PSNR). The value of these metrics is dataset-dependent and they are not a real indicator of their discriminative potential for face recognition purposes. Furthermore, for supervised Super Resolution training, the most common approach is to take a high resolution image dataset and use bilinear interpolation to downscale the images, in order to be able to train the models.

Using artificial downsampling strategies present a domain disparity problem between native very low resolution images from the surveillance feed and the synthetically downscaled images. Moreover, the datasets for this task do not contain only face images, most contain miscellaneous images such as buildings, animals, objects, etc. This hinders performance for face hallucination purposes. Most of the modern methods in this area are neural network-based.

TABLE 3. Summary of super resolution approaches discussed in this section. Traditional approaches have been mostly tested with constrained face recognition datasets, while deep learning approaches have been tested primarily with unconstrained face recognition datasets. Most of these deep learning approaches have been only tested on the UCCS dataset and have made a tremendous leap on recognition performance and different architecture proposals for training and transferring the knowledge from the recognition performance to the super resolution components and across native and synthetic images.

Method	Approach	Reported Metric
S2R2 matching (2008) [36]	Compute and optimize a super resolution matrix by measuring the similarity of a reconstruction in the low resolution feature space, while adding smoothness to the super-resolved image.	Multi-PIE identification accuracy: 84.1% for 12×12 probles, 62.8% for 6×6 probes. Outperforms downsampling the HR image to probe resolution and bilinear interpolation matching on FERET. No efficiency metrics reported.
Uiboupin et al. Sparse Representation (2016) [96]	Based on Hidden Markov Model + SVD components [68]. Uses downsampling operator, blurring kernel, dictionary with face images and natural imagery to improve reconstruction.	FERET mean accuracy: 21.60%. No efficiency metrics reported.
Coherent Local Linear Reconstruction (2010) [42]	Use CCA to model neighboring images accross resolutions and project to HR space using PCA vectors.	PSNR on CAS-PEAL face database [27]: 31.18%, outperforms PCA reconstruction. No efficiency metrics reported.
2D CCA Face Image SR (2014) [1]	Iteratively solves the eigenvalue problem for CCA in directions X and Y for two left and right projection matrices which are used for upscaling.	CUHK dataset [99] recognition accuracy: 99.31%. Inference time: 1.38 seconds per image on a desktop 2.4GHz processor.
Deep Cascaded Bi-Network (2016) [125]	Uses a dual input branch network design to process high frequency spatial priors and the input LR image. Estimates deformation coefficients to generate the final super-resolved image. Cascaded training employed.	PSNR on Multi-PIE: 35.65. Face hallucination step: 3.84 seconds for four cascades on a single core i7-4790 CPU.
Cascaded Super-Resolution with Identity Prior (C-SRIP) (2016) [30]	For each scaling factor, it uses ResNet-based reconstruction modules and independent SqueezeNets for the identity constraint. The reconstruction modules are placed in a cascade fashion.	On LFW: 27.164 PSNR and 0.8171 SSIM. On CelebA: 26.028 PSNR and 0.7945 SSIM. Reconstruction runtime: 30ms for a single image on an NVIDIA Titan X. 30M parameter count [77].
Single Network with SR (2016) [100]	Pre-trains a super resolution sub-network in an unsupervised fashion, then a supervised fine-tuning step is performed for face recognition.	UCCS dataset rank-1 recognition accuracy: 53.69%. No efficiency metrics reported.
Robust Partially Coupled Network (2016) [100]	Fully coupled super resolution network with downsampled HR to ground truth HR image reconstructions and LR to HR image reconstruction. The huber loss [43] was used for improving VLR face recognition performance.	UCCS dataset rank-1 recognition accuracy: 59.03%. No efficiency metrics reported.
Complement Super Resolution and Identity (CSRI) (2018) [14]	Two branch network with shared network weights between the synthetic LR face image branch with supervised super resolution training and the Native LR face images without ground truth. Update both subnetworks with the classification loss.	Tinyface rank-1 recognition performance: 44.8%. No efficiency metrics reported.
Feature Adaptation Network (FAN) (2019 [112]	Dual branch architecture which encodes and concatenates a descriptor from both VLR and HR. Performs super resolution afterwards and creates a the final disentangled descriptor.	SCFace mean recognition accuracy: 90.27%. Inference time performance: 0.016 secods on Nvidia TITAN X GPU.
Dual Directed Capsule Network (2019) [90]	Extract image features using a convolution and project the VLR image features closer to the HR image feature centroid. Super resolve the projected images using a capsule network [80] architecture with a reconstruction FC module.	UCCS recognition accuracy: 95.81%. No efficiency metrics reported.
Compact Super-Resolution Network (ComSupResNet) (2020) [77]	Single branch architecture design with a 2-layer VLR feature extraction module. Asymmetric number of ResNet reconstruction blocks for each scaling factor and cascaded reconstruction modules greatly the reducing parameter count.	On LFW: 26.22683 PSNR and 0.7901 SSIM. On CelebA: 26.5572 PSNR and 0.8059 SSIM. 1M Parameter count.

To mitigate the synthetic dataset and native dataset discrepancy problem, approaches using hybrid datasets (synthetic and native) have been proposed such as [14].

a: CLASSICAL METHODS FOR SUPER RESOLUTION-BASED LOW RESOLUTION FACE RECOGNITION

The **S2R2 matching** [36] proposed by Hennings-Yeomans *et al.* features simultaneous super resolution and recognition. This method performs the synthesis by first super resolving an image with an SR matrix, then using another LR matrix to downsample the images and compare them in the low resolution space. The second component of the optimization measures the smoothness of the super-resolved image. The third component measures the difference between the features extracted from the HR ground truth and those of the super-resolved image.

Other methods for sparse representations include Yang *et al.* [110] and Zeyde *et al.* [114], which inspired the later work of **Uiboupin *et al.*** [96]. The authors proposed a sparse representation method that uses two different dictionaries: one with natural images and face images, and another with face images only. The recognition part is a 7-state Hidden Markov Model, for seven facial components, with SVD coefficients for feature extraction and recognition,

based on [68]. The LR images were modeled as a linear combination of a blurring kernel and downsampling operator, optimizing the reconstruction from the ground-truth high resolution images.

Methods based on Canonical Correlation Analysis (CCA) have also been proposed, such as **Coherent Local Linear Reconstruction SR (CLLR-SR)** [42] by Huang *et al.* and **2D CCA Face Image SR** [1] by An *et al.* On CLLR-SR the authors reconstruct particular facial details and the entire face by using CCA to model the relationships between neighboring images across resolutions. The objective is to project the face image features into a coherent subspace from PCA vectors, then making this transformation of the LR features more correlated to the HR features using the base vectors obtained by the CCA step. The latter approach proposed by An *et al.*, used 2D CCA [51], instead of 1D CCA. The 2D CCA approach yields a better recognition performance in the CAS-PEAL-R1 and CUHK datasets. 2D CCA does not reshape the image data into 1D vectors, utilizing 2D PCA [46] as a base. The projection coefficient is divided into left and right projections, both for each dimension, and is later optimized. The authors reported an average time for super resolving one face image at 1.38 seconds on a 2.4GHz CPU. This is an improvement in super resolution performance from the spatial representation method [47], which had the closest face recognition performance in the reported datasets.

b: DEEP LEARNING METHODS FOR SUPER RESOLUTION-BASED VERY LOW RESOLUTION FACE RECOGNITION

In a study conducted by Wang *et al.* [100], they proposed and evaluated the various image recognition models. In the **Single Network with Super Resolution Pre-training** model, the authors pre-trained an unsupervised super resolution network and posteriorly fine-tuned it with the supervised recognition component (two fully connected layers and softmax classifier on top). The model that yielded the best recognition performance was the **Robust Partially Coupled Network** model. They noted that pre-training using Super Resolution methods was insufficient for recognition purposes and that data augmentation or data adaptation is needed for improving face recognition accuracy performance. In this model, they built on top of the previous one, and added an HR to HR reconstruction such that the filter information gets transferred to the LR to HR reconstruction subnetwork. This makes it a partially-coupled super resolution network. After this, they used the Huber loss [43] instead of the MSE loss to further improve face recognition performance, due to its lower sensitivity to outliers.

The **Deep Cascaded Bi-Network** [125] train using a cascaded process. This cascaded process is done by sequentially incrementing the scaling factors, usually by $2\times$, in each super-resolution step. The Deep Cascaded Bi-Network introduces the concept of a high frequency priors, generated using a mean face template used to estimate the average warping of the training images. The network design features two input branches, one for processing the LR image and the other one to process the LR image with the high frequency priors. Afterwards, the resulting maps are gated with the mean face template and the estimated deformation coefficients to generate a third feature map. This third feature map, the image residual, is multiplied with the output of the previous two branches. The final hallucinated face image is the addition of these two outputs. The authors reported an improvement in PSNR performance over bicubic interpolation, PCA, and CSDN [101] on the MultiPIE dataset. The reported runtime for performing face hallucination using four cascades is of 3.84 seconds on a single core i7-4790 CPU with the authors noting that the bottleneck is the gated subnetwork with higher-dimensional super-resolved feature maps.

The **Cascaded Super-Resolution with Identity Prior (C-SRIP)** [30] consists of three sequential super resolution modules for a cascaded training methodology. Each module is a ResNet-based subnetwork with a Convolution, Leaky-ReLU, and sub-pixel convolution at the end to super-resolve the image. These modules are trained with separate pre-trained SqueezeNets to introduce identity priors at each resolution. The architecture for this approach is described on Figure 8. The proposed final loss function is the sum of the cross-entropy loss for the predicted identity from the SqueezeNets at each scale and a multi-scale SSIM loss. The authors reported better PSNR and SSIM results than LapSRN [49] and SRGAN [50] for the LFW and CelebA datasets, when using only the CASIA-WebFace dataset to train. The authors also report an average runtime of 30ms per image for super resolving a single image using an NVIDIA Titan X GPU.

The **Complement Super-Resolution and Identity (CSRI)** [14] method, detailed in Fig. 9, uses two subnetworks with two parallel branches: one processing the synthetic low resolution images and the high resolution ground truth counterpart and one processing the native low resolution image recognition only. The key aspect of this architecture is the shared parameters from both branches. The authors reported results for their proposed TinyFace benchmark dataset. The complement super resolution learning strategy (shared weights) yields an 8% Rank-1 accuracy increase over an independent weights strategy. They report a 10.1% of Rank-1 accuracy increase when training the recognition network jointly with the super resolution network, as opposed to performing these tasks separately.

The **Feature Adaptation Network (FAN)** [112] bridges the domain gap using a dual branch architecture, disentangling face component feature learning and posteriorly decoding and re-encoding the face image. This architecture is a GAN-based approach, shown in Fig. 10. The authors propose to train the network using native and synthetical VLR image pairs. The synthetical pairs are created by randomly downscaling the HR images. Encoding face and non-identity features separately allows for optimizing the closely related facial identity features, while disregarding the rest at the loss function level. As such, the identity features are the only ones

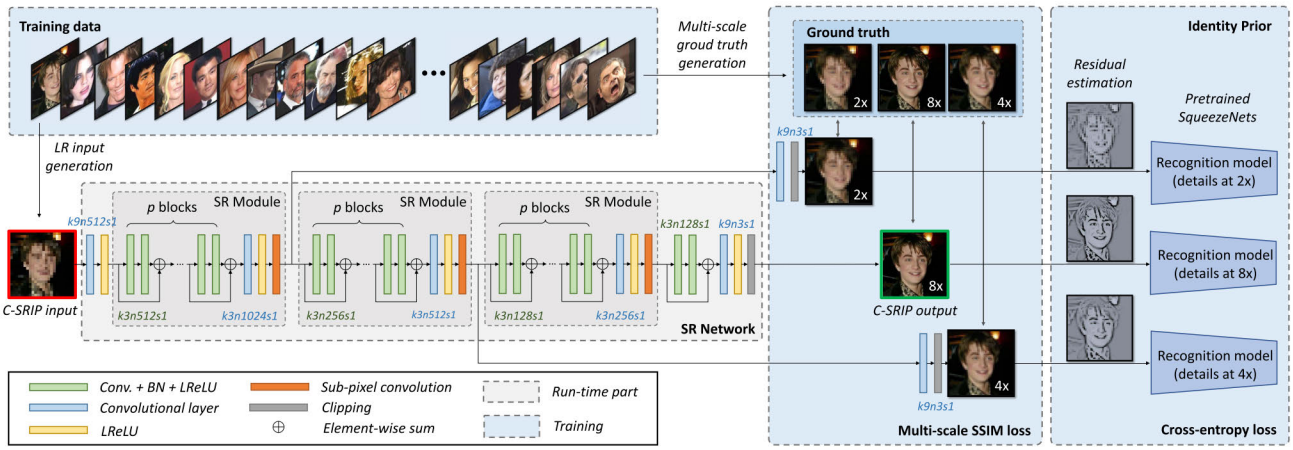

FIGURE 8. Architecture of the Cascaded Super-Resolution with Identity Prior (C-SRIP) network, taken from [30]. This design features the sequential Super Resolution modules to train in a cascade fashion, with separate pretrained SqueezeNets for every magnification factor. The Super Resolution modules are based on ResNet blocks.

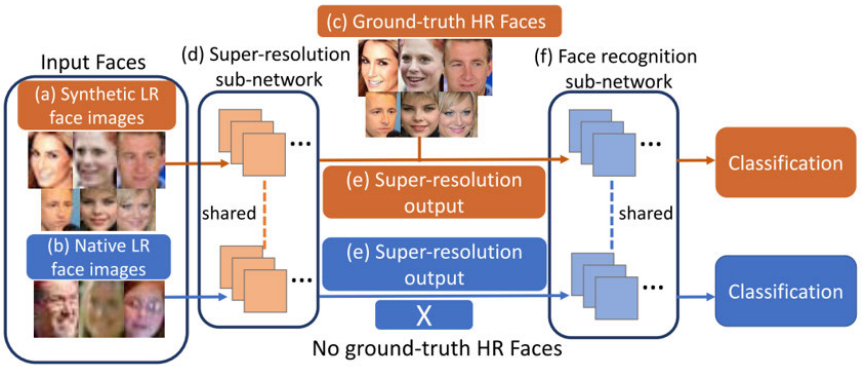

FIGURE 9. Overview of the Complement Super-Resolution and Identity Network (CSRI), taken from [14]. The network is comprised of two subnetworks: one for synthetic LR images with supervised training and one for Native LR face images without supervision for super resolution. Both of these networks share their parameters, learning from the classification accuracy and the super resolution accuracy.

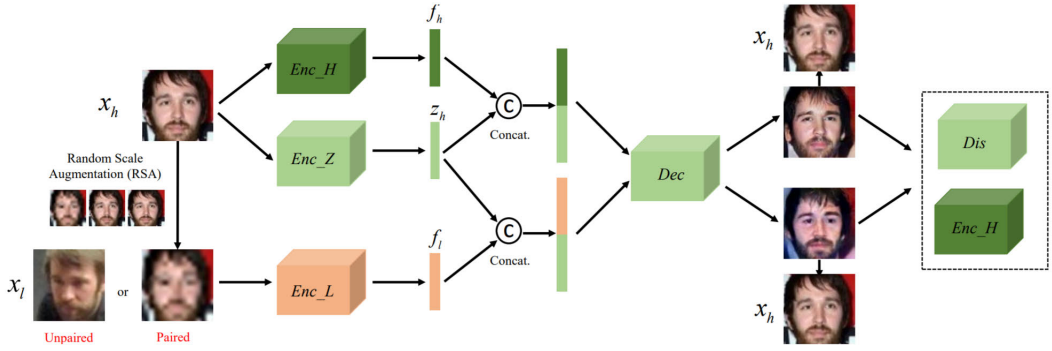

FIGURE 10. Architecture overview of the Feature Adaptation Network (FAN), taken from [112]. This method consists of two input branches for HR and LR images, extracts features for both of them, and generates two descriptors. These descriptors are for the HR image and the LR + HR images. After that, it performs two reconstructions: one for identity features and another one for non-identity features.

used for classification. The rest of the features are used only for reconstruction. The authors reported an average accuracy of 90.27% in the SCFace dataset, and an inference time of 0.016 seconds for a single face image in an NVIDIA Titan X GPU.

Another recently proposed method is the **Dual Directed Capsule Network for Very Low Resolution Image Recognition** [90], depicted in Fig. 11. Firstly, this method utilizes native low resolution images and upscales them using bilinear interpolation for training. It also downscales the native

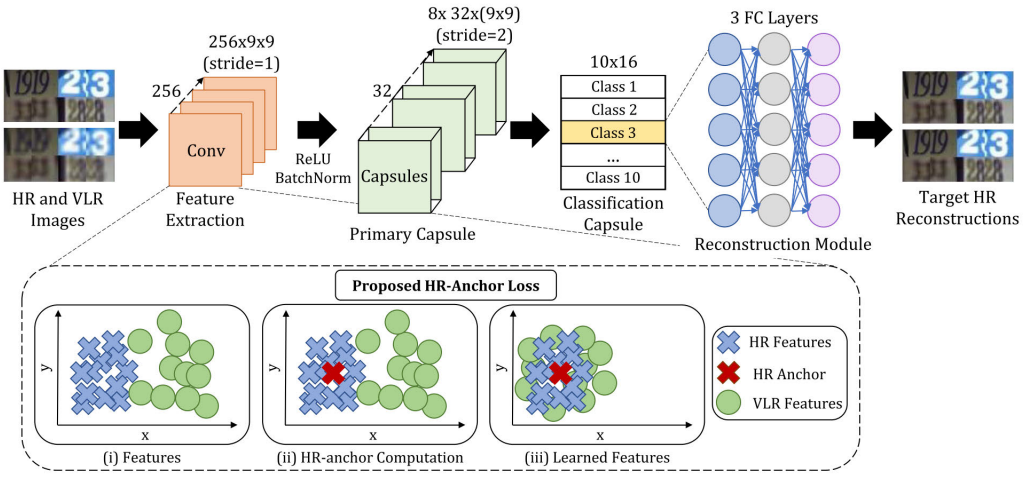

FIGURE 11. Overview of the Dual Directed Capsule Networks for Very Low Resolution Image Recognition, taken from [90]. The authors proposed to project the extracted VLR image features closer to the HR image feature centroid using an HR-Anchor loss as a first step. Then, they used these projected features to feed the capsule network component, which performs the super resolution reconstruction with three fully connected layers.

high resolution images to low resolution. The authors utilize a feature extraction strategy by taking any high resolution information as an ''anchor'' to guide the feature extraction of low resolution native images. This is done by introducing a novel ''High Resolution-anchor'' loss function and also by propagating the recognition gradient in the feature extraction and posterior super resolution stages. The anchor value is learned using every high resolution sample of its class and then is used to modify the extracted low resolution feature such that it gets closer to the anchor value. An important note is that the method does not use the low resolution information to learn the anchor at any time. Even though this method uses projection methods akin to coupled mapping techniques, the network is trained with a reconstruction loss at the end, with the image reconstruction network section at the end. This approach was tested on various recognition datasets, one of them being the UCCS dataset, which is also representative of the problem at hand, achieving a Rank-1 accuracy of 95.81%.

The **Compact Super-Resolution Network (ComSupResNet)** [77] is a single network architecture which extracts VLR features using only two layers, with the first one comprising of a 9×9 filter with boundary padding. Afterwards, the reconstruction modules are presented in a cascade training fashion, incrementing the scale factors by $2\times$ after each module. The internal reconstruction blocks in each module have residual connections. However, the number of blocks in at each module is different, resulting in a decreased number of parameters. Figure 12 illustrates this architecture. The authors report an improvement on PSNR and SSIM performance against C-SRIP and LapSRN for the CelebA dataset. They also achieve the second-best PSNR and SSIM performance on the LFW dataset against C-SRIP with a fewer number of parameters (1M vs 30M parameters).

c: DISCUSSION ON SUPER RESOLUTION APPROACHES

The most critical limitation of super resolution methods is the increased training and testing time. By the nature of the approach, these methods need to upsample and then perform classification to be effective. This encoding process, decoding and re-encoding is costly, so much more on deep learning-based approaches. This is the case of popular methods based on Generative Adversarial Networks such as [50] and [113]. Deep learning super resolution approaches rely on two or more subnetworks to process HR and VLR input images. In the particular case of FAN, it is only possible to perform inference in real-time using expensive GPUs such as the Nvidia Titan X. Efficient super resolution methods for face recognition is still an opportunity area. As noted by [72], some methods include down-sampling and posterior deconvolution to alleviate the computational load. By doing so, the dimensionality of the intermediate representations is heavily reduced, making it possible to run inference in real-time using GPUs. Such is the case of [23] and [87]. However, the advantages of super resolution methods are the novel architectures such as CSRI, FAN, and Dual Directed CapsNet. These innovations come in the form of distilling synthetic and native VLR knowledge using attention layers (CSRI, RPCN), face feature vector components (FAN), and feature projection strategies before reconstruction (Dual Directed CapsNet). One of the most important notions of these methods is using HR anchors to optimize VLR feature extractors to this richer HR feature space.

B. HOMOGENEOUS EFFICIENT APPROACHES FOR VERY LOW RESOLUTION FACE RECOGNITION

Deep learning methods have shown to be effective for general-purpose recognition tasks. However, their complexity

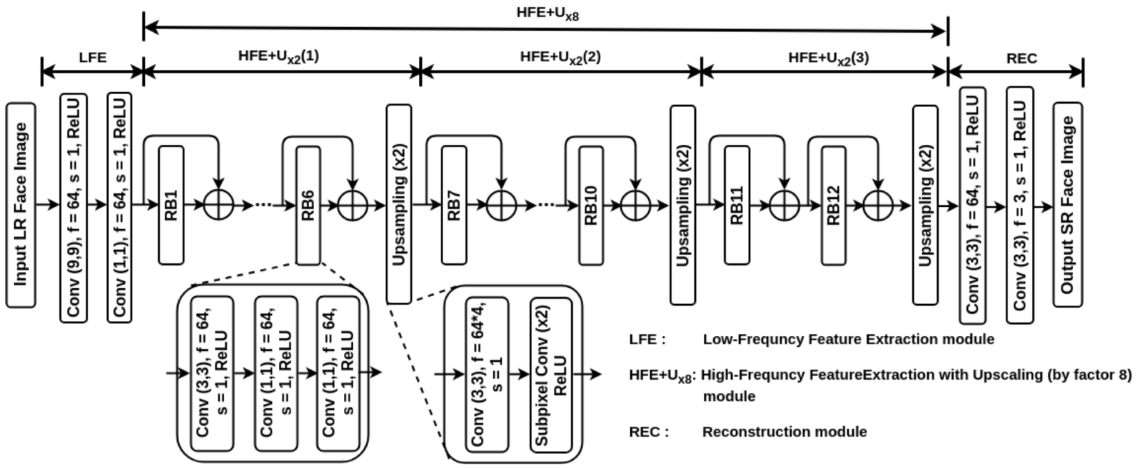

FIGURE 12. The Compact Super-Resolution Network (ComSupResNet), taken from [77], features a single architecture design for super resolving VLR images. The VLR extraction module features two layers while the reconstruction modules feature a variable number of residual blocks, resulting in a 1M parameter count.

requirements make them unfeasible for using them on real-time scenarios. In consequence, methods for achieving real-time performance in other computer vision tasks on embedded devices have emerged, such as **SqueezeNet** [44], **ShuffleNet** [123], **ShuffleNetV2** [64], **MobileNet** [40], **MobileNetV2** [83], **MobileNetV3** [39], and **VarGNet** [121]. This set of methods were proposed for efficiently solving general computer vision tasks such as image recognition, object detection, and others. These methods are commonly called Lightweight Convolutional Neural Networks. Most recently, face recognition variants of these methods were proposed. We will discuss these methods in-depth in the next subsection II-B1.

Fig. 13 shows a multiply-addition operations (MAdds) benchmark against accuracy performance for these computer vision tasks to give a general idea of where they stand in between each other. MobileNetV3 currently claims to be the network with the best efficiency-accuracy network for general-purpose computer vision tasks as per their evaluation [39]. This metric gives us a general idea of the relative performance between each other. However, the accuracy generally depends on a specific dataset and problem.

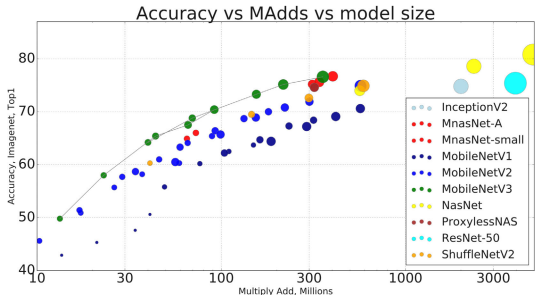

FIGURE 13. Accuracy against MAdds benchmark for lightweight general-purpose networks for computer vision, taken from [39]. Currently, MobileNetv3 shows the best trade-off between multiply-addition operations against accuracy, followed closely by MnasNet and MobileNetv2.

1) LIGHTWEIGHT CONVOLUTIONAL NEURAL NETWORKS FOR FACE RECOGNITION

Lightweight neural networks tailored specifically for the face recognition task have been proposed, such as **MobileFaceNet** [13], **ShuffleFaceNet** [66], and **VarGFaceNet** [107].

In general, these techniques are based on lightweight general-purpose CNN architectures. In order to have a better efficiency-accuracy trade-off, these lightweight networks employ the following techniques: using grouped convolutions and shuffling the output channels to make to reduce the number of operations and share information across different input and output channels, using variable groups of grouped convolutions to balance between information retaining and complexity, point-wise 1×1 convolutions to reduce depth channels and computational complexity while filtering, having low-dimension embeddings before the fully connected layers, using strides instead of max pooling operations to reduce complexity and retain more information directly from the data and using inverted bottleneck structures to reduce the number of parameters and compact the network channels again to match the input channels.

Table 4 shows an overview of the most recent face recognition networks based on mobile CNN architectures, the techniques they use, and their computational footprint in FLOPs. These face recognition-specific approaches follow several guidelines for optimized face recognition performance such as: using Global DepthWise convolutions instead of global average pooling [13], using additive-angular margin-based loss functions [21], and lightweight design ideas. These ideas aim to retain as much fine-grained information as possible by avoiding the use of max pooling [66]. These techniques have proven to be beneficial for face recognition performance and may not be for general computer vision tasks.

TABLE 4. Summary of some of the most recent and most successful efficient lightweight CNN architectures for face recognition. ShuffleFaceNet greatly improves efficiency performance with its compact 128-D feature vector size and a lesser deep architecture while maintaining comparable accuracy performance with VarGFaceNet and MobileFaceNet.

Method	Approach	Efficiency optimizations	Complexity remarks
MobileFaceNet [13]	Based on MobileNetV2 [83]	Global DephtWise convolution instead of Global Average Pooling layer, stride=1 after conv1, 1280-D feature vector	933.3M FLOPs
ShuffleFaceNet [66]	Based on ShuffleNetV2 [64]	Global DephtWise convolution after Conv5, PReLU activation, add strides and eliminate pooling at conv1, compact 128-D feature vector	577.5M FLOPs
VarGFaceNet [107]	Based on VarGNet [121] Knowledge distillation	Teacher-student network architecture, set channel number as constant in a group, variable number of groups, PReLU activation, point-wise conv before FC layer, 512-D feature vector	Teacher model: 24GFLOPs Student Model: 1022M FLOPs

VarGFaceNet [107] is a lightweight method based on VarGNet. In contrast to VarGNet, VarGFaceNet firstly expands the channels to 1024, using 1×1 convolutional layers to preserve information. After that, it uses variable group convolutions and point-wise convolutions to reduce the intermediate representations. The authors managed to reduce run-time on CPU and GPU to 31 fps on the method variant that they tested on the SCface dataset, where they reported a rank-1 accuracy of 43.5%. It does not achieve the same performance as other heavier Deep Learning methods, but it does outperform the classical hand-crafted methods that appeared at the start of this decade and earlier deep learning-based methods that were starting to emerge.

2) KNOWLEDGE DISTILLATION AND QUANTIZED NETWORKS

Approaches based on teacher-student networks for training and efficient inference, named knowledge distillation methods, have also emerged in recent years. VarGFaceNet [107], discussed in the previous subsection, uses knowledge distillation for training. This method uses recursive knowledge distillation with the Angular Distillation loss function. The selected teacher network is a ResNet architecture, which is used for feature vector extraction. These feature vectors are then used in the VarGFaceNet loss function to draw the feature vectors generated by the lightweight network closer to those of the teacher network. The angular distillation methodology uses similarity measurements such as L2 or cosine similarity to score both feature vectors. An attempt for efficient very low resolution face recognition using knowledge distillation was proposed in [28]. Their method uses a teacher model based on VGGFace and uses a student network with a simpler design for inference. The authors also used manual resizing to train the student network. The output of the teacher network uses a graph-based approach to accept and reject persons that do not appear in training. The approach performs better than other methods from the state of the art, such as [91] and [70], but not from the ones mentioned in the previous Coupled Mappings section, even classical hand-crafted approaches. We consider that the most significant drawbacks of this approach are the fact that there is no re-projection and that the images available for training in the low resolution space are the resized gallery images. When testing on a very challenging database such as the SCface one, it does not achieve desirable performance.

Another approach to reduce inference complexity has been quantization, such as the works from [79], [109], and [45]. These approaches focus on reducing parameter number representation by limiting the number of representation bits, where floating-point convolutional layers can be quantized into binary convolutional layers. In consequence, this reduces complexity at the knowledge representation level. Efficiency is mainly achieved by using bit-wise convolutional layers and by changing activation functions to binary operators. However, these approaches present a challenge with balancing the k-bit quantization representation, the network depth, and signal preservation. Recent literature has several points to consider to improve quantized network performance [65]. These considerations are: using learnable scaling factor vectors, residual connections for quantized layer blocks, using the PReLU activation function, among others.

3) DISCUSSION ON EFFICIENT APPROACHES FOR HOMOGENEOUS VLR FACE RECOGNITION

The most important advantage of lightweight convolutional neural networks is the efficiency-accuracy trade-off. Most of these networks achieve it by firstly expanding the channels and then using fast downsampling strategies. These fast downsampling strategies come in the way of point-wise convolutions or using strides directly closer to the input layer. Extending knowledge to the different channels in the intermediate representations is also important. In this respect, both VarGFaceNet and ShuffleFaceNet are based on different strategies for propagating knowledge to different channels, where the variable group convolutions yield better accuracy performance. These networks feature a strong and robust feature extraction, due to them being deep enough for reaching a FLOPs threshold (1.2GFLOPs) while not degrading the image representations. This is achieved by using linear combinations of blocks instead of the traditional convolution and

TABLE 5. Face identification rank-1 accuracy results for the SCface dataset. Compiled from [54] and [67]. Each column marks the horizontal distance of the camera to the subject, where d1 of 4.2 meters is the farthest from the subject and, as such, the most challenging one for heterogeneous face recognition purposes. The networks marked with "FT" denote that they were fine-tuned for the SCface dataset. The models marked with an asterisk * were not trained or fine-tuned in the SCface dataset but rather the MS-Celeb-1M dataset. [32].

Method	d1 (4.2m)	d2 (2.6m)	d3 (1.0m)	Mean accuracy
SCface [29]	1.82	6.18	6.18	4.73
CLPM [52]	3.46	4.32	3.08	3.62
CSCDN [101]	6.99	13.58	18.97	13.18
SSR [111]	7.04	13.2	18.09	12.78
L2softmax [78]	9.20	18.80	16.80	14.93
CCA [102]	9.79	14.85	20.69	15.11
DCA [33]	12.19	18.44	25.53	18.72
LM Softmax [117]	14.00	16.00	18.00	16.00
AM Softmax [98]	14.80	20.8	18.4	18.00
C-RSDA [18]	15.77	18.08	18.46	17.44
RIDN [19]	23.0	66.0	74.0	24.96
LDMDS [108]	62.7	70.7	65.5	66.30
VGG-Face* [67]	41.3	75.5	88.8	68.53
LightCNN* [67]	35.8	79.0	93.8	69.53
Centreface* [67]	36.3	81.8	94.3	70.87
VGG-Face-FT [67]	46.3	78.5	91.5	72.10
ResNet50-ArcFace* [67]	48.0	92.0	99.3	79.77
LightCNN-FT [67]	49.0	83.8	93.5	75.43
Centreface-FT [67]	54.8	86.3	95.8	78.97
FAN* [113]	62.0	90.0	94.8	82.27
ShuffleFaceNet* [67]	55.5	95.3	99.3	83.37
MobileFaceNetV1* [40]	57.0	95.3	99.8	84.03
ResNet50-ArcFace-FT [67]	67.3	93.5	98.0	86.27
MobileFaceNetV2* [67]	68.3	97.0	99.8	88.37
DCR-FT [62]	73.3	93.5	98.0	88.27
TCN-ResNet-FT [115]	74.6	94.9	98.6	89.37
FAN-FT [112]	77.5	95.0	98.3	90.27
ShuffleFaceNet-FT [67]	86.0	99.5	99.8	95.10
MobileFaceNetV2-FT [67]	**95.3**	**100.0**	**100.0**	**98.43**

max pooling operations. Due to their intended application, the network depth and width become the most significant limitations of these networks. Furthermore, these fast downsampling strategies are not the best when dealing with native low resolution imagery due to the inherent loss of data. In the case of ShuffleFaceNet, the more aggressive stride step can lead to additional loss of information. As such, it is necessary to use different upsampling methods to simulate different low resolution scenarios in order to overcome these limitations.

III. RESULTS AND DISCUSSION ON STATE-OF-THE-ART FACE RECOGNITION PERFORMANCE ON UNCONSTRAINED SURVEILLANCE SCENARIOS

In this section, we present, discuss, and analyze the state-of-the-art results of the methods mentioned in the previous sections in the context of unconstrained very low resolution face recognition. We present results for the SCFace and UCCS-Face datasets in particular, as they are very representative of the heterogeneous VLRFR problem with the research challenges mentioned in Section III-C. We also present results for the QMUL-SurvFace and QMUL-TinyFace, which are newer benchmarks representing the native VLR homogeneous face recognition problem. These datasets contain native very low resolution imagery of 32 × 32 pixels of facial region-of-interest or below, which make them suitable for representing the unconstrained VLR FR problem.

TABLE 6. Face verification results on the UCCS dataset, taken from [16].

Method	TAR@FAR 30%	10%	1%	0.1%	AUC (%)
DeepID2 [16]	93.1	83.4	61.7	37.9	93.8
CentreFace [16]	**99.6**	**97.0**	**87.8**	**75.5**	**99.0**
FaceNet [16]	98.2	93.8	79.4	63.4	97.8
VGGFace [16]	97.1	90.6	72.4	55.1	96.7
SphereFace [16]	94.0	84.9	60.2	24.7	94.1

TABLE 7. Face identification results on the UCCS dataset, taken from [90] and [54], for matching HR (80 × 80) vs. VLR (16 × 16) images.

Method	Mean Acc(%)
Robust Partially Coupled Nets [100]	59.03
Selective Knowledge Distillation [28]	67.25
LMSoftmax for VLR [54]	64.90
L2Softmax for VLR [54]	85.00
Centreface [54]	93.40
DualDirectedCapsNet [90]	**95.81**

A. RESULTS AND DISCUSSION ON VLR FACE RECOGNITION PERFORMANCE

We compiled the mean recognition accuracy results for the SCface dataset in Table 5 to better illustrate the state-of-the-art heterogeneous VLR FR panorama. The SCface dataset contains 130 subjects and 4,160 images from different surveillance camera distances: d1(4.2m), d2(2.6m), and d3(1.0m). For this dataset, d1 is the most challenging setting, where the average facial region-of-interest is below 32 × 32 pixels.

For the UCCS dataset, we show the results of state-of-the-art face recognition methods for face verification and face identification, shown in Table 6 and Table 7. The UCCS dataset setting represents a real-world surveillance setting, where the image resolution varies from VLR to HR.

For the homogeneous variant of the problem, we show the face identification results of the QMUL-TinyFace dataset in Table 8 and the face verification results of the QMUL-Survface dataset in Table 9. The TAR@FAR verification results were calculated as per the specifications in the original paper [16] and the provided code [15]. The QMUL-TinyFace dataset contains unconstrained face images from the web with 5,139 identities and 169,403 images. The QMUL-Survface contains native surveillance images from 15,573 subjects and 463,507 images. The image input resolution of both datasets is variable but always below 32 × 32 pixels.

For the results reported in this section, we discuss three critical areas of opportunity regarding the state-of-the-art concerning face recognition performance: training methodologies, generalization capabilities, and type of approach.

TABLE 8. Face identification results on the QMUL-TinyFace dataset, taken from [67]. ShuffleFaceNet and MobileFaceNet were trained on the MS1-Celeb-A dataset and fine-tuned on the QMUL-TinyFace training set.

Method	Rank-1	Rank-20	Rank-50	mAP
DeepID2 [14]	17.4	25.2	28.3	12.1
SphereFace [14]	22.3	35.5	40.5	16.2
VGG-Face [14]	30.4	40.4	42.7	23.1
CentreFace [14]	32.1	44.5	48.4	24.6
ShuffleFaceNet [67]	43.1	58.9	64.5	34
MobileFaceNet [67]	**48.7**	**63.9**	**68.2**	**40.3**

TABLE 9. Face verification results on the QMUL-SurvFace for TAR@FAR 1% and 0.10%, taken from [67]. ShuffleFaceNet and MobileFaceNet were trained on the MS1-Celeb-A dataset and fine-tuned on the QMUL-SurvFace training set.

Method	TAR@FAR		AUC	Mean Acc
	1%	0.1%		
VGG-Face [16]	20.1	4	85.0	78
DeepID2 [16]	28.2	13.4	84.1	76.1
SphereFace [16]	34.1	15.6	85.0	77.6
FaceNet [16]	40.3	12.7	93.5	85.3
CentreFace [16]	**53.3**	26.8	**94.8**	**88.0**
ShuffleFaceNet [67]	38.5	11.9	89.9	82.3
MobileFaceNet [67]	52.9	**33.1**	89.9	83.2

1) DISCUSSION ON TRAINING METHODOLOGIES

Face identification and verification performance are heavily affected by the experimental methodology of every study. Even though we can see that Lightweight CNN approaches achieve the best mean accuracy results, there are still areas of opportunity present within their training and testing methodologies. In order to achieve the best accuracy in the SCface dataset, the authors of [67] pre-trained the networks in the MS1-Celeb-1M dataset and then fine-tuned them using the SCface dataset with different upsampling strategies. In the rest of the evaluations, these different upsampling methodologies were not employed and, as such, we do not see any performance increase. In other methods, such as in Deep Coupled ResNet [62], the images from the SCface dataset were upsampled to three different resolutions, to add robustness to the learned representation, effectively synthesizing the dataset to three times its original size with no previous pre-training. Another method, DMDS [108], randomly selects 50 subjects from the SCface dataset and uses them to train, without upscaling the VLR images. This generates a gap in evaluating the real limitations of the methods. In consequence, there always exists a compromise when selecting a scaling approach for every different method, thus, making them difficult to compare even when using the same datasets. This is true for both the heterogeneous and homogeneous variants of the problem.

We recognize that one of the limitations of deep learning methods is the need of having very large datasets in order to train the networks effectively, which is why we support the idea of pre-training the networks on other face recognition datasets such as LFW [41] or MS-Celeb-1M [32] and other benchmark datasets. As such, another area of opportunity exists in analyzing if a particular pre-trained dataset yields better results for VLR face recognition before fine-tuning to a specific benchmark dataset. As per the results of Tables 5, 6, 8, 9, we can see that VGG Face [74] is unable to learn effective representations for the very low resolution settings even after fine tuning. In contrast, newer and more efficient feature extraction methods can these representations very well. Using VGG Face-like architectures as a base for VLRFR is not an effective strategy and not all the seminal face recognition CNN architectures perform well under this context even after fine-tuning. Such is the case of the TCN variant with VGG Face used in the feature extraction process.

2) DISCUSSION ON GENERALIZATION CAPABILITIES

Furthermore, to bring these methods to a real-world application scenario, the generalization capabilities of these methods need to be tested as completely as possible for different evaluation metrics. Cross-dataset evaluations are needed when fine-tuning to a particular dataset, to avoid reporting results based on the CNN dataset memorization. This would lead to a more fair analysis of which methods effectively mitigate the challenges presented in Section III-C and compare them under the same generalization conditions. Due to the extremely challenging nature of the problem at hand, it is essential that the learned representations are as robust as possible for an open-set identification scenario. In Tables 5, 9, and 8, we can see independent evaluations for VGG-Face, CentreFace, ShuffleFaceNet, and MobileFaceNet. However, the results are higher on d1 for the SCface dataset for ShuffleFaceNet and MobileFaceNet than the rest of the datasets, even when fine-tuning, due to the interpolation methods used in [67]. This clearly shows the limitations of the state-of-the-art methods at VLR without simulating VLR conditions beforehand. Furthermore, verification metrics of the SurvFace and UCCS datasets in Tables 9 and 6 show that even though the mean accuracy and the AUC can be very high, the True Acceptance Rate at 0.1% remains a challenge.

3) DISCUSSION ON APPROACH TYPES

We can also appreciate from Table 5 that the top performing solutions in average accuracy performance for the SCface dataset are CNN-based approaches. Furthermore, the best approaches in terms of mean accuracy performance are the robust homogeneous feature extraction using lightweight CNN architectures. CNN-based methods tend to generalize very well at distances d2 and d3, which are the lesser challenging distances compared to d1. Explaining why they yield a higher mean accuracy than traditional methods. Traditional methods tend to be more consistent across the three distance settings, with LDMDS [108] standing out as the best for Coupled Mappings. This method has a consistent accuracy performance across camera distances, reaching a 62.7% accuracy for the most challenging scenario. This method effectively demonstrates that enforcing the large margins between

TABLE 10. Inference time for Lightweight Convolutional Neural Networks, by CPU inference descending time in milliseconds. The methods yielding inference time of less than 100ms are viable for servicing at least one surveillance camera in real-time. At this time, ShuffleFaceNet with 1.5 of depth multiplier [66] provides the best mean accuracy-efficiency trade-off for real-time CPU performance as per the efficiency results in this table and the results of the previous subsection.

Network	# Params. (Millions)	2× GTX 1080ti	GTX 1080Ti	GTX 1660Ti	Quadro P2000	Laptop GTX 1050Ti	Laptop Intel i7 7700HQ
Light CNN - 4 [106]	6.8 M	5.49 ms	12.67 ms	14.09 ms	41.00 ms	55.50 ms	2,653.76 ms
Light CNN - 9 [106]	8.1 M	6.22 ms	14.36 ms	15.88 ms	40.96 ms	56.17 ms	2,106.72 ms
VGG [89]	144.9 M	3.39 ms	7.83 ms	108.97 ms	25.17 ms	35.29 ms	1,523.40 ms
VGGFace [74]	41.1M	3.99 ms	9.22 ms	14.24 ms	19.64 ms	21.61 ms	433.39 ms
Resnet100 [34]	65.2M	2.76 ms	5.26 ms	12.96 ms	48.14 ms	59.61 ms	285.64 ms
Light CNN - 29 [106]	31.0 M	2.01 ms	4.63 ms	2.38 ms	7.95 ms	7.93 ms	126.71 ms
VarGFaceNet [107]	4.9 M	0.85 ms	1.48 ms	3.48 ms	5.10 ms	27.09 ms	126.59 ms
MobileNetv2 [83]	**1.8 M**	0.80 ms	1.46 ms	4.50 ms	11.08 ms	14.76 ms	103.69 ms
MobileNetv1 [40]	3.2 M	**0.69** ms	1.21 ms	1.88 ms	5.56 ms	20.06 ms	98.99 ms
VarGNet [121]	4.2 M	0.68 ms	1.18 ms	2.16 ms	**4.21** ms	5.42 ms	70.40 ms
MobileFaceNetV2 [13]	2.0 M	0.88 ms	1.48 ms	3.27 ms	5.55 ms	7.28 ms	62.45 ms
MobileFaceNetV1 [13]	3.3 M	0.74 ms	1.31 ms	1.61 ms	4.91 ms	8.23 ms	53.49 ms
ShuffleNet - 2.0 [64]	5.3 M	1.10 ms	1.96 ms	18.79 ms	25.05 ms	N/A	42.92 ms
ShuffleFaceNet - 2.0 [66]	4.5 M	1.00 ms	1.77 ms	2.41 ms	6.36 ms	6.95 ms	37.46 ms
ShuffleNet-1.5 [123]	2.5 M	0.77 ms	1.33 ms	2.77 ms	5.29 ms	12.25 ms	32.98 ms
ShuffleFaceNet-1.5 [66]	2.6 M	0.77 ms	**1.34 ms**	**1.86 ms**	4.75 ms	**4.68 ms**	**29.08 ms**

classes is an effective strategy, a similar notion to the additive margin loss functions [98] and [21], used for face recognition CNNs. Super resolution methods tend not to be present in this table because they are trained and tested in other datasets, such as the UCCS dataset. Furthermore, most of the time, these approaches are focused on improving the SSIM and PSNR metrics rather than recognition performance. However, a clear standout is the FAN architecture, which demonstrates the effectiveness of using a GAN-like approach by super resolving images after performing the disentangled feature learning steps. Table 7 shows face identification results in the UCCS dataset for the Super Resolution methods Robust Partially Coupled Nets and Dual Directed Capsule Network. The Dual Directed Capsule Network is a clear standout for its accuracy performance in this challenging scenario, outperforming CentreFace, performing closer to MobileFaceNet in the QMUL-SurvFace verification results of Table 9.

B. RESULTS AND DISCUSSION ON EFFICIENCY PERFORMANCE

From the methods in the previous face recognition performance subsection, we report the run-time for the efficient lightweight convolutional neural networks in Table 10. This efficiency table gives a general panorama of where lightweight neural networks stand for real-world inference time performance.

When comparing efficiency performance, it is very important to run tests standardized to hardware architectures. Hardware-agnostic metrics such as FLOPs and the number of parameters of a network are often used as indicators to compare network efficiency between proposed architectures. This poses a problem since these metrics do not translate linearly to real run-time performance metrics at any specific hardware configuration. Other considerations, such as the number of times any given architecture has to access memory, GPU/CPU memory size, and memory bandwidth, are significant bottlenecks that affect real-time performance. Furthermore, some authors from earlier literature do not refer to the specific hardware configuration used when reporting time performance in seconds. Even though it is hard to compare the performance between different hardware, even across different CPU generations from the same vendor, it still provides a better idea of which hardware implementation can run the proposed methods at any given scenario successfully.

Focusing on real-time hardware performance, Table 10 shows specific hardware configurations for GPUs and one laptop CPU. Using the same hardware configurations as the authors of [66], we included run-time evaluations for VarGFaceNet, which was featured at the LFR@ICCV2019 Challenge [22], and its base network VarGNet [121]. For affordable hardware, it is possible to achieve real-time recognition performance using low-power GPUs such as a Laptop GTX 1050Ti for almost all the networks described in the table, except for ShuffleFaceNet, which has a larger memory footprint. MobileFaceNetV1, MobileFaceNetV2, and ShuffleFaceNet are the best candidates for running a recognition application in real-time on affordable laptop CPU hardware. Approaches such as MobileNetV1 and ShuffleFaceNet are able to service two cameras at the same time, at least at ten frames-per-second for each one of them, where the accuracy performance detailed in Table 5 favors ShuffleFaceNet. The compact 128-D face descriptor of ShuffleFaceNet favors a substantial increase in performance, while preserving the identity information, favoring a fine-tuned scenario. MobileFaceNetV2 generalizes better, as per Table 5 and can achieve real-time performance for servicing one camera on the Laptop CPU described. As such, MobileFaceNetV2 is a better choice in

terms of accuracy performance, if we use the SCface dataset for fine-tuning and compare face descriptors for other identities foreign to that specific dataset.

Studying the structures of these networks, we can observe that the critical element to balancing the accuracy and efficiency trade-off using is focusing the recognition process on the center part of the aligned face image, using a fast downsampling strategy, such as the one in ShuffleFaceNet, and using a low-dimension face descriptor.

In general, traditional Coupled Mapping methods are more efficient than Super Resolution methods. Deep Learning-based approaches have surpassed the accuracy of traditional methods, however, an interesting case is LMCM [116]. The authors report an inference time of 8.5 microseconds on an older CPU than the one used in our test. This approach is more efficient than any of the Deep Learning approaches shown in Table 10. It also holds a better accuracy performance at the d1 scenario than the lightweight CNNs without fine-tuning, except for MobileFaceNetV2. In the case of Super Resolution methods, the efficiency performance is worse due to the need for more complex networks to up-scale the image to HR resolution and then perform identification or verification tasks. The FAN [112] approach, a Super Resolution method, has an inference time of 0.016s for inference in a much more powerful single GPU than the ones used in our tests, the Nvidia Titan X GPU. This inference time is comparable to the inference time performance of ShuffleNet with 1.5 depth on a laptop GTX 1050Ti GPU, far less powerful than the Titan X GPU.

C. DISCUSSION ON THE RESEARCH CHALLENGES AFFECTING PERFORMANCE ON UNCONSTRAINED VERY LOW RESOLUTION FACE RECOGNITION SCENARIOS

In this section, we discuss and analyze the specific research challenges at very low resolution settings. These challenges include dataset availability for real-world examples, lack of discriminative features, domain discrepancy, and the current landscape on efficient solutions.

1) LACK OF DATASET AVAILABILITY FOR REAL-WORLD NATIVE EXAMPLES AT VERY LOW RESOLUTIONS

One of the hardest challenges to train and evaluate at VLR native settings is the small number of publicly available datasets of native real-world VLR imagery. In the case of homogeneous face recognition, datasets such as TinyFace [14] and QMUL-Survface [16], are publicly available. For the heterogeneous variant, datasets such as SCface [29], Point and Shoot (PaSC) [4], and UCCSface [84], are publicly available. Another dataset, the IJB-S dataset [48], is not available to the public at the time of writing. These datasets were just assembled in recent years to represent a more grounded application for face identification on surveillance scenarios.

Due to the limited availability for real-world datasets, most of the evaluations at VLR were performed on datasets such as LFW [41], CASIA Webface [111], YouTube Faces [105], and Celeb-A [56]. However, when comparing the performance of these same models on other very challenging datasets, such as SCface [29], these models perform very poorly in terms of accuracy. The same effect happens when testing models trained on datasets with data from controlled environments, such as CMU-PIE [31] and FERET [76].

To improve the VLR robustness of the methods using deep face recognition, several augmentation strategies were proposed in recent years. For example, the authors in [28] rely on bilinear interpolation to synthesize the dataset for very low resolution. However, this method by itself is not effective at unconstrained settings. Lu *et al.* [62], on the other hand, resized the images into three different resolutions, to allow the network to learn more mappings due to the varying low resolution present in real-world applications. The study performed by Martínez-Díaz *et al.* [67], concluded that in order to use synthetic datasets, it is necessary to employ various downsampling strategies, to effectively simulate different real-world settings.

Due to the domain gap between synthetic and native datasets, it is harder for super resolution methods to generalize and become effective in true unconstrained scenarios. Super resolution methods are supervised-learning methods by nature, where usually a ground truth higher resolution image is needed in order to learn a mapping for synthetic low resolution to high resolution, in an effort to match the target HR domain. Attempting to remedy this domain gap, the authors of [14] have opted to use synthetic datasets combined with native images. They map the super resolution relationship using the synthetic dataset and hallucinating an HR counterpart of the native VLR face images. The achieve this by sharing weights between these two super resolution sub-networks and keeping the classification sub-networks with their own weights exclusively.

2) LACK OF DISCRIMINATIVE INFORMATION AND POSE VARIATIONS

Another challenge in this context is the lack of discriminative information at very low resolutions. Due to the lighting conditions, pose and blurriness, it becomes harder to extract useful features for classification purposes. Current low resolution face recognition methods do not use face alignment for optimizing performance. Usually, face alignment is a dataset pre-processing step for training any method [32] or the recognition method relies on its ability to generalize identities from its feature extraction capabilities. This leaves an area of opportunity for methods to include face normalization modules to further improve recognition performance. As evidenced in [26], synthesizing pose and expression variations can lead to an improvement in poses from ± 60 degrees and beyond. Convolutional Neural Networks can be used to effectively generate feature maps estimating pose and expression to further aid in face hallucination or classification. However, the act of encoding and decoding each image at the synthesis step can negatively affect inference runtime performance, as in the case of Super Resolution approaches.

In order to extract more robust features, most VLR face recognition methods usually tailor the loss function and optimization steps of the network to the VLR problem, along with designing multiple branch architectures. In the homogeneous VLR face recognition scenario, previous methods used feature extraction techniques based on SIFT [57], LBP [24], or Gabor [17]. Most of the methods are based on residual network architectures or in architectures that already work for general-purpose recognition problems, such as AttentionNets. A study conducted by Sandler *et al.* [82] explains the importance of the internal network layer resolution and the need for designing networks for the specific resolution of the problem at hand. In their paper, they introduce the concept of isometric networks, which retain their size throughout the network. This design yields a better performance regardless of the input resolution. However, they note that at extreme resolutions (14×14 and below), input resolution does matter as it affects the rest of the internal network layer sizes.

3) DOMAIN GAP BETWEEN HIGH RESOLUTION AND LOW RESOLUTION: MATCHING FEATURES FROM DIFFERENT SPACES

In the heterogeneous variant of the problem, for datasets such as Point-and-Shoot [4], UCCS [84], and SCface [29], a domain gap exists between images taken in HR controlled environments and the VLR images taken from the surveillance cameras. The images from surveillance cameras can present varying lighting conditions, blurriness, heavy pose variations, and varying resolutions. Extracting features from varying resolutions in particular becomes a problem. A single image from the native space has different dimensions from the rest of the images, including those from the gallery reference images. Firstly, the input image must be standardized to one or more predefined sizes. It becomes a decision for the image scaling process, where super resolution methods aim to scale the native images to the HR gallery image domain. Secondly, features extracted from the HR gallery images have richer and clearer information [63] than those from native VLR images. As such, the extracted features from a given algorithm greatly vary from one another. Coupled mappings and homogeneous feature extraction methods are the two strategies that have a more robust abstraction of the extracted features to match them under the same domain conditions. However, this domain gap does not exist in the homogeneous variant of the problem. In this variant, the recognition step is performed from images coming from the same domain. Datasets such as QMUL-Tinyface [14] and QMUL-Survface [16] are suitable for training these homogeneous approaches.

4) EFFICIENCY CHALLENGES

Generally, the inference time for classical Coupled Mappings methods is short, even using CPU only [116]. These methods perform very accurately on constrained scenarios. However, when testing them on unconstrained low resolution face recognition scenarios, their accuracy performance rapidly decays. Newer and more accurate deep learning methods for high resolution unconstrained face recognition [85], [93] are not suitable for run-time applications due to their increased computation requirements for inference. Furthermore, these methods do not perform as accurately in very low resolution unconstrained face recognition scenarios, as other state-of-the-art alternatives do [19], [62], [67]. To achieve state-of-the-art accuracy deep learning methods such as [62] and [14], use solutions based on multi-architecture networks, which still have heavy computation requirements. Due to the general efficiency requirements for deep learning applications, mainly on mobile and embedded systems, areas such as lightweight face recognition and quantization have emerged as well [45], [79]. These methods explore deep learning techniques using alternative data representations, such as binary and k-bit floating-point, with small accuracy performance penalty. However, no solutions aiming to bridge the domain gap present on VLR face recognition or using lightweight network design principles have been proposed. The only efficient approach specific for VLR face recognition was proposed by Ge *et al.* [28], which uses Knowledge Distillation.

Furthermore, the vast majority of the methods utilized for VLR face recognition report results based on accuracy and very few do on efficiency. The authors that do report run-time performance, sometimes omit the employed hardware configuration. Omitting this information makes efficiency assessments challenging to perform and compare against other methods.

D. DISCUSSION ON FUTURE RESEARCH DIRECTIONS

As per the reported results and previous discussion regarding face recognition performance in identification, verification, and efficiency metrics, we discuss future research directions in this section. We cover areas of opportunity of current state-of-the-art approaches and detail how they can significantly contribute to the very low resolution face recognition area.

1) CAPSULE NETWORK OPPORTUNITY AREAS ON EFFICIENCY AND FACE RECOGNITION ACCURACY PERFORMANCE

For the particular context of VLR FR, the Dual Directed Capsule Network approach has shown potential by outperforming other methods in the VLR FR Super Resolution literature such as the Robust Partially Coupled Network and CentreFace for VLR in terms of accuracy performance [90]. However, as shown in the rest of the results for CentreFace and the SurvFace dataset, the TAR@FAR=1% verification results still present a challenge. Furthermore, efficient Capsule Network architectures have not been proposed at the time of writing. This completely disregards the novel generalization ideas behind this approach type. The Capsule Network [38] concept proposed by Hinton *et al.* and later refined by Sabour *et al.* [80] is a niche research technique with much potential. Capsule networks have the potential for extracting focused information from different components.

These extractors, called capsules, later concatenate their output with the other capsules, effectively building a more robust descriptor. Some of the guidelines for Capsule Networks, or CapNets, include rejecting the idea of max pooling due to the loss of information and instead model activity vectors, which for any given class yields a vector of values. This vector is later used as a representation for posterior classification. However, they represent a challenge for mobile and embedded systems due to the cost of performing classification with high-dimensional feature vectors. Efficient CapsNet research is not present in the state of the art and can benefit from adding robustness and more accurate performance for face recognition in embedded systems. Early research on efficient Capsule Networks [122] has suggested that the major bottleneck is the dynamic routing procedure (which determines which capsule is a vector going to connect). This procedure is very intensive on memory calls. The authors have proposed a novel low-level framework and CapsNet architecture design to remedy this problem; however, further research is needed in this area.

2) LEVERAGING CNN MULTI-BRANCH ARCHITECTURES AND LOSS FUNCTIONS TAILORED FOR VLFR

The domain gap relationship and evaluation can be modeled at the CNN architecture level and the loss function level. Modeling these relationships has shown a more effective and more generalized feature extraction. Methods such as FAN [112], CSRI [14], among others, have highlighted the importance and effectiveness that inputs for VLR and HR imaging, synthetic or native and even sharing parameters for extracting features across different domain imagery. This principle needs to be extended to the very robust and efficient feature extraction capabilities of the more recent lightweight convolutional neural networks for face recognition. As of now, we can only utilize lightweight CNNs by up-scaling the very low resolution, effectively creating a synthetic image, not representative of the real domain. We identify this as a critical area of opportunity for these already robust feature extractors. We are confident that introducing and processing native VLR and HR images separately and using the efficient design principles of lightweight CNNs can aid generalization considerably. The most common loss functions used for the rest of the deep learning methods are Centerloss (Centreface) [104], Cross-Entropy Softmax [10], and Arcface [21]. These loss functions encourage the inter-class discriminative ability and a homogenized feature extraction process. However, they do not model or consider the native domain space of the heterogeneous source images in the same way some Coupled Mappings and Super Resolution methods do. The CNN methods using tailored loss functions for VLR and HR imagery are Deep Coupled ResNet [62], TCN [115], CSRI [14], and Dual Directed Capsule Networks [90]. For instance, the coupled mapping loss of DCR consists of a combination of softmax loss and center loss for HR and LR feature sets independently and a Euclidean loss for all the extracted feature vectors to the center of each domain. This kind of loss functions effectively enforce the network to learn feature representations based on the data from the same domain only. This principle makes the feature extraction process robust to cross-domain discrepancies.

3) UNTAPPED POTENTIAL OF KNOWLEDGE DISTILLATION APPROACHES

Knowledge Distillation approaches can provide knowledge from a more complex neural network architecture to a more simple one, used only at the testing phase. We have already discussed that methods, such as DCR, FAN, and Dual Directed CapsNet, are not suitable for real-time applications. However, they provide generalization capabilities not present in lightweight convolutional neural networks. We consider that Knowledge Distillation approaches [25], [92] can provide competent generalizations using the heavier networks. This methodology allows to leverage the robust and efficient feature extraction capabilities of the most successful lightweight convolutional neural networks. The heavier networks act as teacher networks to transfer the domain knowledge needed for lighter networks to extract features more robust to resolution and unconstrained condition changes.

IV. CONCLUSION

In this paper, we have reviewed the most successful approaches for unconstrained very low resolution face recognition, while discussing the limitations and advantages of each approach type in the state-of-the-art. We discussed the factors affecting accuracy and inference time performance and the caveats of using different training methodologies. We analyzed the impact of bridging the domain gap at the architecture level, loss function design, and image representation level. With this in mind, we have also discussed the most important tendencies in the deep learning convolutional neural networks as a whole, with approaches such as Capsule Networks, CNN Multi-branch architectures, and Knowledge Distillation.

REFERENCES

[1] L. An and B. Bhanu, "Face image super-resolution using 2D CCA," *Signal Process.*, vol. 103, pp. 184–194, Oct. 2014.

[2] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (SURF)," *Comput. Vis. Image Understand.*, vol. 110, no. 3, pp. 346–359, Jun. 2008.

[3] P. N. Belhumeur, J. P. Hespanha, and D. J. 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.

[4] J. R. Beveridge, K. W. Bowyer, P. J. Flynn, S. Cheng, P. J. Phillips, D. S. Bolme, B. A. Draper, G. H. Givens, Y. M. Lui, M. N. Teli, H. Zhang, and W. T. Scruggs, "The challenge of face recognition from digital point-and-shoot cameras," in *Proc. IEEE 6th Int. Conf. Biometrics, Theory, Appl. Syst. (BTAS)*, Sep. 2013, pp. 1–8.

[5] S. Biswas, G. Aggarwal, P. J. Flynn, and K. W. Bowyer, "Pose-robust recognition of low-resolution face images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 3037–3049, Dec. 2013.

[6] S. Biswas, K. W. Bowyer, and P. J. Flynn, "Multidimensional scaling for matching low-resolution face images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 10, pp. 2019–2030, Oct. 2012.

[7] B. J. Boom, G. M. Beumer, L. J. Spreeuwers, and R. N. J. Veldhuis, "The effect of image resolution on the performance of a face recognition system," in *Proc. 9th Int. Conf. Control, Autom., Robot. Vis.*, 2006, pp. 1–6.
[8] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks)," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 1021–1030.
[9] J. B. Camiña, M. A. Medina-Pérez, R. Monroy, O. Loyola-González, L. A. P. Villanueva, and L. C. G. Gurrola, "Bagging-RandomMiner: A one-class classifier for file access-based masquerade detection," *Mach. Vis. Appl.*, vol. 30, no. 5, pp. 959–974, Jul. 2019.
[10] J. Cao, Z. Su, L. Yu, D. Chang, X. Li, and Z. Ma, "Softmax cross entropy loss with unbiased decision boundary for image classification," in *Proc. Chin. Autom. Congr. (CAC)*, Nov. 2018, pp. 2028–2032.
[11] Z. Cao, Q. Yin, X. Tang, and J. Sun, "Face recognition with learning-based descriptor," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 2707–2714.
[12] C. Zhou, Z. Zhang, D. Yi, Z. Lei, and S. Z. Li, "Low-resolution face recognition via simultaneous discriminant analysis," in *Proc. Int. Joint Conf. Biometrics (IJCB)*, Oct. 2011, pp. 1–6.
[13] S. Chen, Y. Liu, X. Gao, and Z. Han, "MobileFaceNets: Efficient CNNs for accurate real-time face verification on mobile devices," in *Proc. Chin. Conf. Biometric Recognit.*, in Lecture Notes in Computer Science, 2018, pp. 428–438.
[14] Z. Cheng, X. Zhu, and S. Gong, "Low-resolution face recognition," pp. 1–16, Nov. 2018, *arXiv:1811.08965*. [Online]. Available: http://arxiv.org/abs/1811.08965
[15] Z. Cheng, X. Zhu, and S. Gong. (2018). *QMUL-SurvFace: Surveillance Face Recognition Challenge*. [Online]. Available: https://qmul-survface.github.io/
[16] Z. Cheng, X. Zhu, and S. Gong, "Surveillance face recognition challenge," 2018, *arXiv:1804.09691*. [Online]. Available: http://arxiv.org/abs/1804.09691
[17] C. Liu and H. Wechsler, "Gabor feature based classification using the enhanced Fisher linear discriminant model for face recognition," *IEEE Trans. Image Process.*, vol. 11, no. 4, pp. 467–476, Apr. 2002.
[18] Y. Chu, T. Ahmad, G. Bebis, and L. Zhao, "Low-resolution face recognition with single sample per person," *Signal Process.*, vol. 141, pp. 144–157, Dec. 2017.
[19] D. Zeng, H. Chen, and Q. Zhao, "Towards resolution invariant face recognition in uncontrolled scenarios," in *Proc. Int. Conf. Biometrics (ICB)*, Jun. 2016, pp. 1–8.
[20] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "RetinaFace: Single-shot multi-level face localisation in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5203–5212.
[21] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4690–4699.
[22] J. Deng, J. Guo, D. Zhang, Y. Deng, X. Lu, and S. Shi, "Lightweight face recognition challenge," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 1–9.
[23] C. Dong, C. C. Loy, and X. Tang, "Accelerating the super-resolution convolutional neural network," in *Computer Vision—ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham, Switzerland: Springer, 2016, pp. 391–407.
[24] D.-c. He and L. Wang, "Texture unit, texture spectrum, and texture analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 28, no. 4, pp. 509–512, Jul. 1990.
[25] C. N. Duong, K. Luu, K. G. Quach, and N. Le, "ShrinkTeaNet: Million-scale lightweight face recognition via shrinking teacher-student networks," 2019, *arXiv:1905.10620*. [Online]. Available: https://arxiv.org/abs/1905.10620
[26] C. Fu, Y. Hu, X. Wu, G. Wang, Q. Zhang, and R. He, "High-fidelity face manipulation with extreme poses and expressions," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 2218–2231, 2021.
[27] W. Gao, B. Cao, S. Shan, X. Chen, D. Zhou, X. Zhang, and D. Zhao, "The CAS-PEAL large-scale Chinese face database and baseline evaluations," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 38, no. 1, pp. 149–161, Jan. 2008.
[28] S. Ge, S. Zhao, C. Li, and J. Li, "Low-resolution face recognition in the wild via selective knowledge distillation," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 2051–2062, Apr. 2019.
[29] M. Grgic, K. Delac, and S. Grgic, "SCface—Surveillance cameras face database," *Multimedia Tools Appl.*, vol. 51, no. 3, pp. 863–879, Feb. 2011.
[30] K. Grm, W. J. Scheirer, and V. Struc, "Face hallucination using cascaded super-resolution and identity priors," *IEEE Trans. Image Process.*, vol. 29, pp. 2150–2165, 2020.
[31] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-PIE," *Image Vis. Comput.*, vol. 28, no. 5, pp. 807–813, May 2010.
[32] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "MS-Celeb-1M: A dataset and benchmark for large-scale face recognition," in *Proc. ECCV*, 2016, pp. 87–102.
[33] M. Haghighat and M. Abdel-Mottaleb, "Low resolution face recognition in surveillance systems using discriminant correlation analysis," in *Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2017, pp. 912–917.
[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[35] R. He, J. Cao, L. Song, Z. Sun, and T. Tan, "Adversarial cross-spectral face completion for NIR-VIS face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 5, pp. 1025–1037, May 2020.
[36] P. H. Hennings-Yeomans, S. Baker, and B. V. K. V. Kumar, "Simultaneous super-resolution and feature extraction for recognition of low-resolution faces," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.
[37] C. Herrmann, D. Willersinn, and J. Beyerer, "Low-resolution convolutional neural networks for video face recognition," in *Proc. 13th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Aug. 2016, pp. 221–227.
[38] G. E. Hinton, A. Krizhevsky, and S. D. Wang, "Transforming auto-encoders," in *Proc. 21th Int. Conf. Artif. Neural Netw. (ICANN)*. Berlin, Germany: Springer-Verlag, 2011, pp. 44–51.
[39] A. Howard, M. Sandler, B. Chen, W. Wang, L.-C. Chen, M. Tan, G. Chu, V. Vasudevan, Y. Zhu, R. Pang, H. Adam, and Q. Le, "Searching for MobileNetV3," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1314–1324.
[40] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*. [Online]. Available: http://arxiv.org/abs/1704.04861
[41] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Dept. Comput. Sci., Univ. Massachusetts, Amherst, MA, USA, Tech. Rep. 07-49, Oct. 2007.
[42] H. Huang, H. He, X. Fan, and J. Zhang, "Super-resolution of human face image using canonical correlation analysis," *Pattern Recognit.*, vol. 43, no. 7, pp. 2532–2543, Jul. 2010.
[43] P. J. Huber, "Robust estimation of a location parameter," *Ann. Math. Statist.*, vol. 35, no. 1, pp. 73–101, Mar. 1964.
[44] F. N. Iandola, M. W. Moskewicz, K. Ashraf, S. Han, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50× fewer parameters and <1 mb model size," 2017, *arXiv:1602.07360*. [Online]. Available: https://arxiv.org/abs/1602.07360
[45] B. Jacob, S. Kligys, B. Chen, M. Zhu, M. Tang, A. Howard, H. Adam, and D. Kalenichenko, "Quantization and training of neural networks for efficient integer-arithmetic-only inference," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2704–2713.
[46] J. Yang, D. Zhang, A. F. Frangi, and J.-y. Yang, "Two-dimensional PCA: A new approach to appearance-based face representation and recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 1, pp. 131–137, Jan. 2004.
[47] J. Yang, J. Wright, T. Huang, and Y. Ma, "Image super-resolution as sparse representation of raw image patches," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.
[48] N. D. Kalka, B. Maze, J. A. Duncan, K. OrConnor, S. Elliott, K. Hebert, J. Bryan, and A. K. Jain, "IJB–S: IARPA janus surveillance video benchmark," in *Proc. IEEE 9th Int. Conf. Biometrics Theory, Appl. Syst. (BTAS)*, Oct. 2018, pp. 1–9.
[49] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Fast and accurate image super-resolution with deep Laplacian pyramid networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 11, pp. 2599–2613, Nov. 2019.

[50] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 105–114.

[51] S. H. Lee and S. Choi, "Two-dimensional canonical correlation analysis," *IEEE Signal Process. Lett.*, vol. 14, no. 10, pp. 735–738, Oct. 2007.

[52] B. Li, H. Chang, S. Shan, and X. Chen, "Low-resolution face recognition via coupled locality preserving mappings," *IEEE Signal Process. Lett.*, vol. 17, no. 1, pp. 20–23, Jan. 2010.

[53] P. Li, L. Prieto, D. Mery, and P. Flynn, "Face recognition in low quality images: A survey," May 2018, *arXiv:1805.11519*. [Online]. Available: http://arxiv.org/abs/1805.11519

[54] P. Li, L. Prieto, D. Mery, and P. J. Flynn, "On low-resolution face recognition in the wild: Comparisons and new techniques," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 8, pp. 2000–2012, Aug. 2019.

[55] S. Z. Li and A. K. Jain, *Handbook of Face Recognition*, 2nd ed. New York, NY, USA: Springer, 2011.

[56] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3730–3738.

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

[58] O. Loyola-González, "Black-box vs. white-box: Understanding their advantages and weaknesses from a practical point of view," *IEEE Access*, vol. 7, pp. 154096–154113, 2019.

[59] O. Loyola-González, A. López-Cuevas, M. A. Medina-Pérez, B. Camiña, J. E. Ramírez-Márquez, and R. Monroy, "Fusing pattern discovery and visual analytics approaches in tweet propagation," *Inf. Fusion*, vol. 46, pp. 91–101, Mar. 2019.

[60] O. Loyola-González, M. Medina-Pérez, D. Hernández-Tamayo, R. Monroy, J. Carrasco-Ochoa, and M. García-Borroto, "A pattern-based approach for detecting pneumatic failures on temporary immersion bioreactors," *Sensors*, vol. 19, no. 2, p. 414, Jan. 2019.

[61] O. Loyola-González, R. Monroy, M. A. Medina-Pérez, B. Cervantes, and J. E. Grimaldo-Tijerina, "An approach based on contrast patterns for bot detection on Web log files," in *Advances in Soft Computing*, I. Batyrshin, M. Martínez-Villaseñor, and H. P. Espinosa, Eds. Cham, Switzerland: Springer, 2018, pp. 276–285.

[62] Z. Lu, X. Jiang, and A. Kot, "Deep coupled ResNet for low-resolution face recognition," *IEEE Signal Process. Lett.*, vol. 25, no. 4, pp. 526–530, Apr. 2018.

[63] Y. M. Lui, D. Bolme, B. A. Draper, J. R. Beveridge, G. Givens, and P. J. Phillips, "A meta-analysis of face recognition covariates," in *Proc. IEEE 3rd Int. Conf. Biometrics, Theory, Appl., Syst.*, Sep. 2009, pp. 1–8.

[64] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "Shufflenet V2: Practical guidelines for efficient CNN architecture design," in *Computer Vision—ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham, Switzerland: Springer, 2018, pp. 122–138.

[65] B. Martinez, J. Yang, A. Bulat, and G. Tzimiropoulos, "Training binary neural networks with real-to-binary convolutions," 2020, *arXiv:2003.11535*. [Online]. Available: https://arxiv.org/abs/2003.11535

[66] Y. Martindez-Diaz, L. S. Luevano, H. Mendez-Vazquez, M. Nicolas-Diaz, L. Chang, and M. Gonzalez-Mendoza, "ShuffleFaceNet: A lightweight face architecture for efficient and highly-accurate face recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 1–8.

[67] Y. Martínez-Díaz, H. Méndez-Vázquez, L. S. Luevano, L. Chang, and M. González-Mendoza, "Lightweight low-resolution face recognition for surveillance applications," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 5421–5428.

[68] H. Miarnaeimi and P. Davari, "A new fast and efficient HMM-based face recognition system using a 7-state HMM along with SVD coefficients," *Iranian J. Electr. Electron. Eng.*, vol. 4, nos. 1–2, pp. 46–57, 2008.

[69] S. P. Mudunuri, S. Sanyal, and S. Biswas, "GenLR-Net: Deep framework for very low resolution face and object recognition with generalization to unseen categories," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, p. 60209.

[70] S. P. Mudunuri, S. Venkataramanan, and S. Biswas, "Dictionary alignment with re-ranking for low-resolution NIR-VIS face recognition," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 4, pp. 886–896, Apr. 2019.

[71] M. Asim, Z. Ming, and M. Y. Javed, "CNN based spatio-temporal feature extraction for face anti-spoofing," in *Proc. 2nd Int. Conf. Image, Vis. Comput. (ICIVC)*, Jun. 2017, pp. 234–238.

[72] K. Nguyen, C. Fookes, S. Sridharan, M. Tistarelli, and M. Nixon, "Super-resolution for biometrics: A comprehensive survey," *Pattern Recognit.*, vol. 78, pp. 23–42, Jun. 2018.

[73] S. Ouyang, T. Hospedales, Y.-Z. Song, X. Li, C. C. Loy, and X. Wang, "A survey on heterogeneous face recognition: Sketch, infra-red, 3D and low-resolution," *Image Vis. Comput.*, vol. 56, pp. 28–48, Dec. 2016.

[74] 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. Swansea, U.K.: BMVA Press, Sep. 2015, pp. 41.1–41.12.

[75] C. Peng, N. Wang, J. Li, and X. Gao, "Re-ranking high-dimensional deep local representation for NIR-VIS face recognition," *IEEE Trans. Image Process.*, vol. 28, no. 9, pp. 4553–4565, Sep. 2019.

[76] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The FERET evaluation methodology for face-recognition algorithms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 10, pp. 1090–1104, Oct. 2000.

[77] A. Rai, V. Chudasama, K. Upla, K. Raja, R. Ramachandra, and C. Busch, "ComSupResNet: A compact super-resolution network for low-resolution face images," in *Proc. 8th Int. Workshop Biometrics Forensics (IWBF)*, Apr. 2020, pp. 1–6.

[78] R. Ranjan, C. D. Castillo, and R. Chellappa, "L_2-constrained softmax loss for discriminative face verification," *CoRR*, vol. abs/1703.09507, pp. 1–10, Mar. 2017.

[79] M. Rastegari, V. Ordonez, J. Redmon, and A. Farhadi, "XNOR-Net: ImageNet classification using binary convolutional neural networks," in *Proc. Eur. Conf. Comput. Vis.*, in Lecture Notes in Computer Science, 2016, pp. 525–542.

[80] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst. (NIPS)*. Red Hook, NY, USA: Curran Associates, 2017, pp. 3859–3869.

[81] M. Sajjad, S. Khan, T. Hussain, K. Muhammad, A. K. Sangaiah, A. Castiglione, C. Esposito, and S. W. Baik, "CNN-based anti-spoofing two-tier multi-factor authentication system," *Pattern Recognit. Lett.*, vol. 126, pp. 123–131, Sep. 2019.

[82] M. Sandler, J. Baccash, A. Zhmoginov, and A. Howard, "Non-discriminative data or weak model? On the relative importance of data and model resolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 1036–1044.

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

[84] A. Sapkota and T. E. Boult, "Large scale unconstrained open set face database," in *Proc. IEEE 6th Int. Conf. Biometrics, Theory, Appl. Syst. (BTAS)*, Sep. 2013, pp. 1–8.

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

[86] J. Shi and C. Qi, "From local geometry to global structure: Learning latent subspace for low-resolution face image recognition," *IEEE Signal Process. Lett.*, vol. 22, no. 5, pp. 554–558, May 2015.

[87] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1874–1883.

[88] S. Siena, V. N. Boddeti, and B. V. K. V. Kumar, "Coupled marginal Fisher analysis for low-resolution face recognition," in *Computer Vision—ECCV 2012. Workshops and Demonstrations*, A. Fusiello, V. Murino, and R. Cucchiara, Eds. Berlin, Germany: Springer, 2012, pp. 240–249.

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

[90] M. Singh, S. Nagpal, R. Singh, and M. Vatsa, "Dual directed capsule network for very low resolution image recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 340–349.

[91] M. Singh, S. Nagpal, M. Vatsa, R. Singh, and A. Majumdar, "Identity aware synthesis for cross resolution face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 479–488.

[92] D. Svitov and S. Alyamkin, "MarginDistillation: Distillation for margin-based softmax," 2020, *arXiv:2003.02586*. [Online]. Available: https://arxiv.org/abs/2003.02586
[93] 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. (CVPR)*, Jun. 2014, pp. 1701–1708.
[94] L. Trottier, P. Giguere, and B. Chaib-Draa, "Parametric exponential linear unit for deep convolutional neural networks," in *Proc. 16th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2017, pp. 207–214.
[95] M. A. Turk and A. P. Pentland, "Face recognition using eigenfaces," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jan. 1991, pp. 586–587.
[96] T. Uiboupin, P. Rasti, G. Anbarjafari, and H. Demirel, "Facial image super resolution using sparse representation for improving face recognition in surveillance monitoring," in *Proc. 24th Signal Process. Commun. Appl. Conf. (SIU)*, May 2016, pp. 437–440.
[97] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Dec. 2001, p. I.
[98] F. Wang, J. Cheng, W. Liu, and H. Liu, "Additive margin softmax for face verification," *IEEE Signal Process. Lett.*, vol. 25, no. 7, pp. 926–930, Jul. 2018.
[99] X. Wang and X. Tang, "Face photo-sketch synthesis and recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 11, pp. 1955–1967, Nov. 2009.
[100] Z. Wang, S. Chang, Y. Yang, D. Liu, and T. S. Huang, "Studying very low resolution recognition using deep networks," pp. 4792–4800, 2016, *arXiv:1601.04153*. [Online]. Available: http://arxiv.org/abs/1601.04153
[101] Z. Wang, D. Liu, J. Yang, W. Han, and T. Huang, "Deep networks for image super-resolution with sparse prior," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 370–378.
[102] Z. Wang, W. Yang, and X. Ben, "Low-resolution degradation face recognition over long distance based on CCA," *Neural Comput. Appl.*, vol. 26, no. 7, pp. 1645–1652, Oct. 2015.
[103] A. R. Webb, "Multidimensional scaling by iterative majorization using radial basis functions," *Pattern Recognit.*, vol. 28, no. 5, pp. 753–759, May 1995.
[104] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Computer Vision—ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham, Switzerland: Springer, 2016, pp. 499–515.
[105] L. Wolf, T. Hassner, and I. Maoz, "Face recognition in unconstrained videos with matched background similarity," in *Proc. CVPR*, Jun. 2011, pp. 529–534.
[106] X. Wu, R. He, Z. Sun, and T. Tan, "A light CNN for deep face representation with noisy labels," 2015, *arXiv:1511.02683*. [Online]. Available: https://arxiv.org/abs/1511.02683
[107] M. Yan, M. Zhao, Z. Xu, Q. Zhang, G. Wang, and Z. Su, "VarGFaceNet: An efficient variable group convolutional neural network for lightweight face recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1–8.
[108] F. Yang, W. Yang, R. Gao, and Q. Liao, "Discriminative multidimensional scaling for low-resolution face recognition," *IEEE Signal Process. Lett.*, vol. 25, no. 3, pp. 388–392, Mar. 2018.
[109] H. Yang, M. Fritzsche, C. Bartz, and C. Meinel, "BMXNet: An open-source binary neural network implementation based on MXNet," in *Proc. 25th ACM Int. Conf. Multimedia (MM)*, New York, NY, USA, 2017, pp. 1209–1212.
[110] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Trans. Image Process.*, vol. 19, no. 11, pp. 2861–2873, Nov. 2010.
[111] 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
[112] X. Yin, Y. Tai, Y. Huang, and X. Liu, "FAN: Feature adaptation network for surveillance face recognition and normalization," 2019, *arXiv:1911.11680*. [Online]. Available: https://arxiv.org/abs/1911.11680
[113] X. Yu, B. Fernando, R. Hartley, and F. Porikli, "Semantic face hallucination: Super-resolving very low-resolution face images with supplementary attributes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 11, pp. 2926–2943, Nov. 2020.
[114] R. Zeyde, M. Elad, and M. Protter, "On single image scale-up using sparse-representations," in *Curves and Surfaces*, J.-D. Boissonnat, P. Chenin, A. Cohen, C. Gout, T. Lyche, M.-L. Mazure, and L. Schumaker, Eds. Berlin, Germany: Springer, 2012, pp. 711–730.
[115] J. Zha and H. Chao, "TCN: Transferable coupled network for cross-resolution face recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 3302–3306.
[116] J. Zhang, Z. Guo, X. Li, and Y. Chen, "Large margin coupled mapping for low resolution face recognition," in *PRICAI 2016: Trends in Artificial Intelligence* (Lecture Notes in Computer Science), vol. 3157, C. Zhang, H. W. Guesgen, and W.-K. Yeap, Eds. Berlin, Germany: Springer, 2016, pp. 661–672.
[117] K. Zhang *et al.*, "AIM 2019 challenge on constrained super-resolution: Methods and results," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 3565–3574.
[118] K. Zhang, Z. Zhang, C.-W. Cheng, W. H. Hsu, Y. Qiao, W. Liu, and T. Zhang, "Super-identity convolutional neural network for face hallucination," in *Proc. Eur. Conf. Comput. Vis.*, Lecture Notes in Computer Science, 2018, pp. 196–211.
[119] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.
[120] P. Zhang, X. Ben, W. Jiang, R. Yan, and Y. Zhang, "Coupled marginal discriminant mappings for low-resolution face recognition," *Optik*, vol. 126, no. 23, pp. 4352–4357, Dec. 2015.
[121] Q. Zhang, J. Li, M. Yao, L. Song, H. Zhou, Z. Li, W. Meng, X. Zhang, and G. Wang, "VarGNet: Variable group convolutional neural network for efficient embedded computing," 2019, *arXiv:1907.05653*. [Online]. Available: http://arxiv.org/abs/1907.05653
[122] X. Zhang, S. L. Song, C. Xie, J. Wang, W. Zhang, and X. Fu, "Enabling highly efficient capsule networks processing through a PIM-based architecture design," in *Proc. IEEE Int. Symp. High Perform. Comput. Archit. (HPCA)*, Feb. 2020, pp. 542–555.
[123] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6848–6856.
[124] P. Zhao, F. Zhang, J. Wei, Y. Zhou, and X. Wei, "SADG: Self-aligned dual NIR-VIS generation for heterogeneous face recognition," *Appl. Sci.*, vol. 11, no. 3, p. 987, Jan. 2021.
[125] S. Zhu, S. Liu, C. C. Loy, and X. Tang, "Deep cascaded bi-network for face hallucination," in *Computer Vision—ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham, Switzerland: Springer, 2016, pp. 614–630.

LUIS S. LUEVANO (Member, IEEE) received the B.Sc. degree in computer science and technology from the Tecnologico de Monterrey, Mexico, and the M.Sc. degree from the Stevens Institute of Technology, in 2018, with a focus on artificial intelligence and computer vision. He is currently pursuing the Ph.D. degree in computer science with the Tecnologico de Monterrey. He has one year and a half of industry experience with Dell Inc., in 2016. He is also a part of the Intelligent Systems Research Group, Tecnologico de Monterrey, where he focused on very low resolution face recognition research as his doctoral thesis topic, striving to close the gap for real-time applications.

LEONARDO CHANG received the bachelor's degree (Hons.) from CUJAE University, Havana, Cuba, in 2007, and the M.Sc. and Ph.D. degrees in computer science from the National Institute of Astrophysics, Optics, and Electronics (INAOE), Mexico, in 2010 and 2015, respectively. From 2007 to 2017, he was a Researcher at CENATAV, Cuba, where he led and contributed to applied researches for commercial face recognition systems, for both constrained and unconstrained environments. He is currently a full-time Researcher and a Professor with the Tecnologico de Monterrey, Mexico. He has published several articles in top journals and several papers in conferences. His research interests include biometrics (mainly focused to face recognition), object recognition, and video-surveillance applications.

HEYDI MÉNDEZ-VÁZQUEZ is graduated in software engineering from the Technological University of Havana José Antonio Echeverría, Cuba, in 2005. She received the Ph.D. degree in automation and computing in the field of face recognition in 2011. Since 2005, she has been a Researcher at CENATAV. She is currently the Head of the Department of Biometric Research. She has more than 50 published articles regarding the development of new methods for automatic face recognition. Her research interests include biometrics, face recognition, and digital image processing.

YOANNA MARTÍNEZ-DÍAZ is graduated in software engineering from the Technological University of Havana José Antonio Echeverría, Cuba, in 2010. She received the Ph.D. degree in automation and computing in the field of face recognition with the Technological University of Havana José Antonio Echeverría, in 2018. She is currently a Researcher with the Department of Biometric Research, Advanced Technologies Application Center (CENATAV). She has coauthored more than 20 articles in international journals and more than 20 papers in conferences. Her research interests include biometrics, image and video processing, and face recognition.

MIGUEL GONZÁLEZ-MENDOZA received the Ph.D. degree in artificial intelligence from the LAAS—INSA, Toulouse, France, in 2004. He was responsible of the Intelligent Systems Research Group, from 2007 to 2015, and the Head of the Graduate Programs on computer sciences at the Tecnologico de Monterrey, Mexico, from 2005 to 2016. He currently works as a Research Professor. He is interested in machine learning, data management, and computer vision applications, in which he has advised 31 Ph.D. thesis and 24 master's thesis. He leads several Mexican and European Commission research projects (as a Local Chair). He has authored more than 100 articles in JCR, congresses, and book chapters. He is the Former President of the Mexican Society for Artificial Intelligence for the period 2015–2018.