Wiley
IET Biometrics
Volume 2024, Article ID 7886911, 23 pages
https://doi.org/10.1049/2024/7886911

Review Article

A Survey on Automatic Face Recognition Using Side-View Face Images

Pinar Santemiz, Luuk J. Spreeuwers, and Raymond N. J. Veldhuis

Faculty of EEMCS, Data Management and Biometrics, University of Twente, P.O. Box 217 7500AE, Enschede, Netherlands

Correspondence should be addressed to Pinar Santemiz; p.santemiz-1@utwente.nl

Received 13 December 2023; Revised 20 April 2024; Accepted 31 May 2024

Academic Editor: Claus Vielhauer

Face recognition from side-view positions poses a considerable challenge in automatic face recognition tasks. Pose variation up to the side-view is an issue of difference in appearance and visibility since only one eye is visible at the side-view poses. Traditionally overlooked, recent advancements in deep learning have brought side-view poses to the forefront of research attention. This survey comprehensively investigates methods addressing pose variations up to side-view and categorizes research efforts into feature-based, image-based, and set-based pose handling. Unlike existing surveys addressing pose variations, our emphasis is specifically on extreme poses. We report numerous promising innovations in each category and contemplate the utilization and challenges associated with side-view. Furthermore, we introduce current datasets and benchmarks, conduct performance evaluations across diverse methods, and explore their unique constraints. Notably, while feature-based methods currently stand as the state-of-the-art, our observations suggest that cross-dataset evaluations, attempted by only a few researchers, produce worse results. Consequently, the challenge of matching arbitrary poses in uncontrolled settings persists.

1. Introduction

One of the few problems still challenging in face recognition is face recognition across extreme poses. Many applications successfully apply face recognition in various domains, such as surveillance, forensics, access control, commercial uses, intelligent human–computer interfaces, and entertainment [1].

The most prominent advantage of automatic face recognition is its nonintrusive nature, which enables the recognition of uncooperative subjects. This property is important for real-world applications where people are rarely recorded under precisely controlled conditions. These applications bring new challenges such as pose and illumination variations, low resolution, and occlusion [1].

When pose variation is large, image regions suffer from distortions, discontinuities, and self-occlusions. In side-view poses, these factors are so significant that the symmetry of the face can no longer be an assumption. Moreover, the variation in facial texture and shape is nonlinear, resulting in the loss of semantic correspondences.

This study specifically targets side-view poses and reviews both still image- and video-based face recognition studies from that perspective. We comprehensively analyze mainstream pose-invariant face recognition algorithms and compare as many state-of-the-art works as possible. Initially, individuals were enrolled with frontal poses and tested with face images captured from diverse viewpoints. Researchers addressed this challenge by handling pose alignment at the feature level or synthesizing faces in different views [2]. Over the years, several datasets have become standard for extreme pose studies, including CMU Multi-PIE [3], celebrities in frontal-profile (CFP) [4], IJB-A [5], YouTube celebrities (YTC) [6], and YouTube faces (YTF) [7]. We include studies tested on these datasets and compare their performances. The availability of large-scale data has led to the development of set-based methods that compare sets of face images in arbitrary poses. Moreover, advancements in deep learning methods have encouraged face recognition to extend performance evaluations to extreme poses.

Previously, several surveys have been published that specifically focus on pose variations [1, 8, 9]. Furthermore, some recent papers covering advances in deep learning algorithms and video-based face recognition [10, 11, 12] present a general idea of the progress in pose invariant face recognition.

But we have not seen a clear overview of the side-view. Extreme poses are mentioned as a part of these surveys, and the reported information is limited. In Table 1, we summarize survey papers addressing pose variation and discuss their limitations.

The pose invariant face recognition approaches can be categorized based on their pose handling methods, as shown in Figure 1, namely:

(1) *Feature-based* pose handling approaches that search for pose-robust description methods.
(2) *Image-based* pose handling approaches that synthesize faces in similar poses as a means of alignment.
(3) *Set-based* pose handling approaches that compare sets of images and combine these techniques.

In the literature, there is some inconsistency in the definition of "side-view." Some early methods limit side-view to poses up to 60° [19]. On the other hand, some define poses around 75° – 90° as "profile" [20, 21, 22, 23], whereas for some, the term "profile" refers to the curve representing the outline of the face [24, 25]. We connect the definition to visibility instead of the pose angle to avoid this confusion. We define side-view poses as images with significant pose variations where precisely one of the eyes is not visible due to self-occlusion. This approximately corresponds to 75° – 90°. We illustrate the appearance changes across poses in Figure 2.

We will review available studies and investigate the progress and challenges in different scenarios. At the end of each section, we offer a discussion to summarize the relevant research. Following the structure in Figure 1, we will examine feature-based pose handling methods in Section 2, image-based methods in Section 3, and set-based methods in Section 4. We will present available datasets and benchmarks, compare the performances of these methods in Section 5, and present our final conclusion in Section 6.

2. Feature-Based Pose Handling

Many pose-invariant face recognition methods focus on handling pose variation at the feature level. The aim is to extract pose-robust features or project the samples to a subspace where the correlation between the same subject under pose variation is maximum. Early methods within this category use manually designed features robust to pose variation up to some degree. Later studies applied machine learning techniques such as similarity learning and deep learning.

2.1. Engineered Features. Inspired by frontal face recognition methods, initial studies used engineered features to describe facial image semantic correspondences. In feature engineering, researchers employ domain knowledge to select and transform the most relevant variables within raw data. One advantage of this approach is that these methods do not require training. However, they cannot handle considerable pose variation, and their performance depends highly on alignment. Therefore, these algorithms commonly compare poses in similar views, assuming the closest match belongs to the same identity.

Some initial studies use the complementary information provided by the facial profile to align and compare faces [24, 25, 26, 27]. Most profile-based algorithms propose to align the faces using a tangent-based normalization method [24, 26] or the curvatures [25]. Once aligned, recognition is performed by comparing these curvatures. However, such information is not robust to pose variation. An informative facial profile can only be obtained on side-view and get poor results even under small view changes.

Later studies [28, 29] aim to extract features that use texture information in addition to shape. Since holistic approaches depend more on pixel-wise correspondences, these methods commonly divide the image into local regions to ensure robustness to small pose variations. Some of the most popular description methods are Gabor wavelet [30], SIFT [31], SURF [32], HOG [33], and LBP [34]. Sengupta et al. [4] compared these methods and showed that they perform poorly in large pose variation.

2.2. Similarity Learning. Similarity learning refers to the process of training models to learn a similarity function to determine if two objects belong to the same class. Ding et al. [23] proposed a multitask learning approach named multitask feature transformation learning (MtFTL) to break down recognition into pose-based feature transformations. They use a local descriptor called dual cross patterns (DCP) [35] to describe the visible face region and propose patch-based partial representation (PBPR) to form a set of patch-level DCP features. Using MtFTL, they learn transformation dictionaries to map patch features of different poses into a common discriminative space.

Similarly, distance metric learning aims to construct a function to map objects into an embedded space where similar objects have smaller distances with respect to each other than dissimilar objects. Lu et al. [36] proposed jointly optimizing the face and pose verification tasks using a pose-robust metric learning framework. Their model learns two metrics, one for pose verification and one for face verification, and uses the projection matrices to calculate the similarity scores.

2.3. Deep Learning. Deep learning is a machine learning approach where multiple layers of nodes transfer input data into higher level features to achieve distinctive representations. The availability of datasets that are rich in size and variety enabled deep learning methods to dominate the research due to their ability to handle pose and general high performance.

In 2014, DeepID [37] and DeepFace [38] dramatically improved the state-of-the-art performance on the LFW benchmark [18]. The following year Parkhi et al. [39] proposed VGG-face, which was trained on 2.6 million faces from the Internet Movie Database. The same year Schroff et al. [40] presented FaceNet that learns a deep convolutional network to map face images to a compact Euclidean space.

A deep learning architecture for pose invariant face recognition can be seen in Figure 3. In general, these methods

Table 1: Summary of surveys addressing pose variations.

Publication	Focus	Summary	Limitations
Zhao et al. [8]	Face recognition	This paper covers still- and video-based face recognition, elaborating on the key steps such as detection, alignment, feature extraction, and recognition. A dedicated section on pose variation reporting challenges provides an overview of the initial methods concerning pose invariant face recognition	The reviewed papers are predominantly evaluated using the FERET [13], FRVT [14], XM2VTS [15], and CMU-PIE [16] datasets. These datasets are outdated and have a limited number of subjects. The methods reviewed, which emphasize pose-invariant face recognition, do not encompass pose variations up to the side-view, and their effectiveness is primarily limited to small pose variations
Zhang and Gao [9]	Pose-invariant face recognition	This survey focuses on face recognition across different poses and categorizes the approaches into general algorithms, 2D techniques, and 3D methods. General algorithms use feature-based pose handling methods to handle various image variations. On the other hand, 2D and 3D methods specifically target pose variations and utilize image-based pose handling techniques	The performances of the methods are mainly tested on the FERET [13] and CMU-PIE [16] datasets. The reviewed methods do not handle extreme pose variations
Barr et al. [17]	Face recognition from video	Barr et al. [17] reviewed face recognition from video and focus on challenges arising from unconstrained settings, including pose variation. They categorize methods into two groups: set-based and sequence-based. Studies involving manifold modeling, 3D modeling, super-resolution, and frame selection are grouped under the set-based category, while methods utilizing temporal cues for recognition are placed under the sequence-based category	This paper offers a comprehensive overview of video-based methods but only mentions pose variation as a challenge without explicitly providing the performances of the reviewed methods in addressing pose variation. Additionally, they provide a list of video datasets, although most are outdated
Ding and Tao [1]	Pose-invariant face recognition	This paper concentrates on pose-invariant face recognition and categorizes existing methods into four groups: feature extraction approaches, multiview subspace learning approaches, face synthesis approaches, and hybrid approaches. The first two groups encompass feature-based pose handling methods. The face synthesis section includes image-based pose handling techniques, further categorized into 2D and 3D techniques	Ding and Tao provided a comparison of the methods based on their performances on the FERET [13], CMU Multi-PIE [3], and LFW [18] datasets. However, these tests do not encompass side-view poses, and the evaluation does not include deep learning methods. Additionally, the survey provides a list of standard datasets used for pose-invariant face recognition. Most datasets in the list are outdated except for the CMU Multi-PIE [3] and IJB-A [5]
Taskiran et al. [12]	Face recognition	This survey summarizes image- and video-based face recognition methods, provides a comprehensive taxonomy of face recognition systems, and reviews popular datasets. Furthermore, it covers deep learning-based methods and emphasizes their robustness regarding extreme poses	The focus of this paper is not on pose variation; therefore, the information regarding the performances of the methods under pose variation is very limited
Wang and Deng [11]	Deep face recognition	Wang and Deng review recent developments on deep face recognition methods and provide insight into available algorithm designs. They also summarize available face datasets and evaluation protocols regarding deep learning	The information specifically on pose variation is limited. The paper includes a section concerning technical challenges in deep face recognition where they underline the difficulty in cross-pose face recognition
Wang et al. [10]	Face recognition	In this paper, Wang et al. summarize face recognition algorithms and delve into the details of deep learning methods. They provide an extensive analysis of these methods, evaluation protocols, and the impact of data distribution	This paper categorizes methods based on their strategies and related applications. Pose is only mentioned when the reviewed paper focuses on pose handling. However, there is no discussion or comparison regarding pose variation

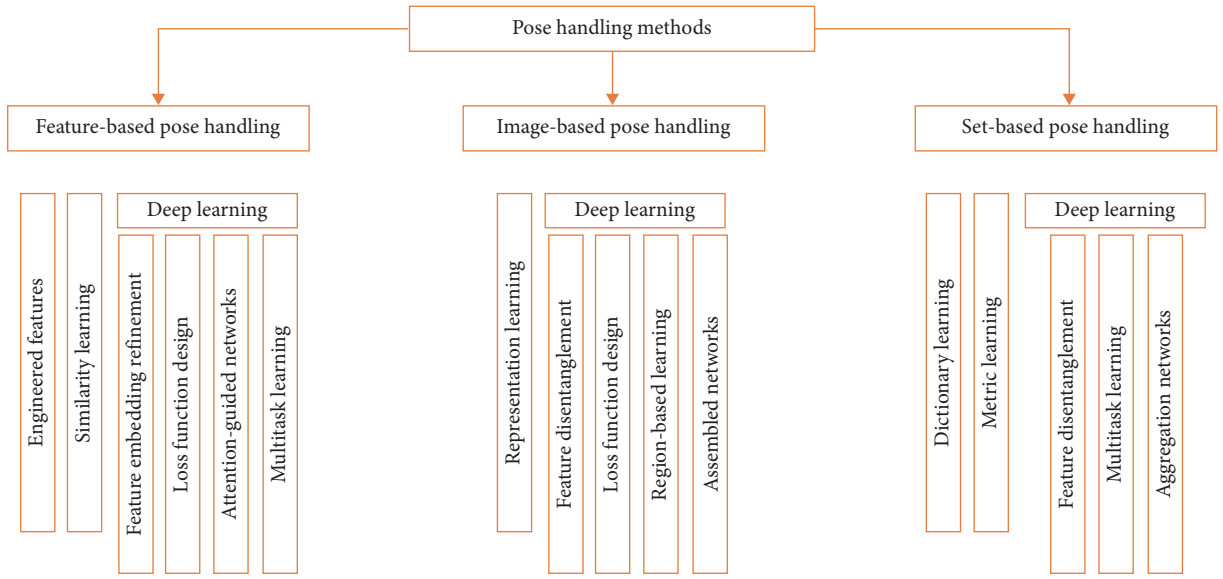

FIGURE 1: Pose invariant face recognition methods are categorized based on their pose handling approaches.

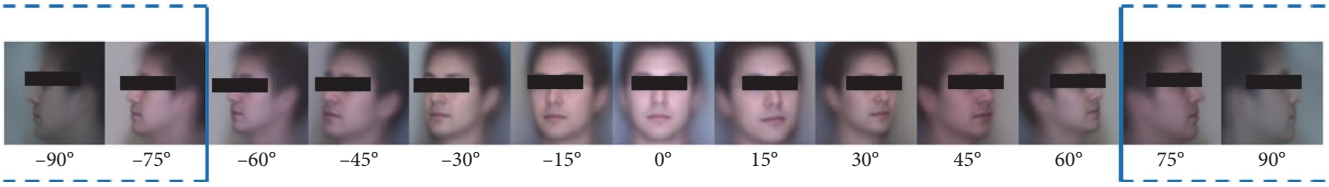

FIGURE 2: Appearance changes across pose variation from −90° to +90° using the average faces in the CMU Multi-PIE. In these examples, only at poses around 75°–90°, one of the eyes is occluded and regarded as side-view images.

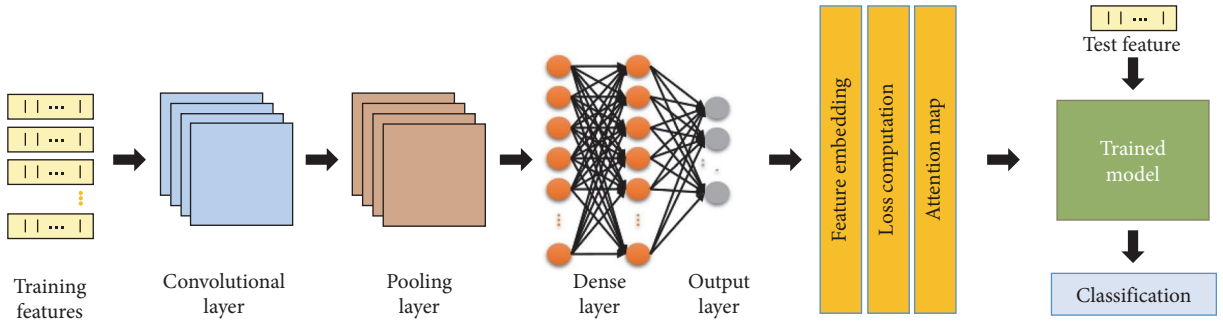

FIGURE 3: Deep learning architecture for pose invariant face recognition. Generally, networks that deal with face recognition under pose are convolutional neural networks (CNN). The general architecture can be improved by introducing several training principles to automatically enhance specific layers or operations. A common method is to add an extra step for feature embedding refinement. A variety of loss functions are available to train the algorithm, and a better loss function design can improve the performance. Also, an attention module can force the network to focus on specific data parts containing more information.

improve convolutional neural networks (CNN) to enhance specific layers or operations according to the requirements. Available techniques are categorized in Figure 1. One common method is to add a feature embedding refinement step to enhance the descriptive capacity of the model. Some methods focus on a better loss function design to improve performance. Also, an attention module can force the network to focus on specific data parts containing more information. Finally, multitask learning is introduced that assembles multiple networks and learns different tasks jointly.

2.3.1. Feature Embedding Refinement. To improve the descriptive characteristic of the deep features, some studies propose feature embedding refinement methods to find a

mapping that transforms deep features into a new space where the minimum distances between class centers are maximized. Unlike the conventional metric learning approach, the embedding step is treated as an additional layer to improve the network during training.

Zhang et al. [41] proposed a feature-level alignment method using a single hidden layer neural network to find mappings from any pose to random faces with predefined poses. They apply a sparsity constraint to reduce the dimensionality and concatenate generated vectors to obtain pose-invariant features. Sankaranarayanan et al. [42] proposed a DCNN method that couples a CNN-based approach with a low dimensional triplet probability embedding (TPE) step. Their embedding approach consists of TPE and triplet distance embedding (TDE) and achieves more discriminative projections of deep features. Chen et al. [43] combined Fisher vector representation and deep convolutional network (DCNN) features to represent faces for unconstrained face verification tasks. Their method first extracts deep features using a pretrained DCNN model and learns a Gaussian mixture model representing the features. Then, they perform Fisher Vector encoding and metric learning for verification.

Wang et al. [44] presented a method that combines deep features with a commercial off-the-shelf (COTS) matcher. In their method, they first reduce the number of candidates using a fast k-NN search and fuse the deep learning and COTS matcher scores for identification. Cao et al. [45] investigated a method to model the transformation from side-view to frontal poses in the deep feature space. They use a deep residual mapping (DREAM) block and adaptively add residuals to a deep representation. The residuals are generated via a few additional weight layers, and their number increases with extreme pose variation. Shi and Jain [46] proposed probabilistic face embeddings (PFEs) to model each face image as a multivariate Gaussian distribution encoded by deep neural networks and apply probabilistic embedding for verification. They evaluate their method on seven benchmarks and achieve state-of-the-art performances on the IJB-A [5] and the YTF [7]. Despite a relatively low performance on the CFP [4], they conclude that PFEs can serve as an indicator of the discriminative quality of face images.

2.3.2. Loss Function Design. To explore the distribution of faces in the hyperspace manifold, an alternative idea is to modulate the loss function. While training CNN features, the last layer is dedicated to penalizing the error, and the choice of the loss function can greatly affect the performance.

Rao et al. [47] proposed a DCNN-based face recognition system using the Knot Magnify (KM) loss to enhance the influence of rare complex samples during training to handle the imbalance in training sets. Zheng et al. [48] presented a feature normalization approach for deep networks to augment standard loss functions. They use ring loss to normalize each feature using a convex norm constraint and significantly improve robustness. As another feature aligning method, He et al. [49] learned recognition-oriented alignment using the proposed deformable face net (DFN) method. In this approach, they design deformable convolution modules for learning the geometric transformations adaptively according to pose. They designed two loss functions, namely displacement consistency loss (DCL) and identity consistency loss (ICL), to minimize the variation caused by different poses.

Zhang et al. [50] proposed pose-guided margin loss (PGM-Face). They transform features into the target pose while preserving the identity information using a pose-guided representation transfer network (PGRT-Net). Their method clusters each face feature according to the head pose using the discrete Gaussian distribution label (DGDL) and apply a pose-guided representation transfer network for rotation at the feature level. They obtain the highest accuracy on the CMU Multi-PIE [3]. Zhang et al. [51] modified the triplet loss and derive the triplet angular loss (TAL) to obtain a pose-robust feature embedding. Their method aims to normalize features using TAL so that faces under varying poses can be clustered around the corresponding frontal class center. They attain competing performances on the CFP [4].

2.3.3. Attention Guided Networks. As another feature refinement method, some studies propose focusing on specific data parts that may contain more important information depending on the context. They use attention modules [52] to find and enhance significant parts of the feature. Wang and Guo [53] aimed to identify informative regions to extract local feature information at multiple scales and proposed the local and multi-scale convolutional neural networks (LS-CNN). In this work, channel attention and spatial attention are fused to achieve better performance. It is one of the highest performing methods on the CFP [4]. He et al. [54] proposed a method called PCSAFace where they apply angular regularization to cluster centers of identities and introduce channel and spatial attention to refine face features. Huang and Ding [55] presented a frontalization method where they learn feature embeddings between nearby poses and transform features progressively. They use the channel attention module to modulate the loss function to capture the common deep feature elements at different poses. Their method is the highest performing feature-based method on the CFP [4].

2.3.4. Multitask Learning. Due to the promising performances of individual networks, some studies introduce multitask learning and assemble multiple networks to learn different tasks jointly. In general, these methods aim to disentangle side factors such as pose variation from the identity classification task at higher levels of the network and feed each task into different sub-networks to achieve task-specific classification.

One such method is proposed by Xiong et al. [56], where they use a conditional convolutional neural network (c-CNN) forest for multimodal face recognition. Their method learns a decision tree jointly with an ensemble of pose-specific c-CNN models to explore the underlying modality distribution. It is one of the earliest studies tested on all views of the CMU Multi-PIE [3]. Peng et al. [57] presented a feature disentanglement method to separate the temporal-variant factors like pose and expression from the identity. They use a 3D morphable model to generate face images in multiple poses for data augmentation and then employ a deep neural network VGGFace [39] to jointly learn the identity and nonidentity

features to achieve a rich feature embedding. Finally, they use metric learning for feature disentanglement to find the identity and the pose.

Yin and Liu [58] explored multitask learning and proposed a pose-directed multitask CNN. In their method, they learn pose-specific identity features and apply nonlinear mapping to estimate the correct identity. Naveen and Sivakumar [59] implemented an ensemble convolutional neural network (ECNN) for face recognition. They train multiple pose-specific models and combine predictions of each model by automatically assigning loss weights using the Bat algorithm (BA).

2.4. Discussion. Handling pose variation at the feature level is a fast and reliable method that requires less computational power. Initial methods focus on extracting descriptive features using shape and texture to ensure robustness to pose variations. These methods depend highly on alignment performance and require reliable face and landmark detection. Due to the nonlinear changes under extreme pose variation, these methods are not very robust to pose variations and perform poorly. Over the years, many improvements have been observed by adapting deep learning networks to pose invariant face recognition problems. With deep learning, it became possible to obtain robust feature extraction and classification through extensive training.

Studies that use feature-based pose handling predominantly tested their system on the CFP [4] benchmark. Among many proposed methods, attention-guided networks achieve the highest performance. The second most effective method is the loss function design. Both these methods achieve more defining features by enhancing the more informative parts of the data or penalizing the error. In Section 5, we compare the performances of feature-level alignment approaches in detail. We observe that feature-based methods currently stand as the state-of-the-art, and many image-based and set-based approaches include feature-based alignment to improve their performances. This indicates that a robust feature-level pose alignment can be achieved using deep learning after fine-tuning the network design, depending on the task and data.

3. Image-Based Pose Handling

Extensive research on face recognition highlights that pose variation can significantly impact recognition performance more than within-subject variations [1]. To address this, researchers reconstruct faces in arbitrary poses, which is a challenging task due to the self-occlusions and nonlinear changes evident in the 2D projection of the face under different view changes. Resynthesized images may suffer from severe artifacts and texture losses after significant pose variations. Therefore, several methods guess or eliminate missing information by improving representation methods or models for face synthesis. While early methods handle these challenges using representation learning techniques, with the advances in deep learning technologies, it is now possible to generate photorealistic images of the face and extract identity-preserving features robustly.

3.1. Representation Learning. In representation learning, algorithms discover features to model the nonlinear change of the facial texture caused by pose variation. We categorize algorithms in this context as those that initially augment data by generating images at different viewpoints and subsequently concentrate on representation learning methods for recognition. One of the earliest methods that explore the synthesis of the 3D face is proposed by Blanz and Vetter [60], where they use single images to derive a morphable face model. Inspired by this method, early techniques apply 3D Morphable Models (3DMM) to align the face.

One such method is proposed in Zhang et al. [61], where use frontal and side-view mug shots to learn 3D face shapes, synthesize faces in arbitrary poses and compare reconstructed images using PCA and LBP. Another approach for gallery augmentation is the probabilistic facial expression recognition generic elastic model (PFER-GEM) [62] for constructing a 3D model from a single frontal image. Moeini and Moeini [62] rendered rotated face images from each gallery image for the probe image's estimated pose angle, extract features using dual-tree complex wavelet transform (DT-CWT), and perform recognition using support vector machines (SVMs).

Crispell et al. [63] estimated the pose using 2D landmark locations and render synthetic training data exhibiting various poses and illumination. The optimal alignment is adaptively chosen among randomly rendered images using a deep neural network-based recognition system. Hu et al. [64] introduced LDF-Net to learn the displacement field, reflecting the pixels' shifting relationship during frontalization. They combine the advantages of 2D and 3D methods to train the displacement field network, which acts as a pose estimator while transforming the image to the frontal pose. For final recognition, they use linear discriminant analysis (LDA). Liang et al. [65] proposed a mugshot-based face recognition method. They use the frontal- and side-view images to generate a 3D face model and synthesize face images in predefined poses to enlarge the gallery set. They compare an arbitrary view probe image to the generated gallery images and find the image with the highest similarity to decide its identity using state-of-the-art deep learning-based face classifiers.

3.2. Deep Learning. Since its first appearance, deep learning techniques have demonstrated robust invariance to factors that arise in uncontrolled environments, such as pose or illumination [38]. The first contribution was fast and accurate face or landmark detection, which enhances the capacity to model the face. Later, deep learning was adapted to face modeling, giving rise to realistic face reconstruction systems. The generative adversarial networks (GAN) [66] are the most popular. As depicted in Figure 4, they consist of a generative network that synthesizes an image and a discriminative network that decides if the image is genuine. Several strategies have been studied to improve performance, such as feature disentanglement, loss function design, region-based learning, and assembled networks.

3.2.1. Feature Disentanglement. In feature disentanglement methods, an auxiliary module is included in the system to disentangle identity information from accompanying variations

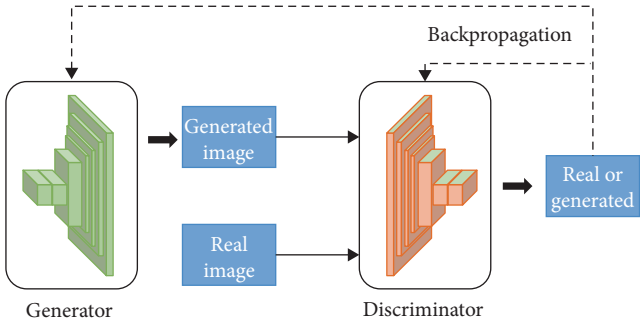

FIGURE 4: GAN architecture. A GAN network combines a generator network synthesizing the image and a discriminator network that decides whether the constructed image is real or fake.

caused by pose or illumination changes. Unlike feature embedding refinement methods, feature disentanglement occurs in the synthesis stage. To achieve this, a conditional encoder–decoder network is usually employed as the generative part of the GAN.

Tran et al. [67, 68] proposed the DC-GAN and improved their method to present DR-GAN to synthesize and discriminate faces in arbitrary poses. Their network consists of an encoder–decoder structured generator and a discriminator that is designed as a multi-task CNN with three components for real/fake, identity, and pose classification. FF-GAN proposed by Yin et al. [21] incorporate the elements from a deep 3D morphable model (3DMM) to a two-way fused encoder-decoder generator. Using 3DMM coefficients as priors to the GAN provides control over shape, expression, and texture. Rong et al. [69] improved the typical GAN network in the proposed feature-improving GAN (FI-GAN) by including an intermediate module called the feature-mapping block to the encoder–decoder type generator. It is similar to DREAM-block [45] and transforms intermediate features into the frontal space.

Fu et al. [70] proposed a conditional encoder–decoder network for high-resolution extreme face manipulation. They use two encoders dedicated to the input and the boundary images to estimate pose and expression and a proxy network to disentangle structure and texture to feed into the decoder. Marriott et al. [71] integrate a 3DMM into the GAN in their proposed 3D GAN to learn a nonlinear texture model with shape, texture, expression, pose, and background parameters via conventional GAN training. They achieve the highest accuracy on the CFP [4]

While many methods rely on estimating the pose degree or 3D information, Zhong et al. [72] used an end-to-end deep learning method to learn the most suitable 2D transformation parameters at face detection. They add a localization network to predict a bounding box that defines the geometric transformation and guides warping. Similarly, Wang et al. [73] used a lightweight Pseudo Profile Facial-Generator (PPFG) in their LS-CNN. They preserve the pose of the input and make minor pixel-wise modifications to generate pseudo-facial images with a small embedding distance to the corresponding frontal image.

3.2.2. Loss Function Design. Some methods altered the loss function design to achieve identity information unaffected by pose changes and enhanced the identity information. Zhao et al. [74, 75] presented dual-agent generative adversarial network (DA-GAN), where they modify standard GAN by using dual-agent architecture as the discriminator to combine the prior knowledge from the side-view faces' distribution with the domain knowledge of faces. They combine pose perception loss, identity loss, and adversarial loss for artifact removal. Tang et al. [76] proposed a modified version of DR-GAN [67] called the RL-WGAN, where the discriminator is built upon the Wasserstein loss, and the identity/pose classifiers are separated from the discriminator.

3.2.3. Region-Based Learning. In extreme pose variation, face reconstruction suffers from nonlinear facial appearance, self-occlusions causing texture loss, and reconstruction errors. However, these changes are not affecting all local regions equally. Therefore, exploring the combination of holistic and local descriptors in the generator can form a more robust description of the face.

Huang et al. [77] proposed TP-GAN for photorealistic frontal view synthesis. Their method extracts four landmark-located patches and trains separate networks to rotate each patch and the holistic image to its corresponding frontal view. They combine the information from the global and the local pathways to generate the final synthesis. Hu et al. [78] directly use the landmark heatmaps as pose embeddings in their couple-agent pose-guided generative adversarial network (CAPG-GAN). They concatenate the source image, the source pose embedding, and the target pose embedding to synthesize the target view image in arbitrary poses.

Zhao et al. [79, 80] proposed the pose invariant model (PIM) to unify a dual-path GAN called face frontalization subnet (FFN), and a discriminative learning subnet (DLN). FFN consists of a global generator and a local generator with four subnetworks dedicated to center-cropped local patches of the left eye, the right eye, the nose, and the mouth. Luo et al. [81] proposed the FA-GAN consisting of a geometric preserving module and a face disentanglement module (FDM). They use graph convolutional networks (GCNs) to jointly explore different face regions' spatial and semantic relations.

An alternative way to enhance local information is by using attention modules. Yin et al. [82] presented dual-attention GAN (DA-GAN) for frontalization and introduce two complementary attention mechanisms, a self-attention-based generator and a face-attention-based discriminator, to handle local features better. In the GSP-GAN, Luan et al. [83] modeled the generator as an autoencoder with an encoder for feature extraction and a decoder with a self-attention block to focus on more important regions while preserving face geometry. In their deep attention-based face frontalization (DAFF) method, Shao et al. [84] learned an attention branch simultaneously with a frontalization branch and guide the generator to pay more attention to the foreground regions.

3.2.4. Assembled Networks. Recent studies propose to tackle pose variations by assembling multiple networks. While some methods focus on improving reconstruction by training multiple networks for separate tasks, others combine networks to enhance the descriptive quality of the extracted features.

Researchers encounter a challenge in dealing with artifacts and texture losses that occur after substantial pose variations, leading to the proposal of several methods to restore the image. Deng et al. [85] concentrated on UV mapping, which is a texture mapping process used in 3D modeling to project the surface of a 3D model to a 2D image. They proposed an adversarial UV completion framework (UV-GAN) and feed an incomplete UV texture learned by the 3DMM to the GAN that learns how to reconstruct the missing regions.

Cao et al. [86] proposed 3D-aided duet generative adversarial networks (AD-GAN) to break down the face synthesis into two tasks using a face normalizer GAN and a face editor GAN. In normalization, the network estimates the correspondence field of texture mapping to learn the face geometry while reconstructing the frontal face. Then in the editing stage, the pose is predicted, and the face is resynthesized in the desired views. Later in [87], they extend their idea to present the high-fidelity pose invariant model (HF-PIM) and introduce adversarial residual dictionary learning (ARDL) to supervise warping and a multiperception guided loss to handle the noise introduced by multiple training images.

Xu et al. [88] introduced gated deformable face synthesis network and use a dual network as the generator to learn a deformation prior. Using a soft-gate mechanism, they fuse deformable and primitive features. Tu et al. [89] aimed to restore high-quality faces from low-quality in-the-wild images with their proposed MDFR model. Their system consists of a dual-agent generator, a 3D-based pose normalization module, a pose-conditioned discriminator, and an identity-conditioned discriminator. They apply task-integrated training for the generator by combining a face restoration subnet and a face frontalization subnet. Zeng et al. [90] proposed a deep shape reconstruction and texture completion network (SRTCNet). Their system consists of a correspondence network to align the image to a 2D shape, an inpainting network to complete occluded areas, and a shaping network to reconstruct the 3D geometry.

Some studies aim to measure the reconstruction quality and use this information to guide recognition. Shao et al. [91] trained from multiple domains and present well-advised pose normalization network (WAPNN) to handle feature quality variance caused by domain differences. They train a face normalization network and apply adaptive weight selection to learn domain classification simultaneously. Then, they extract features from both the original side-view image and the synthesized frontal image and apply feature fusion guided by quality scores. Liu et al. [92] presented cross-pose generative adversarial networks (CP-GAN), which consists of an autoencoder as the generator and a siamese convolutional network as the discriminator. They fuse the deep representations of the original and the reconstructed images using face rotation as the quality coefficient.

Encouraged by the reconstruction performance of GANs for small view changes, several studies propose progressive frontalization and break down the face synthesis operation into a series of near-view-based rotation tasks. Liu et al. [93] presented a multitask network named the PPN-GAN, where they sequentially assemble multiple GANs specialized for rotation tasks under 30° pose variances. They decide on the starting synthesis stage by estimating the pose of each testing sample using the geometry of five facial landmarks. Xu et al. [94] proposed the FFlowGAN to generate different views progressively. At each step, they estimate the correspondence between the input and the target face using a face flow module to guide warping for 15° rotation. Using a self-selective face fusion method, they handle the reconstruction artifacts and refine the synthesized image. Instead of detecting the pose of the input image, they iterate until they reach the optimal view.

Some recent studies combine multiple networks to handle rotation in uncontrolled environments, each specialized for different rotation tasks. Zhang et al. [95] combined two generators for nonfrontal-to-frontal synthesis and frontal-to-nonfrontal synthesis in the proposed CCFF-GAN. One shared discriminator forces the two generators to be the inverse mappings of each other. Similarly, Zhang et al. [96] presented IPG-GAN where they mutually learn the identity and the head pose. They reconstruct faces according to the gallery and test poses, compare both views and fuse the results. Taherkhani et al. [97] presented the coupled conditional generative adversarial network (CP-GAN) that consists of two conditional GAN-based subnetworks dedicated to frontal and side-view domains and aimed to find the projection of side-view and frontal images into a shared embedding space.

3.3. Discussion. Image-based pose handling methods aim to generate faces in different views to enable comparison in similar conditions. The reconstruction performance of early methods depended highly on face and landmark detection and tried to eliminate the effect of reconstruction errors using representation learning algorithms. More recent methods enable deep learning architectures to ensure photorealistic synthesis. Considering realistic face reconstruction, GANs dominate the research area with their capacity to improve alignment, augment the training data, or directly minimize the interclass variation.

Studies mainly focus on face frontalization and are tested on CMU Multi-PIE [3] and the CFP [4]. Compared to feature-based or set-based methods, they perform poorly under uncontrolled settings. Most systems cannot achieve state-of-the-art performances on the CFP [4]. In contrast, their performance is consistently superior on CMU Multi-PIE [3], where pose variation is more significant, but a more controlled setting is in place. It is worth mentioning that the main focus of GAN is obtaining realistic reconstruction, and recognition is only handled as a second task. A more in-depth performance comparison can be found in Section 5.

Among image-based methods, the ones that achieve the highest accuracies [73, 94, 96] argue that increasing the pose angles leads to higher reconstruction error. To tackle this

problem, they either fuse comparisons under different views [96] or combine multiple reconstructions under small rotations progressively to achieve higher rotation [94]. Alternatively, Wang et al. [73] only make small pixel-wise changes to decrease the embedding distance between the images preserving their original poses. These findings suggest that, despite ensuring identity-preserving details, the performance of these methods is notably affected by cluttered backgrounds and reconstruction errors, particularly when facing substantial pose variation.

4. Set-Based Pose Handling

In set-based face recognition, the objective is to analyze sets of images containing multiple observations that capture various characteristics of the same subject. These studies are distinguished by their testing approach rather than training strategies. In surveys regarding video-based face recognition [11, 17], feature-based and image-based techniques are categorized under one-to-many augmentation or many-to-one normalization, while set-based methods involve many-to-many comparisons. Most methods commonly integrate feature-based and image-based techniques to enhance the available features or to learn the intrinsic 3D geometric information [17]. Early studies concentrated on dictionary learning to describe faces as linear combinations of primary feature vectors or metric learning to transfer the data to a more descriptive feature space. Recent advancements have leaned toward deep learning, where several techniques are proposed to improve the network design.

Image sets containing varied poses provide a ground for comparison under similar views. To ensure that, some studies disentangle features or apply multitask learning to separate pose from identity. On the other hand, not all images in a set offer the same level of information, and diverse conditions can impact overall performance. Consequently, some methods introduce aggregation networks to enhance image or feature components containing reliable information. Unfortunately, it is not always evident whether these systems favor frontal views over extreme poses. Nevertheless, they deserve attention since they address extreme pose variation and report state-of-the-art performances on the IJB-A [5].

4.1. Dictionary Learning. To handle the data limitation, dictionary learning algorithms are proposed to describe faces as a linear combination of some basic feature vectors. These vectors are called dictionary atoms and are learned from training samples to form a matrix called the dictionary [98]. An illustration is given in Figure 5. Dictionary learning algorithms are computationally efficient for a moderate training size. However, the dictionary may become inefficient with large training sets and easily overfit due to the noisy instances.

In an early method, Schroff et al. [99] created a face library with significant variations and represented each face by generating image signatures as an ordered list based on the similarities to the images in the face library. One advantage of this method is that it handles pose variations implicitly. Zhang et al. [100] presented a joint dynamic sparse representation method, where they compute class-specific

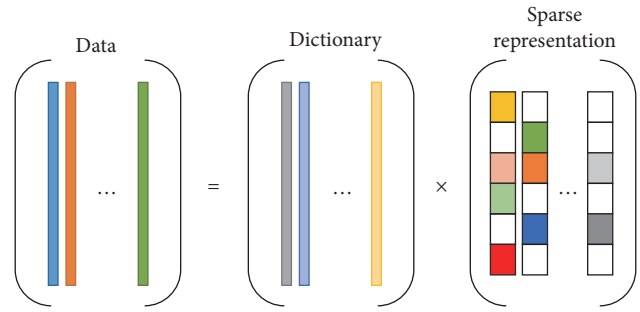

FIGURE 5: In sparse dictionary learning, each data element is represented as a linear combination of a sparse vector and a matrix called the dictionary that is learned from the training samples. In this case, each data element corresponds to a set of images of a subject.

sparsity patterns and generate multiview probe faces to associate images of similar viewpoints in different sets.

The distance measure in most image set classification methods uses either the distance between specific samples of the two sets or the similarity of the sets' structures. Hu et al. [101] introduced a between-set distance measure, where they model the image sets using affine hulls and compute the distance between sparse approximated nearest points of two image sets. Later, they improved their method in [102] by introducing a convex formulation to handle the image sets' nonlinear structures.

Bhatt et al. [103] combined temporal and intrapersonal variations to form a discriminative video signature as an ordered list of still-face images. They use a dictionary of normalized face images and generate a ranked list by ordering them based on the similarity to the input image. They cluster each list to enhance similarities between images in comparable viewing conditions. Similarly, Chen et al. [104] constructed structured dictionaries that implicitly encode temporal, pose, and illumination information. They cluster each face video into subsequences containing images with a specific pose or lighting conditions. Then, they learn subdictionaries for each partition and combine these to form the sequence-specific dictionary. To eliminate the effect of pose variation, Qiu and Chellappa [19] proposed a compositional dictionary learning approach. Their method decomposes each face image into identity, pose, and illumination using sparse representations and applies feature-level pose alignment to compare images of different views.

Bahrampour et al. [105] expanded the dictionary learning method for feature-level fusion and proposed a joint sparse representation method. Their approach treats each view range as a distinct modality and simultaneously learns multipose, supervised dictionaries to generate corresponding feature vectors to be fused using the joint sparse representation method. Hajati et al. [106] combined texture and motion and presented a new derivative sparse representation approach (DSR) for face and texture recognition. They use short-length videos to construct local linear subspaces and sparsely represent them on the Grassmann manifold using the novel DSR approach. Lu et al. [107] proposed the deep simultaneous feature and dictionary learning D-SFDL to simultaneously learn discriminative features and dictionaries

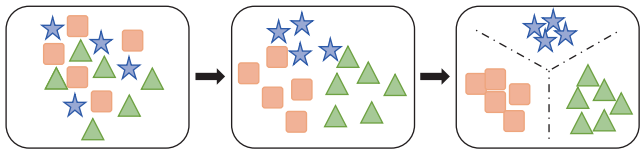

Figure 6: Metric learning process. We represent the classes of each set with different symbols. The aim of metric learning is to find a mapping to an embedded space where the similarity of objects is preserved while dissimilar objects are separated.

for image set-based face recognition. They use deep learning to model the nonlinearity information of faces and learn multiple hierarchical nonlinear transformations and class-specific dictionaries to improve recognition performance.

4.2. Metric Learning. Metric learning is a feature embedding method that transforms the data into a more descriptive feature space, as depicted in Figure 6. In the presence of multiple samples, the algorithm can better learn the distinctive characteristics of each class.

Wang and Chen [108] proposed the manifold discriminant analysis that model image sets as local linear models and separate them using a hierarchical divisive clustering approach. They learn a linear discriminant function to transform the local models and calculate the similarity between two sets by finding the minimum pairwise local model distance. Later, Wang et al. [109] improved this method and presented the covariance discriminative learning approach that models image-set with its covariance matrix and maps it from the Riemannian manifold to Euclidean space. Hayat et al. [110] used a class-specific deep reconstruction model to describe the image sets' geometric structure. Their method automatically detects faces and performs PCA whitening followed by LBP [34] to extract features. Then they learn a deep reconstruction model for each class using Gaussian restricted Boltzmann machines (GRBMs). They reconstruct each image, and using three voting strategies, they evaluate the reconstruction error to determine its class. Haghighat et al. [111] combined feature vectors extracted from multiple views using a feature fusion method named the discriminant correlation analysis (DCA). They treat each view as a different modality and apply DCA to find the optimal linear combinations for each modality that maximize the pairwise correlation between two feature sets.

Chen et al. [112] trained a DCNN model to represent faces and derived the joint Bayesian metric for unconstrained face verification. Their approach models the joint distribution of feature vectors as a Gaussian and measures the similarity of two sets using the interclass and intraclass log-likelihood ratio. The template adaptation method proposed by Crosswhite et al. [113] is a transfer learning approach combining deep CNN features with a simple SVM classifier. They train two linear SVM classifiers for the probe and gallery templates at test time and average the scores to find the final similarity score.

Bodla et al. [114] proposed a set-based face recognition method that fuses the features generated by two different DCNN models using weighted average pooling and projects them to a more descriptive nonlinear space. They train a joint Bayesian metric learning [112] using tight bounding boxes and triplet probabilistic embedding (TPE) [42] with loose bounding boxes. If one network is more robust to pose variation, the fusion network retains this property and produces an enhanced representation.

Wang et al. [115] proposed to represent the image set as the Gaussian mixture model and present an alternative discriminative learning approach on the Riemann manifold of Gaussians. They develop two learning frameworks: the kernel-based framework that projects data to some Euclidean space, maximizing interclass differences, and the graph-based framework that describes the geometric structure by constructing a graph associating the Gaussians.

Cheng et al. [116] proposed the duplex metric learning (DML) method for image set classification. In their approach, there are two progressive metric-learning stages. In the first stage, they train a discriminative stacked autoencoder (DSAE) to obtain new feature mappings. In the second stage, they train an image classifier for set-based face recognition. They achieve state-of-the-art performance on the YTF [7]. Duan et al. [117] achieved similar performance with their proposed UniformFace. In this study, they introduce a loss function to distribute the class centers uniformly on the feature space. Combined with SoftMax loss, the uniform loss function constrains interclass relationships by assuming equally distributed class centers on the hyperspace manifold and maximizing the minimum distance between class centers.

4.3. Deep Learning. Sets of images containing diverse poses and illumination conditions offer a foundation for extracting more comprehensive information about subjects compared to single-image-based systems. In this context, numerous studies have employed deep learning for feature disentanglement or multitask learning to separate identity characteristics. Furthermore, aggregation networks are proposed to exploit features or image components in order to choose informative parts.

4.3.1. Feature Disentanglement. Incorporating feature disentanglement into deep learning is a widely studied method that inserts an additional module in the system to isolate and enrich identity characteristics. These methods commonly use an encoder–decoder network to disentangle factors like pose variation from identity-preserving characteristics. The presence of multiple samples for each class facilitates learning the prominent elements that describe the identity and improves recognition performance.

The collaborative random faces proposed by Shao et al. [118] is a feature-level alignment method that uses supervised autoencoders to map images with variant poses to both the frontal pose and multiple random signals for augmenting features. They ensure within-class compactness and between-class separability using the Fisher criterion and synchronize multiple autoencoders to disentangle the identity feature from the pose. Alternatively, Xue et al. [119] proposed a video-based face recognition system where they fit a 3D morphable model on the frames of a video to extract the UV texture and conducted UV completion using a GAN-based network. They render

frontal views and extract deep features, which they average to describe the set. Their method is among the best-performing methods on the YTF [7].

4.3.2. Multitask Learning. As previously mentioned, in multitask learning multiple networks are combined to learn different tasks jointly. This approach facilitates the separation of factors attributed to pose variation from identity at higher levels of the network.

Ranjan et al. [120] presented a multitask learning framework and improve their method in [121] to simultaneously detect the face and estimate the key points and pose. They align each face to a canonical view using the detected landmarks and apply a TPE for classification. Later in [122], they added the deep pyramid single shot face detector (DPSSD) to their system and the crystal loss function for training the face verification network. They apply a score-level fusion to generate a set-based representation and average the similarities obtained by the network.

Zhao et al. [123] presented the multiprototype network (MPNet) model to learn multiple prototype face representations from the image sets. They apply a Dense SubGraph (DSG) learning subnet to cluster each face set into subcategories sharing similar conditions that are not predefined. Once each image set is partitioned into multiple prototypes, they apply MPNet to compare image sets at the prototype level. They achieve the highest verification rate on the IJB-A [5] and the highest recognition accuracy on the YTF [7].

An et al. [124] proposed the adaptive pose alignment method to align each face to a related template according to facial poses and apply a feature normalization method that uses deep representations. Their approach determines the optimal number of alignment templates and their corresponding yaw angles using the k-means algorithm. They estimate the pose using the five landmarks and perform a canonical transformation to align the face to the corresponding template. They represent each image set by averaging features and unit normalizing. They achieve one of the highest verification scores on the IJB-A [5].

Some studies tackle pose variation by reconstructing faces in multiple views and assembling pose-specific networks. AbdAlmageed et al. [22] computed the pose angle using detected landmarks and reconstruct multiple poses using 3D rendering to augment the training and the gallery dataset. They generate pose-specific CNNs using real and synthesized images for frontal, half-profile, and near-side-view poses. Then, they compare image pairs of two sets and fuse the similarity scores using SoftMax weights. Masi et al. [125] proposed a similar method and synthesized face images in challenging appearance variations from original 2D images using different 3D generic shapes. Later, they propose a more sophisticated method called pose-aware models (PAM) [20, 126], where they detect facial landmarks to estimate the facial pose and align each image to five reference poses. They learn multiple pose-specific CNNs for recognition and compute the pairwise similarity scores. Then, they apply score pooling using SoftMax to get the final similarity score.

4.3.3. Aggregation Network. Aggregation, as a technique, involves merging multiple data streams into a single, more informative entity. In set-based face recognition, the images within a set are not equally informative, and particularly in uncontrolled settings, certain images are significantly influenced by variant conditions. Aggregation networks aim to prioritize informative features or image components among many noisy candidates.

Chowdhury et al. [127] used a bilinear CNN model to describe images and combine features with detected semantic image patches using a feature-level pooling technique. They train a linear SVM classifier and use max-pooling over the SVM scores to select the representation that best describes the set. Goswami et al. [128] presented a frame selection approach based on entropy using discrete wavelet transform. They use deep learning to select the feature-rich frames, feeding the representations learned from an autoencoder to a deep Boltzmann machine (DBM). They assign a feature-richness score to each frame to weight pairwise comparisons.

Yang et al. [129] proposed a deep CNN network and aggregated the features with two attention blocks to form a compact and discriminative face representation. Their system uses an adaptive weighting method that automatically learns to enhance high-quality images with discriminative characteristics. Zhao et al. [130] proposed a set-to-set distance measure to calculate the similarities of deep features using the cosine similarity metric. They apply kNN-average pooling to find the similarity between the two sets. Their pooling method performs like a weighting procedure for frame selection.

Rao et al. [131] combined metric learning with adversarial learning to learn a discriminative aggregation network (DAN). Instead of aggregating at the feature level, they use a GAN-based framework that directly eliminates images containing misleading information and generates more discriminative face images. They feed these images into a deep feature extraction network and apply mean-pooling to find the feature representing the set. They obtain the highest recognition accuracy on the YTC [6]. Gong et al. [132] proposed a component-wise feature aggregation network (C-FAN) and added an aggregation module to the end of a CNN-based network. The aggregation module assigns a quality score to the feature components that serve as weights for feature pooling.

Arachchilage and Izquierdo [133] devised a hierarchical clustering method named ClusterFace to guide classification based on the feature space distribution. They introduce an intermediate learning step that first groups images under similar conditions and then finds gradual and incremental associations to assign samples of the same identity to a single cluster. They use these clusters to represent image sets and perform a nearest-neighbor search for recognition. Liu et al. [134] introduced a dependency-aware attention control (DAC) network. They formulate a Markov decision process (MDP), combining the DAC network with an attention control module to find the optimal weights for each image. The weighting depends on the quality of the image and the diversity of the image set to retain images with extreme poses. They also propose a metric

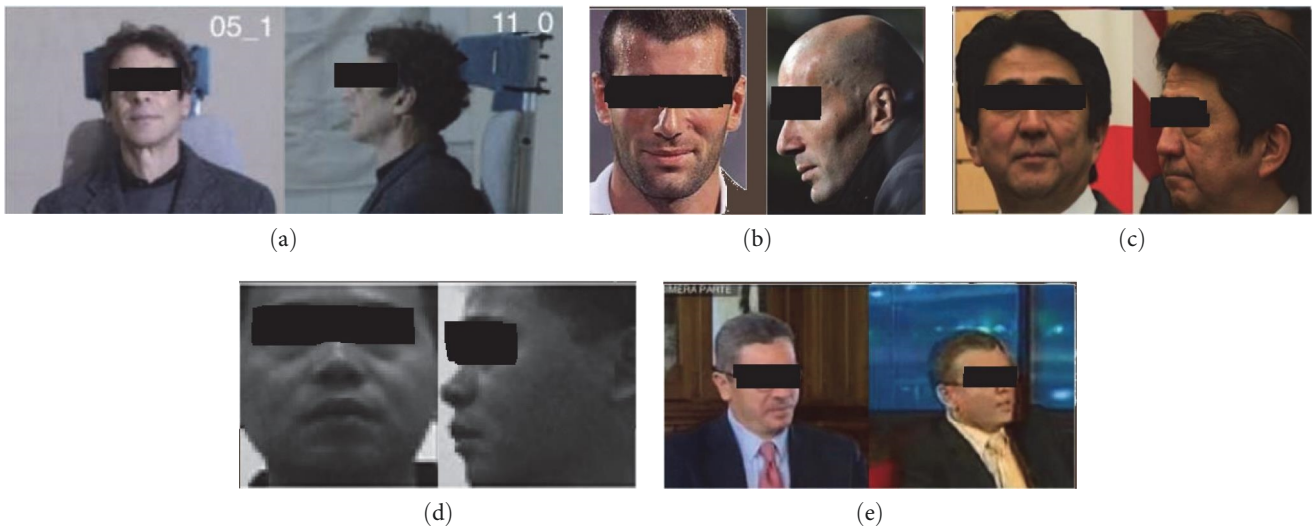

FIGURE 7: Sample images illustrating pose variations from datasets: (a) CMU Multi-PIE [3], (b) celebrities in frontal-profile (CFP) [4], (c) the IARPA Janus Benchmark-A (IJB-A) [5], (d) YouTube celebrities (YTC) dataset [6], and (e) YouTube faces (YTF) dataset [7].

learning-based pose-guided representation method for pose alignment and compare similar pose groups to find the similarity between two sets.

Bai et al. [135] proposed a component-wise aggregation method where they establish the spatial structure of the feature maps to extract local features. They compare local information in various images using a local feature enhancement network (LFENet) and eliminate parts suffering from occlusion or low quality. Using an attention map, they transfer enhanced local information to the related region and form an aligned face descriptor. By average pooling, they find a single feature representing the image set.

4.4. Discussion. In set-based face recognition, the availability of multiple images provides the possibility to portray different characteristics of the same subject. Early methods proposed dictionary learning methods to describe faces as linear combinations of primary feature vectors. These methods perform poorly regarding extreme poses since linear methods are not well suited to describe the nonlinearity of the face.

Another method that has been widely studied is metric learning, which transfers the data to a new feature space, increasing the distance between elements from different classes. Some recent methods [116, 117] apply deep metric learning techniques where they improve the loss function design and achieve state-of-the-art performances on the YTF [7]. This is consistent with our previous observations that the loss function can greatly affect performance.

Alternatively, image sets containing variant poses establish a ground for comparison in similar views. To ensure that, some studies untangled features to separate pose from identity and performed recognition after pose alignment. Most recent feature disentanglement methods [119, 123, 124] consistently attain superior performances. Among all set-based methods, the MPNet [123] achieves the highest accuracy both on the YTF [7] and the IJB-A [5].

A method that recently gained popularity is aggregation. Especially in uncontrolled settings, not all images within a set are equally informative, and the variant conditions can affect the overall performance poorly. Therefore, some methods aim to enhance images or feature parts that contain reliable information. Some aggregation methods [131, 133, 135] present high results on the YTC [6] and IJB-A [5]. Unfortunately, these methods do not explicitly mention if their system favors frontal views over extreme poses.

5. Datasets and Benchmark Results

To elevate research on pose-invariant face recognition, various extensive datasets capturing extreme poses have been collected, including CMU Multi-PIE [3], CASIA [136], CFP [4], YTC [6], and YTF [7]. Most image-based studies conduct tests on the CMU Multi-PIE [3] due to the controlled setting and considerable pose variation. Specifically for the frontalization studies, Sengupta et al. [4] published the CFP dataset and included a protocol to encourage researchers for frontal-to-side-view face recognition. Video datasets such as the YTF [7] and YTC [6] were automatically collected from the web to guarantee the natural variability in the pose, illumination, and other factors encountered in everyday life.

In addition, several benchmarks are designed to provide a standardized way to compare the performances of different face recognition algorithms under various conditions, such as changes in lighting, pose, expression, and the presence of occlusions. A face recognition benchmark is a standardized, well-defined set of datasets, protocols, and evaluation metrics. Well-known face recognition benchmarks include the labeled faces in the wild (LFW) [18] and IARPA Janus Benchmark-A (IJB-A) [5]. Recent studies that focus on face recognition in uncontrolled settings either prefer the IJB-A [5] benchmark or the video datasets YTF [7] and YTC [6]. Some illustrative examples from each dataset are given in Figure 7.

5.1. CMU Multi-PIE. The CMU Multi-PIE [3] is one of the most widely used face datasets addressing challenges due to pose and illumination variation. The dataset contains a total of 755, 370 images featuring 337 subjects, taken from 15 viewpoints and under 20 distinct illumination conditions. Subjects are recorded in up to four recording sessions over a span of 6 months. This dataset is an extension of the CMU-PIE [16], which includes 68 subjects and 13 poses.

CMU Multi-PIE is not designed to test a specific task, and the controlled pose variation enables researchers to conduct experiments on different scenarios. Consequently, there is no unified protocol, and most studies adopt the two most popular protocols in the literature. In the first protocol, only images from the first session are used. The systems are trained with the first 150 subjects, and the remaining 100 subjects are used for testing. The second protocol includes all sessions with 200 subjects in the training set and 137 subjects in the test set. In Table 2, we report the frontal-to-side-view identification performances of various studies.

We can report only a few studies that handle pose alignment at feature level [49, 50, 56, 58, 59]. They generally perform lower than image-based methods, except Zhang et al. [50], who propose a new loss function design and achieve the highest accuracy.

Due to the controlled setting and the availability of considerable pose variation, studies that use image-based pose handling methods predominantly evaluate their performance on the CMU Multi-PIE. These studies often focus on GAN-based architectures for frontal image reconstruction. While many GAN-based approaches prioritize realistic reconstruction over face recognition, the results suggest their ability to preserve identity details.

When examining these strategies, a continuous improvement in performance is evident since 2015. Recognition accuracy has risen from 47.26% to 96.23% on the first protocol and from 61.20% to 94.37% on the second protocol. Notably, Zhang et al. [50] reported the highest accuracy on both protocols, employing a feature-based alignment method, and achieve 97.45% and 94.37% mean accuracy on the second protocol at $\pm 75°$ and $\pm 90°$, respectively. Among image-based alignment methods, both assembled networks and feature disentanglement consistently outperform others. It is widely argued that significant pose variation contributes to increased reconstruction error. Consequently, image-based methods achieving the highest performances [73, 94, 96] focus on minimizing rotations, refining feature embedding, or progressively combining multiple reconstructions to attain the desired pose.

5.2. Celebrities in Frontal-Profile (CFP). Sengupta et al. [4] introduced the CFP dataset to support research on frontal to side-view verification problem. The dataset comprises 10 frontal and four side-view images collected from 500 individuals. The authors presented a verification protocol that include both frontal-to-frontal and frontal-to-side-view experiments. Additionally, they conducted experiments to measure human performance on these recognition tasks, establishing a baseline human verification accuracy for the frontal-to-side-view setting at 94.57%. In Table 3, we provide the reported performances of the CFP dataset using the frontal-to-side-view protocol.

When we examine the results since the dataset's initial publication in 2016, a notable and sustained enhancement in performance is apparent. The suggested baseline human performance of 94.57% was surpassed, particularly around the year 2020, largely attributed to advancements in feature-based alignment methods.

CFP stands out as the preferred dataset for feature-based pose alignment studies when compared to other benchmarks. Among feature-based methods, Huang et al. [55] achieved the highest verification accuracy at 98.40% on the frontal-to-side-view protocol by integrating the channel attention module to their networks, effectively addressing pose variation progressively. Another attention-guided network is proposed in [53], utilizing spatial attention to leverage local information alongside global face appearances, resulting in a mean accuracy of 97.17% on the frontal-to-side-view protocol. Designing new loss functions yielded the second-best-performing feature-level pose handling methods. Using triplet angular loss [51] and pose-guided margin loss [50] produced recognition scores of 97.47% and 96.13%, respectively.

Among image-based methods, using GAN-based networks stands out as the most preferred image-based method for frontalization. Marriott et al. [71] integrated a 3D morphable model into the generator of their GAN-based architecture and achieved the best overall performance of 98.51%. Most image-based methods cannot achieve state-of-the-art performance except one feature disentanglement method [87] and one assembled network [75]. GANS aim to synthesize photorealistic images, but many artifacts occur at extreme poses that highly affect recognition performance. Therefore, we can argue that face recognition for extreme pose variation, GANs still have room for improvement, whereas feature-level pose alignment methods prominently establish the state-of-the-art.

5.3. IJB-A. The IARPA Janus Benchmark-A face challenge (IJB-A) [5] is the first version of Janus datasets related to processing unconstrained in-the-wild face images. In the past years, several extensions of IJB-A are released, namely the IJB-B [137] and the IJB-C [138].

IJB-A contains 5, 712 face images and 2, 085 videos of 500 subjects. The data are collected from the Internet, and the bounding boxes for the face region and the ground truth eye and nose locations are manually annotated. IJB-A images cover a wider variety of head poses and far more challenging viewing conditions than LFW [18] and CASIA [136]. IJB-A also introduces a search protocol for open-set face identification and a protocol to measure verification accuracy. There are 10 training and testing splits with 333 subjects in each training split and 167 in the corresponding testing split, where each set is composed of the mixed media type.

In Table 4, we provide the performance outcomes of studies conducted on the IJB-A dataset. Notably, around 2019, multiple studies reported true accept rates (TAR) exceeding 97.0% at a false accept rate (FAR) of 0.01%, thus establishing the state-of-the-art. Among these, the set-based methods introduced by

Table 2: Comparison of rank-1 identification performances (%) of different methods on CMU Multi-PIE.

Protocol	Category	Publication	Method	±75°	±90°
Protocol 1	Feature based	Xiong et al. [56]	C-CNN forest	60.66	47.26
		He et al. [49]	DFN	88.97	84.07
		Zhang et al. [50]	PGM-Face	**98.27**	**96.23**
		Huang et al. [77]	TP-GAN	84.10	64.03
		Zhao et al. [79]	PIM	91.20	75.00
		Hu et al. [78]	CAPG-GAN	87.40	77.10
		Luan et al. [83]	GSP-GAN	78.40	56.90
		Zhao et al. [80]	PIM	91.20	75.00
	Image based	Shao et al. [91]	WAPNN	93.26	77.23
		Rong et al. [69]	FIGAN	90.24	81.17
		Xu et al. [88]	GDFace	93.74	87.93
		Shao et al. [84]	DAFF	94.02	79.35
		Xu et al. [94]	FFlowGAN	96.62	91.62
		Wang et al. [73]	PPFG	**98.94**	**94.07**
		Zhang et al. [96]	IPG-GAN	97.56	94.72
		Zeng et al. [90]	SRTC-Net	91.10	63.20
		Tu et al. [89]	MDFR	88.52	79.83
		Taherkhani et al. [97]	PF-cpGAN	94.20	88.10
Protocol 2	Feature based	Yin and Liu [58]	p-CNN	87.50	83.90
		Naveen and Sivakumar [59]	p-ECNN	89.03	78.05
		Zhang et al. [50]	PGM-Face	**97.45**	**94.37**
	Image based	Yin et al. [21]	FF-GAN	77.20	61.20
		Huang et al. [77]	TP-GAN	77.43	64.64
		Hu et al. [78]	CAPG-GAN	83.05	66.05
		Zhao et al. [79]	PIM	95.00	86.50
		Tang et al. [76]	RL-WGAN	86.10	75.10
		Liu et al. [93]	PPN-GAN	84.61	83.20
		Cao et al. [86]	AD-GAN	95.30	89.70
		Luan et al. [83]	GSP-GAN	82.90	65.60
		Shao et al. [91]	WAPNN	88.17	71.93
		Rong et al. [69]	FIGAN	88.24	77.03
		Yin et al. [82]	DA-GAN	93.24	81.56
		Zhao et al. [80]	PIM	95.00	86.50
		Xu et al. [88]	GDFace	95.28	90.32
		Cao et al. [87]	HF-PIM	96.40	92.32
		Zhang et al. [95]	CCFF-GAN	88.30	73.90
		Luo et al. [81]	FAGAN	94.30	75.90
		Fu et al. [70]	Hi-Fi face manipulation	96.30	87.40
		Xu et al. [94]	FFlowGAN	97.00	93.01
		Zhang et al. [96]	IPG-GAN	**97.87**	**93.47**
		Wang et al. [73]	PPFG	97.60	93.68
		Tu et al. [89]	MDFR	84.93	69.61

We report the cases where the gallery contains one frontal image and the test set contains side-view poses. The highest accuracies for each category are highlighted in bold.

Zhao et al. [123] and An et al. [124] demonstrated the highest performances, reaching 99.1% and 98.95%, respectively. These methods employ a multitask learning approach to categorize each set into subcategories that share similar conditions without predefined specifications. Another noteworthy set-based approach, LFENet by Bai et al. [135], achieved the third-highest performance at 98.93%. This method incorporates an aggregation technique and employs an attention module to enhance the influence of local regions containing more valuable information.

Image-based methods exhibit slightly lower performance compared to set-based methods. Within image-based approaches, the DA-GAN, introduced by Zhao et al. [74, 75], stands out with the highest true acceptance rate (TAR) of 98.90% at

Table 3: Comparison of classification performances of different methods for frontal to profile protocol in CFP.

Category	Publication	Method	EER	Accuracies
Feature based	Sengupta et al. [4]	Human	5.02	94.57
	Sengupta et al. [4]	Deep features	14.97	84.91
	Sankaranarayanan et al. [42]	CNN + TPE	8.85	89.17
	Chen et al. [43]	FV-DCNN	8.00	91.97
	Lu et al. [36]	Joint bayesian	7.10	92.90
	Peng et al. [57]	Disentanglement	—	93.76
	Rao et al. [47]	Center + KM Loss	—	93.39
	Yin and Liu [58]	p-CNN	5.94	94.39
	Shi and Jain [46]	PFE	—	93.34
	Wang and Guo [53]	LS-CNN	—	97.17
	He et al. [54]	PCSA	—	91.80
	Zhang et al. [50]	PGM-Face	—	96.13
	Zhang et al. [51]	KL-TAL	—	97.47
	Huang and Ding [55]	APM	—	**98.40**
Image based	Tran et al. [67]	DR-GAN	—	93.41
	Zhao et al. [79]	PIM	7.69	93.10
	Tran et al. [68]	DR-GAN	—	93.89
	Deng et al. [85]	UV-GAN	—	94.05
	Zhao et al. [75]	DA-GAN2.0	4.61	95.96
	Tang et al. [76]	RL-WGAN	—	92.74
	Cao et al. [86]	AD-GAN	6.26	94.39
	Shao et al. [91]	WAPNN	6.60	93.66
	Rong et al. [69]	FIGAN	—	94.20
	Xu et al. [88]	GDFace	6.71	94.43
	Cao et al. [87]	HF-PIM	5.86	94.71
	Wang et al. [73]	PPFG	5.70	94.10
	Zhang et al. [96]	IPG-GAN	—	94.16
	Zhang et al. [95]	CCFF-GAN	—	94.30
	Shao et al. [84]	DAFF	5.86	94.34
	Marriott et al. [71]	3D-GAN	—	**98.51**
	Taherkhani et al. [97]	PF-cpGAN	7.21	93.78
	Zeng et al. [90]	SRTC-Net	6.39	—

The highest accuracies for each category are highlighted in bold.

FAR = 0.01%. The majority of other image-based methods use GAN-based architectures, but they fall short of reaching the state-of-the-art. A closer examination of feature-based approaches reveals that most of them struggle to surpass a TAR of 96.0% at FAR = 0.01%, with an exception being the feature embedding refinement method known as PFE, proposed by Shi and Jain [46].

In their study, Pal et al. [139] highlighted a significant frontal bias in yaw poses within the IJB-A dataset, asserting that high performance on such biased datasets does not necessarily indicate robustness to extreme poses. They validated this claim by assessing a model excelling on the IJB-A, only to observe a performance decline on the CMU Multi-PIE, characterized by significant pose variation. We corroborate this pattern by examining studies achieving state-of-the-art results on the IJB-A [46, 75], exhibiting only baseline CFP performances. Conversely, studies demonstrating proficiency on the CFP [51, 53] or the CMU Multi-PIE [94] tend to yield relatively lower scores on the IJB-A. In particular, only a handful of researchers ventured into cross-dataset evaluations, achieving poor performances. This suggests that algorithms are often tuned to specific datasets, and cross-dataset evaluations continue to pose a challenge. Moreover, in these studies, the analysis regarding pose variation is limited. Therefore, it remains unclear whether the high performance is due to the algorithm's robustness toward extreme poses. In this regard, cross-dataset evaluations can provide excellent insight into the algorithm's pose handling performance.

5.4. Video Datasets. Several video datasets have been published over the years [12], where few datasets stand out for recent pose-robust face recognition studies. One is the YTC dataset, Kim et al. [6] presented in 2008. YTC includes 1, 910 videos of 47 actors and politicians, where pose illumination and expression are largely uncontrolled. Later in 2011, Wolf et al. [7] presented the YTF dataset, which consists of 3, 425 labeled videos of 1, 595 individuals in challenging, uncontrolled conditions.

Table 4: Comparison of classification performances reported on the IJB-A benchmark.

Category	Publication	Method	1 : 1 verification TAR (%)		1 : *N* identification TPIR (%)		1 : *N* identification accuracies (%)	
			@FAR = 0.001	@FAR = 0.01	@FPIR = 0.01	@FPIR = 0.1	Rank-1	Rank-5
Feature based	Klare et al. [5]	GOTS	19.80	40.60	95.30	76.50	44.30	59.50
	Sankaranarayanan et al. [42]	CNN + TPE	81.30	90.00	75.30	86.30	93.20	—
	Wang et al. [44]	DF-COTS	51.00	72.90	38.50	60.80	82.20	93.10
	Lu et al. [36]	Joint Bayesian	81.40	91.30	—	—	—	—
	Yin and Liu [58]	p-CNN	53.90	77.50	—	—	85.80	93.80
	Cao et al. [45]	DREAM	86.80	94.40	—	—	94.60	96.80
	Zheng et al. [48]	Ring loss	93.22	—	—	—	—	—
	Shi and Jain [46]	PFE	**95.25**	**97.50**	—	—	—	—
	Wang and Guo [53]	LS-CNN	75.50	87.50	—	—	—	—
	Zhang et al. [51]	KL-TAL	90.20	95.60	—	—	96.50	98.00
Image based	Crispell et al. [63]	Dual pose jitter	—	—	73.40	86.00	94.40	98.10
	Tran et al. [67]	DR-GAN	53.90	77.40	—	—	85.50	94.70
	Yin et al. [21]	FF-GAN	66.30	85.20	—	—	90.20	95.40
	Zhao et al. [74]	DA-GAN	93.00	97.60	89.00	94.90	97.10	98.90
	Tran et al. [68]	DR-GAN	78.10	87.20	—	—	92.00	96.10
	Zhao et al. [75]	DA-GAN2.0	**97.30**	**98.90**	**93.90**	**98.20**	**99.00**	**99.50**
	Tang et al. [76]	RL-WGAN	57.70	81.60	—	—	89.30	96.50
	Cao et al. [86]	AD-GAN	88.90	94.60	—	—	95.90	97.00
	Shao et al. [91]	WAPNN	77.00	90.20	82.30	—	96.50	98.70
	Zhao et al. [80]	PIM	87.50	93.30	81.50	89.20	94.40	—
	Cao et al. [87]	HF-PIM	89.90	95.30	—	—	96.40	98.10
	Zhang et al. [95]	CCFF-GAN	72.29	84.07	—	—	98.13	98.93
	Shao et al. [84]	DAFF	75.70	90.00	—	—	95.80	98.20
	Xu et al. [94]	FFlowGAN	93.40	96.90	—	—	92.00	95.00
	Taherkhani et al. [97]	PF-cpGAN	91.20	95.80	—	—	97.60	98.80
Set based	Abdalmageed et al. [22]	Deep multipose	—	78.70	87.60	95.40	84.60	92.70
	Masi et al. [126]	PAM (soft-max)	65.20	82.60	—	—	84.00	92.50
	Chen et al. [112]	DCNN	—	83.80	—	—	90.30	96.50
	Masi et al. [125]	Augmented face	72.50	88.60	—	—	90.60	96.20
	Ranjan et al. [121]	CNN + TPE	82.30	92.20	79.20	88.70	94.70	—
	Crosswhite et al. [113]	Template Adaptation	—	93.90	77.40	88.20	92.80	97.70
	Yang et al. [129]	NAN	88.10	94.10	81.70	91.70	95.80	98.00
	Rao et al. [131]	DAN	91.00	94.10	—	—	98.00	99.00
	Zhao et al. [130]	kNN-aver	85.40	94.50	70.80	85.80	93.00	98.00
	Masi et al. [20]	PAM (pooling)	71.10	84.70	—	—	86.20	94.30
	Gong et al. [132]	C-FAN	91.59	93.97	86.88	92.85	94.57	96.27
	Ranjan et al. [122]	DPSSD	95.20	96.90	92.00	96.20	97.50	98.60
	Duan et al. [117]	UniformFace	92.30	96.90	—	—	97.90	98.80
	An et al. [124]	APA	97.50	98.95	94.87	98.41	99.13	99.59
	Zhao et al. [123]	MPNet	**98.40**	**99.10**	**94.10**	**98.90**	**99.40**	—
	Arachchilage and Izauierdo [133]	Clusterface	86.60	94.23	—	—	94.28	97.05
	Liu et al. [134]	DAC	—	96.30	86.50	94.30	97.60	—
	Bai et al. [135]	LFENet	96.83	98.93	—	—	—	—

We report the true accept rates (TAR) at various false accept rates (FAR) for identification and verification and the rank-1 and rank-5 accuracies for identification results. The highest accuracies for each category are highlighted in bold.

Table 5: Comparison of rank-1 identification rates (%) of different methods for video-based face recognition on YTC and YTF.

Category	Publication	Method	Mean accuracies
YTC			
Set based	Wang and Chen [108]	MDA	67.20 ± 4.00
	Hu et al. [101]	SANP	65.03 ± 5.74
	Hu et al. [102]	KSANP	65.46 ± 5.53
	Wang et al. [109]	COV + PLS	70.10
	Hayat et al. [110]	DRM	72.55 ± 4.74
	Lu et al. [107]	D-SFDL	79.50
	Rao et al. [131]	DAN	**97.70 ± 0.72**
	Wang et al. [115]	DARG-Kernel	77.09
	Cheng et al. [116]	DML	82.80 ± 2.60
	Arachchilage and Izauierdo [133]	Clusterface	91.06
YTF			
Feature based	Schroff et al. [40]	FaceNet	95.12 ± 0.39
	Parkhi et al. [39]	VGG-Face	97.30
	Shi and Jain [46]	PFE	97.36
Image based	Taigman et al. [38]	DeepFace	91.40 ± 1.10
	Zhong et al. [72]	e2e	95.00
Set based	Hayat et al. [110]	DRM	51.45 ± 3.06
	Yang et al. [129]	NAN	95.72 ± 0.64
	Shao et al. [118]	VGG-CRF-SME	93.80 ± 1.30
	Rao et al. [131]	DAN	95.01 ± 0.60
	Cheng et al. [116]	DML	97.70
	Gong et al. [132]	C-FAN	96.50 ± 0.90
	Duan et al. [117]	UniformFace	97.70
	Xue et al. [119]	PCPSFM	97.84
	Zhao et al. [123]	MPNet	**99.10**
	Liu et al. [134]	DAC	96.80 ± 0.38

The highest accuracies for each category are highlighted in bold.

In Table 5, we display the outcomes from YTF and YTC. Video datasets are predominantly employed to assess the effectiveness of set-based methods. Given that YTF is a more recent and comprehensive dataset, recent studies tend to favor it over YTC. On YTC, investigations centered on deep aggregation networks, particularly those presented in studies [131, 133] set the state-of-the-art. Rao et al. [131] achieved the highest performance, reaching 97.70%, with their proposed DAN.

The pioneering deep learning algorithms, including DeepFace [38], VGG-Face [39], and FaceNet [40], were initially evaluated on the YTF dataset, elevating the baseline to a mean accuracy of 97.30%. Zhao et al. [123] introduced a multitask learning approach, achieving the highest performance on YTF at 99.10%. Xue et al. [119] presented a feature disentanglement method, securing the second-highest performance at 97.84%.

Another effective technique yielding promising results is deep metric learning, implemented by Cheng et al. [116] and Duan et al. [117], who achieved state-of-the-art outcomes on the YTF. However, when their performances on the YTC are compared, a notable performance disparity emerges between aggregation networks and deep metric learning methods. This observation aligns with our earlier findings, indicating that many methods experience a decline in performance under cross-dataset evaluations. It is worth noting that aggregation networks aim to emphasize informative features among numerous noisy candidates and may not explicitly specify if their systems favor frontal views over extreme poses. Hence, the findings on the YTF dataset provide a more comprehensive insight into the current state-of-the-art, particularly regarding extreme poses in video datasets.

5.5. Discussion. Over the years, a variety of diverse and extensive datasets capturing extreme poses have been curated to evaluate the effectiveness of studies on pose-invariant face recognition. Among these datasets, the CMU Multi-PIE [3] dataset tackles challenges stemming from pose and illumination variations, while the CFP [4] dataset supports research on frontal-to-side-view verification. For the evaluation of set-based methods, various video datasets are available, including YTC [6] and YTF [7]. Additionally, the IJB-A [5] benchmark is tailored for studies focusing on face recognition in uncontrolled settings.

Studies concentrating on face frontalization are primarily tested on CMU Multi-PIE and CFP, as these datasets offer a more controlled setting. Among the proposed methods, feature-based approaches consistently outperform image-based methods. Notably, attention-guided networks [53, 55]

and designing new loss functions [50, 51] stand out for achieving the highest performance. These methods excel by enhancing the more informative aspects of the data or penalizing errors, resulting in more distinctive features. In contrast, many image-based methods focus on realistic reconstruction while treating recognition as a secondary task. It is argued that increasing the pose angles leads to higher reconstruction errors. Image-based methods achieving the best performances often employ assembled networks to fuse comparisons under different views [96] or progressively combine multiple reconstructions under slight rotations to achieve higher rotation [94].

Set-based methods are predominantly evaluated on IJB-A or video datasets, with multitask learning [123] and feature disentanglement [119] emerging as the state-of-the-art techniques. Recently, aggregation networks [131, 133] have gained popularity among set-based methods, demonstrating the highest accuracy on the YTC dataset. Nevertheless, these approaches prioritize extracting informative features from a multitude of noisy candidates and may not explicitly specify whether their systems favor frontal views over extreme poses. Additionally, we observed that only a limited number of researchers have explored cross-dataset evaluations, resulting in suboptimal performances [46, 51, 53, 75, 94, 116, 117]. This highlights the challenge of algorithms being often tailored to specific datasets, and cross-dataset evaluations remain a persistent challenge. Moreover, pose variation analysis in set-based methods is limited, and cross-dataset evaluations can offer insight into how well the algorithm handles pose variations.

6. Conclusion

Recognizing faces in extreme poses, especially up to the side-view, poses a significant challenge in achieving pose-robust face recognition. Side-view poses bring about self-occlusion, resulting in visibility issues where only one of the eyes is visible, particularly in uncontrolled settings where illumination and image quality are unreliable. In this article, we concentrate on side-view face recognition and review methods designed to handle extreme pose variations.

We categorized the research in this area based on pose handling techniques, including feature-based, image-based, and set-based pose handling. Feature-based methods involve pose-robust feature extraction or embedding strategies to find projections that explain the correlation among similar subjects despite pose variation. Conversely, image-based approaches involve reconstructing faces in various views to enable comparisons under similar conditions. Set-based pose handling methods leverage multiple images to describe the subject's characteristics.

Several datasets capturing extreme poses have been curated over time, including the CMU Multi-PIE [3] for pose and illumination variations, the CFP [4] for frontal-to-side-view verification, and various video datasets such as YTC [6] and YTF [7] for evaluating set-based methods. Furthermore, the IJB-A [5] benchmark is specifically crafted for studies centered around face recognition in uncontrolled settings.

In the realm of pose-robust face recognition, comparing frontal to side-view poses has been the primary focus, with feature-based pose alignment techniques defining the state-of-the-art. Attention-guided networks and loss function design are identified as the most effective techniques. GANs dominate image-based methods, tested mainly on the CMU Multi-PIE and CFP due to the controlled setting and considerable pose variation. However, GANs still have room for improvement compared to feature-based methods, especially in the presence of significant pose variations and cluttered backgrounds.

Recently, face recognition in uncontrolled settings has gained popularity, leading to various set-based strategies. These methods are predominantly evaluated on IJB-A or video datasets, with multitask learning [123] and feature disentanglement [119] emerging as state-of-the-art techniques. The quality of images in face datasets studying pose variation significantly influences algorithm performance. Recently, aggregation networks [131, 133] have gained popularity among set-based methods to enhance feature-rich components in image sets among numerous noisy candidates and demonstrate high accuracy on the YTC dataset. However, these methods do not explicitly specify whether their systems favor frontal views over extreme poses.

Matching arbitrary poses in uncontrolled settings remains challenging, particularly in neglected areas like side-view to side-view recognition, where face rotation lacks associations apart from face symmetry. Additionally, a limited number of researchers have explored cross-dataset evaluations, resulting in suboptimal performances [46, 51, 53, 75, 94, 116, 117]. This emphasizes the challenge of algorithms often being tailored to specific datasets, with cross-dataset evaluations remaining a persistent challenge.

Data Availability

Data acquired by the authors in [29] will be made available upon reasonable request. For data from other cited articles, we refer to the authors of those papers.

Conflicts of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This study was funded by the University of Twente, Netherlands.

References

[1] C. Ding and D. Tao, "A comprehensive survey on pose-invariant face recognition," *ACM Transactions on Intelligent Systems and Technology*, vol. 7, no. 3, pp. 1–42, 2016.

[2] T. Hassner, S. Harel, E. Paz, and R. Enbar, "Effective face frontalization in unconstrained images," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4295–4304, IEEE, Boston, MA, USA, 2015.

[3] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-PIE," *Image and Vision Computing*, vol. 28, no. 5, pp. 807–813, 2010.

[4] S. Sengupta, J. C. Chen, C. Castillo, V. M. Patel, R. Chellappa, and D. W. Jacobs, "Frontal to profile face verification in the wild," in *2016 IEEE Winter Conference on Applications of*

Computer Vision, pp. 1–9, IEEE, Lake Placid, United States, 2016.

[5] B. F. Klare, B. Klein, E. Taborsky et al., "Pushing the frontiers of unconstrained face detection and recognition: IARPA janus benchmark A," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1931–1939, IEEE, Boston, MA, USA, 2015.

[6] M. Kim, S. Kumar, V. Pavlovic, and H. Rowley, "Face tracking and recognition with visual constraints in real-world videos," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, IEEE, Anchorage, AK, USA, 2008.

[7] L. Wolf, T. Hassner, and I. Maoz, "Face recognition in unconstrained videos with matched background similarity," in *CVPR 2011*, pp. 529–534, Colorado Springs, CO, USA, 2011.

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

[9] X. Zhang and Y. Gao, "Face recognition across pose: a review," *Pattern Recognition*, vol. 42, no. 11, pp. 2876–2896, 2009.

[10] X. Wang, J. Peng, S. Zhang, B. Chen, Y. Wang, and Y. Guo, "A survey of face recognition," ArXiv:221213038, pp. 1–59, 2022.

[11] M. Wang and W. Deng, "Deep face recognition: a survey," *Neurocomputing*, vol. 429, pp. 215–244, 2021.

[12] M. Taskiran, N. Kahraman, and C. E. Erdem, "Face recognition past, present and future (a review)," *Digital Signal Processing*, vol. 106, Article ID 102809, 2020.

[13] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The FERET evaluation methodology for face-recognition algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 10, pp. 1090–1104, 2000.

[14] P. J. Phillips, P. J. Grother, R. J. Micheals, D. M. Blackburn, E. Tabassi, and M. Bone, "Face recognition vendor test 2002," IEEE Int SOI Conference, Article ID 44, 2003.

[15] K. Messer, J. Kittler, M. Sadeghi et al., "Face verification competition on the XM2VTS database," in *Audio- and Video-Based Biometric Person Authentication. AVBPA 2003*, vol. 2688 of *Lecture Notes in Computer Science*, pp. 964–974, Springer, Berlin, Heidelberg, 2003.

[16] T. Sim, S. Baker, and M. Bsat, "The CMU pose, illumination, and expression database," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 12, pp. 1615–1618, 2003.

[17] J. R. Barr, K. W. Bowyer, P. J. Flynn, and S. Biswas, "Face recognition from video: a review," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 26, no. 5, Article ID 1266002, 2012.

[18] G. B. Huang, M. Mattar, T. Berg, and E. Learned.Miller, "Labeled faces in the wild: a database for studying face recognition in unconstrained environments," in *Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition*, pp. 1–14, HAL, Images, 2008.

[19] Q. Qiu and R. Chellappa, "Compositional dictionaries for domain adaptive face recognition," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 5152–5165, 2015.

[20] I. Masi, F.-J. Chang, J. Choi et al., "Learning pose-aware models for pose-invariant face recognition in the wild," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 379–393, 2019.

[21] X. Yin, X. Yu, K. Sohn, X. Liu, and M. Chandraker, "Towards large-pose face frontalization in the wild," in *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 4010–4029, IEEE, Venice, Italy, 2017.

[22] W. AbdAlmageed, Y. Wu, S. Rawls et al., "Face recognition using deep multi-pose representations," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1–9, IEEE, Lake Placid, NY, USA, 2016.

[23] C. Ding, C. Xu, and D. Tao, "Multi-task pose-invariant face recognition," *IEEE Transactions on Image Processing*, vol. 24, no. 3, pp. 980–993, 2015.

[24] S. Ding, Q. Zhai, Y. F. Zheng, and D. Xuan, "Side-view face authentication based on wavelet and random forest with subsets," in *2013 IEEE International Conference on Intelligence and Security Informatics*, pp. 76–81, IEEE, Seattle, WA, USA, 2013.

[25] X. Zhou and B. Bhanu, "Human recognition based on face profiles in video," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)-Workshops*, pp. 1–8, IEEE, San Diego, CA, USA, 2005.

[26] G. Pan, L. Zheng, and Z. Wu, "Robust metric and alignment for profile-based face recognition an experimental comparison," in *2005 Seventh IEEE Workshops on Applications of Computer Vision (WACV/MOTION'05)-Volume 1*, pp. 117–122, IEEE, Breckenridge, CO, USA, 2005.

[27] W. Hong, Z. Ping, and G. Zhen-Guo, "An algorithm for automatic side face portrait recognition based on fourier descriptor," in *2008 International Symposium on Information Science and Engineering*, pp. 767–772, IEEE, Shanghai, China, 2008.

[28] R. Rathore, S. Prakash, and P. Gupta, "Efficient human recognition system using ear and profile face," in *2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, pp. 1–6, IEEE, Arlington, VA, USA, 2013.

[29] P. Santemiz, L. J. Spreeuwers, and R. N. J. Veldhuis, "Automatic face recognition for home safety using video-based side-view face images," *IET Biometrics*, vol. 7, no. 6, pp. 606–614, 2018.

[30] C. Liu and H. Wechsler, "Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition," *IEEE Transactions on Image Processing*, vol. 11, no. 4, pp. 467–476, 2002.

[31] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the Seventh IEEE International Conference on Computer Vision*, vol. 2, pp. 1150–1157, IEEE, Kerkyra, Greece, 1999.

[32] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (SURF)," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[33] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, pp. 886–893, IEEE, San Diego, CA, USA, 2005.

[34] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[35] C. Ding, J. Choi, D. Tao, and L. S. Davis, "Multi-directional multi-level dual-cross patterns for robust face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 3, pp. 518–531, 2016.

[36] B. Lu, J. Zheng, J. C. Chen, and R. Chellappa, "Pose-robust face verification by exploiting competing tasks," in *Proceedings - 2017 IEEE Winter Conference on Applications of Computer Vision*, pp. 1124–1132, IEEE, Santa Rosa, United States, 2017.

[37] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1891–1898, IEEE, Columbus, OH, USA, 2014.
[38] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "DeepFace: closing the gap to human-level performance in face verification," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1701–1708, IEEE, Columbus, OH, USA, 2014.
[39] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," Proc of British Machine Vision Conf, pp. 1–12, 2015.
[40] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: a unified embedding for face recognition and clustering," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 815–823, IEEE, Boston, MA, USA, 2015.
[41] Y. Zhang, M. Shao, E. K. Wong, and Y. Fu, "Random faces guided sparse many-to-one encoder for pose-invariant face recognition," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2416–2423, IEEE, 2013.
[42] S. Sankaranarayanan, A. Alavi, C. D. Castillo, and R. Chellappa, "Triplet probabilistic embedding for face verification and clustering," in *2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pp. 1–8, IEEE, Niagara Falls, NY, USA, 2016.
[43] J.-C. Chen, J. Zheng, V. M. Patel, and R. Chellappa, "Fisher vector encoded deep convolutional features for unconstrained face verification," in *2016 IEEE International Conference on Image Processing (ICIP)*, pp. 2981–2985, IEEE, Phoenix, AZ, USA, 2016.
[44] D. Wang, C. Otto, and A. K. Jain, "Face search at scale," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1122–1136, 2017.
[45] K. Cao, Y. Rong, C. Li, X. Tang, and C. C. Loy, "Pose-robust face recognition via deep residual equivariant mapping," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5187–5196, IEEE, Salt Lake City, UT, USA, 2018.
[46] Y. Shi and Jain, "Probabilistic face embeddings," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6901–6910, IEEE, Seoul, Korea (South), 2019.
[47] Q. Rao, B. Yu, Y. Yang, and B. Feng, "Knot magnify loss for face recognition," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, pp. 2396–2400, IEEE, Athens, Greece, 2018.
[48] Y. Zheng, D. K. Pal, and M. Savvides, "Ring loss: convex feature normalization for face recognition," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5089–5097, IEEE, Salt Lake City, UT, USA, 2018.
[49] M. He, J. Zhang, S. Shan, M. Kan, and X. Chen, "Deformable face net: learning pose invariant feature with pose aware feature alignment for face recognition," in *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pp. 1–8, IEEE, Lille, France, 2019.
[50] Y. Zhang, K. Fu, C. Han, P. Cheng, S. Yang, and X. Yang, "PGM-face: Pose-guided margin loss for cross-pose face recognition," *Neurocomputing*, vol. 460, pp. 154–165, 2021.
[51] Z. Zhang, Y. Chen, W. Yang, G. Wang, and Q. Liao, "Triplet angular loss for pose-robust face recognition," in *2021 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, IEEE, Shenzhen, China, 2021.
[52] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," arXiv, 2016, https://arxiv.org/abs/1409.0473.
[53] Q. Wang and G. Guo, "LS-CNN: Characterizing local patches at multiple scales for face recognition," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 1640–1653, 2019.
[54] H. He, W. Zhou, and X. Liu, "PCSAFace: face recognition with parallel channel and spatial attention," in *2021 IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, pp. 930–933, IEEE, Chongqing, China, 2021.
[55] J. Huang and C. Ding, "Attention-guided progressive mapping for profile face recognition," in *2021 IEEE International Joint Conference on Biometrics (IJCB)*, pp. 1–8, IEEE, Shenzhen, China, 2021.
[56] C. Xiong, X. Zhao, D. Tang, K. Jayashree, S. Yan, and T. Kim, "Conditional convolutional neural network for modality-aware face recognition," in *2015 International Conference on Computer Vision*, pp. 3667–3675, IEEE, Chile, Santiago, 2015.
[57] X. Peng, X. Yu, K. Sohn, D. N. Metaxas, and M. Chandraker, "Reconstruction-based disentanglement for pose-invariant face recognition," in *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 1632–1641, IEEE, Venice, Italy, 2017.
[58] X. Yin and X. Liu, "Multi-task convolutional neural network for pose-invariant face recognition," *IEEE Transactions on Image Processing*, vol. 27, no. 2, pp. 964–975, 2018.
[59] P. Naveen and P. Sivakumar, "Adaptive morphological and bilateral filtering with ensemble convolutional neural network for pose-invariant face recognition," *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, no. 11, pp. 10023–10033, 2021.
[60] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3D faces," in *SIGGRAPH '99: Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, pp. 187–194, ACM, 1999.
[61] X. Zhang, Y. Gao, and M. K. H. Leung, "Recognizing rotated faces from frontal and side views: an approach toward effective use of mugshot databases," *IEEE Transactions on Information Forensics and Security*, vol. 3, no. 4, pp. 684–697, 2008.
[62] A. Moeini and H. Moeini, "Real-world and rapid face recognition toward pose and expression variations via feature library matrix," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 5, pp. 969–984, 2015.
[63] D. Crispell, O. Biris, N. Crosswhite, J. Byrne, and J. L. Mundy, "Dataset augmentation for pose and lighting invariant face recognition," arXiv:170404326, pp. 1–9, 2017.
[64] L. Hu, M. Kan, S. Shan, K. Song, and X. Chen, "LDF-Net learning a displacement field network for face recognition across pose," in *2017 IEEE 12th International Conference on Automatic Face & Gesture Recognition*, pp. 9–16, IEEE, 2017.
[65] J. Liang, F. Liu, H. Tu, Q. Zhao, and A. K. Jain, "On mugshot-based arbitrary view face recognition," in *2018 24th International Conference on Pattern Recognition (ICPR)*, pp. 3126–3131, IEEE, Beijing, China, 2018.
[66] I. Goodfellow, J. Pouget-Abadie, M. Mirza et al., "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.
[67] L. Tran, X. Yin, and X. Liu, "Disentangled representation learning GAN for pose-invariant face recognition," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1283–1292, IEEE, Honolulu, HI, USA, 2017.
[68] L. Q. Tran, X. Yin, and X. Liu, "Representation learning by rotating your faces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 12, pp. 3007–3021, 2018.

[69] C. Rong, X. Zhang, and Y. Lin, "Feature-improving generative adversarial network for face frontalization," *IEEE Access*, vol. 8, pp. 68842–68851, 2020.
[70] C. Fu, Y. Hu, X. Wu, G. Wang, and R. Q. H. Zhang, "High-fidelity face manipulation with extreme poses and expressions," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 2218–2231, 2021.
[71] R. T. Marriott, S. Romdhani, and L. Chen, "A 3D GAN for improved large-pose facial recognition," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13440–13450, IEEE, Nashville, TN, USA, 2021.
[72] Y. Zhong, J. Chen, and B. Huang, "Toward end-to-end face recognition through alignment learning," *IEEE Signal Processing Letters*, vol. 24, no. 8, pp. 1213–1217, 2017.
[73] G. Wang, J. Ma, Q. Zhang, J. Lu, and J. Zhou, "Pseudo facial generation with extreme poses for face recognition," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1994–2003, IEEE, Nashville, TN, USA, 2021.
[74] J. Zhao, L. Xiong, P. K. Jayashree et al., "Dual-agent gans for photorealistic and identity preserving profile face synthesis," in *31st Conference on Neural Information Processing Systems (NIPS 2017)*, pp. 66–76, Curran Associates, Inc., Long Beach, CA, USA, 2017.
[75] J. Zhao, L. Xiong, J. Li, J. Xing, S. Yan, and J. Feng, "3D-aided dual-agent gans for unconstrained face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, pp. 2380–2394, 2018.
[76] C.-H. Tang, G.-S. J. Hsu, and M. H. Yap, "Face recognition with disentangled facial representation learning and data augmentation," in *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 1670–1674, IEEE, Taipei, Taiwan, Int Conf on Image Processing, 2019.
[77] R. Huang, S. Zhang, T. Li, and R. He, "Beyond face rotation: global and local perception GAN for photorealistic and identity preserving frontal view synthesis," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2458–2467, IEEE, 2017.
[78] Y. Hu, X. Wu, B. Yu, R. He, and Z. Sun, "Pose-guided photorealistic face rotation," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8398–8406, IEEE, Salt Lake City, UT, USA, 2018.
[79] J. Zhao, Y. Cheng, Y. Xu et al., "Towards pose invariant face recognition in the wild," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2207–2216, IEEE, Salt Lake City, UT, USA, 2018.
[80] J. Zhao, J. Xing, L. Xiong, S. Yan, and J. Feng, "Recognizing profile faces by imagining frontal view," *International Journal of Computer Vision*, vol. 128, no. 2, pp. 460–478, 2020.
[81] M. Luo, J. Cao, X. Ma, X. Zhang, and R. He, "FA-GAN: face augmentation gan for deformation-invariant face recognition," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 2341–2355, 2021.
[82] Y. Yin, S. Jiang, J. P. Robinson, and Y. Fu, "Dual-attention GAN for large-pose face frontalization," in *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pp. 249–256, IEEE, Buenos Aires, Argentina, 2020.
[83] X. Luan, H. Geng, L. Liu, W. Li, Y. Zhao, and M. Ren, "Geometry structure preserving based gan for multi-pose face frontalization and recognition," *IEEE Access*, vol. 8, pp. 104676–104687, 2020.
[84] X. Shao, J. Xing, R. Pan, Z. Li, X. Zhou, and Y. Shi, "Multi-view face recognition using deep attention-based face frontalization," in *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, IEEE, Shenzhen, China, 2021.
[85] J. Deng, S. Cheng, N. Xue, Y. Zhou, and S. Zafeiriou, "UV-GAN: adversarial facial UV map completion for pose-invariant face recognition," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7093–7102, IEEE, Salt Lake City, UT, USA, 2018.
[86] J. Cao, Y. Hu, B. Yu, R. He, and Z. Sun, "3D aided duet GANs for multi-view face image synthesis," *IEEE Transactions on Information Forensics and Security*, vol. 14, pp. 2028–2042, 2019.
[87] J. Cao, Y. Hu, H. Zhang, R. He, and Z. Sun, "Towards high fidelity face frontalization in the wild," *International Journal of Computer Vision*, vol. 128, no. 5, pp. 1485–1504, 2020.
[88] X. Xu, K. Li, C. Xu, and S. He, "GDFace: gated deformation for multi-view face image synthesis," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 7, pp. 12532–12540, 2020.
[89] X. Tu, J. Zhao, Q. Liu et al., "Joint face image restoration and frontalization for recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 3, pp. 1285–1298, 2022.
[90] X. Zeng, Z. Wu, X. Peng, and Y. Qiao, "Joint 3D facial shape reconstruction and texture completion from a single image," *Computational Visual Media*, vol. 8, no. 2, pp. 239–256, 2022.
[91] X. Shao, X. Zhou, Z. Li, and Y. Shi, "Multi-view face recognition via well-advised pose normalization network," *IEEE Access*, vol. 8, pp. 66400–66410, 2020.
[92] J. Liu, Q. Li, M. Liu, and T. Wei, "CP-GAN: a cross-pose profile face frontalization boosting pose-invariant face recognition," *IEEE Access*, vol. 8, pp. 198659–198667, 2020.
[93] L. Liu, L. Zhang, and J. Chen, "Progressive pose normalization generative adversarial network for frontal face synthesis and face recognition under large pose," in *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 4434–4438, IEEE, Taipei, Taiwan, 2019.
[94] Y. Xu, X. Xu, J. Jiao, K. Li, C. Xu, and S. He, "Multi-view face synthesis via progressive face flow," *IEEE Transactions on Image Processing*, vol. 30, pp. 6024–6035, 2021.
[95] Z. Zhang, R. Liang, X. Chen et al., "Semisupervised face frontalization in the wild," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 909–922, 2021.
[96] Y. Zhang, K. Fu, C. Han, and P. Cheng, "Identity-and-pose-guided generative adversarial network for face rotation," *Neurocomputing*, vol. 450, pp. 33–47, 2021.
[97] F. Taherkhani, V. Talreja, J. Dawson, M. C. Valenti, and N. M. Nasrabadi, "Profile to frontal face recognition in the wild using coupled conditional generative adversarial network," *IET Biometrics*, vol. 11, no. 3, pp. 260–276, 2022.
[98] Y. Xu, Z. Li, J. Yang, and D. Zhang, "A survey of dictionary learning algorithms for face recognition," *IEEE Access*, vol. 5, pp. 8502–8514, 2017.
[99] F. Schroff, T. Treibitz, D. Kriegman, and S. Belongie, "Pose, illumination and expression invariant pairwise face-similarity measure via Doppelgänger list comparison," in *2011 International Conference on Computer Vision*, pp. 2494–2501, IEEE, Barcelona, Spain, 2011.
[100] H. Zhang, N. M. Nasrabadi, Y. Zhang, and T. S. Huang, "Joint dynamic sparse representation for multi-view face recognition," *Pattern Recognition*, vol. 45, no. 4, pp. 1290–1298, 2012.
[101] Y. Hu, A. S. Mian, and R. Owens, "Sparse approximated nearest points for image set classification," in *CVPR 2011*, pp. 121–128, IEEE, Colorado Springs, CO, USA, 2011.

[102] Y. Hu, A. S. Mian, and R. Owens, "Face recognition using sparse approximated nearest points between image sets," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 10, pp. 1992–2004, 2012.
[103] H. S. Bhatt, R. Singh, and M. Vatsa, "On recognizing faces in videos using clustering-based re-ranking and fusion," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 7, pp. 1056–1068, 2014.
[104] Y.-C. Chen, V. M. Patel, P. J. Phillips, and R. Chellappa, "Dictionary-based face and person recognition from unconstrained video," *IEEE Access*, vol. 3, pp. 1783–1798, 2015.
[105] S. Bahrampour, N. M. Nasrabadi, A. Ray, and W. K. Jenkins, "Multimodal task-driven dictionary learning for image classification," *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 24–38, 2016.
[106] F. Hajati, M. Tavakolian, S. Gheisari, Y. Gao, and A. S. Mian, "Dynamic texture comparison using derivative sparse representation: application to video-based face recognition," *IEEE Transactions on Human-Machine Systems*, vol. 47, no. 6, pp. 970–982, 2017.
[107] J. Lu, G. Wang, and J. Zhou, "Simultaneous feature and dictionary learning for image set based face recognition," *IEEE Transactions on Image Processing*, vol. 26, no. 8, pp. 4042–4054, 2017.
[108] R. Wang and X. Chen, "Manifold discriminant analysis," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 429–436, IEEE, Miami, FL, USA, 2009.
[109] R. Wang, H. Guo, L. S. Davis, and Q. Dai, "Covariance discriminative learning: a natural and efficient approach to image set classification," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2496–2503, IEEE, Providence, RI, USA, 2012.
[110] M. Hayat, M. Bennamoun, and S. An, "Deep reconstruction models for image set classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 4, pp. 713–727, 2015.
[111] M. Haghighat, M. Abdel-Mottaleb, and W. Alhalabi, "Discriminant correlation analysis: real-time feature level fusion for multimodal biometric recognition," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 9, pp. 1984–1996, 2016.
[112] J. Chen, V. M. Patel, and R. Chellappa, "Unconstrained face verification using deep CNN features," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1–9, IEEE, Lake Placid, NY, USA, 2016.
[113] N. Crosswhite, J. Byrne, C. Stauffer, O. Parkhi, Q. Cao, and A. Zisserman, "Template adaptation for face verification and identification," *Image and Vision Computing*, vol. 79, pp. 35–48, 2018.
[114] N. Bodla, J. Zheng, H. Xu, J. Chen, C. Castillo, and R. Chellappa, "Deep heterogeneous feature fusion for template-based face recognition," in *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 586–595, IEEE, Santa Rosa, CA, USA, 2017.
[115] W. Wang, R. Wang, Z. Huang, S. Shan, and X. Chen, "Discriminant analysis on riemannian manifold of gaussian distributions for face recognition with image sets," *IEEE Transactions on Image Processing*, vol. 27, pp. 151–163, 2017.
[116] G. Cheng, P. Zhou, and J. Han, "Duplex metric learning for image set classification," *IEEE Transactions on Image Processing*, vol. 27, no. 1, pp. 281–292, 2018.
[117] Y. Duan, J. Lu, and J. Zhou, "Uniformface: learning deep equidistributed representation for face recognition," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3410–3419, IEEE, Long Beach, CA, USA, 2019.
[118] M. Shao, Y. Zhang, and Y. Fu, "Collaborative random faces-guided encoders for pose-invariant face representation learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 4, pp. 1019–1032, 2018.
[119] N. Xue, J. Deng, S. Cheng, Y. Panagakis, and S. Zafeiriou, "Side information for face completion: a robust PCA approach," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, pp. 2349–2364, 2019.
[120] R. Ranjan, V. M. Patel, and R. Chellappa, "Hyperface: a deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 1, pp. 121–135, 2019.
[121] R. Ranjan, S. Sankaranarayanan, C. D. Castillo, and R. Chellappa, "An all-in-one convolutional neural network for face analysis," in *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pp. 17–24, IEEE, Washington, DC, USA, 2017.
[122] R. Ranjan, A. Bansal, J. Zheng et al., "A fast and accurate system for face detection, identification, and verification," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 1, no. 2, pp. 82–96, 2019.
[123] J. Zhao, J. Li, X. Tu et al., "Multi-prototype networks for unconstrained set-based face recognition," in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pp. 4397–4403, IJCAI, 2019.
[124] Z. An, W. Deng, J. Hu, Y. Zhong, and Y. Zhao, "APA: Adaptive pose alignment for pose-invariant face recognition," *IEEE Access*, vol. 7, pp. 14653–14670, 2019.
[125] I. Masi, A. Tran, T. Hassner, J. T. Leksut, and G. Medioni, "Do we really need to collect millions of faces for effective face recognition?" in *Computer Vision–ECCV 2016*, vol. 9909 of *Lecture Notes in Computer Science*, pp. 579–596, Springer, Cham, 2016.
[126] I. Masi, S. Rawls, G. Medioni, and P. Natarajan, "Pose-aware face recognition in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4838–4846, IEEE, 2016.
[127] A. R. Chowdhury, T.-Y. Lin, S. Maji, and E. Learned-Miller, "One-to-many face recognition with bilinear CNNs," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1–9, IEEE, Lake Placid, NY, USA, 2016.
[128] G. Goswami, M. Vatsa, and R. Singh, "Face verification via learned representation on feature-rich video frames," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 7, pp. 1686–1698, 2017.
[129] J. Yang, P. Ren, D. Zhang et al., "Neural aggregation network for video face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5216–5225, IEEE, 2017.
[130] J. Zhao, J. Han, and L. Shao, "Unconstrained face recognition using a set-to-set distance measure on deep learned features," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 10, pp. 2679–2689, 2018.
[131] Y. Rao, J. Lu, and J. Zhou, "Learning discriminative aggregation network for video-based face recognition and person re-identification," *International Journal of Computer Vision*, vol. 127, no. 6-7, pp. 701–718, 2019.
[132] S. Gong, Y. Shi, N. Kalka, and A. Jain, "Video face recognition: component-wise feature aggregation network

(c-fan)," in *2019 International Conference on Biometrics (ICB)*, pp. 1–8, IEEE, Crete, Greece, 2019.
[133] S. W. Arachchilage and E. Izauierdo, "Clusterface: joint clustering and classification for set-based face recognition," in *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 1781–1787, IEEE, Milan, Italy, 2021.
[134] X. Liu, Z. Guo, J. You, and B. V. K. V. Kumar, "Dependency-aware attention control for image set-based face recognition," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 1501–1512, 2019.
[135] Z. Bai, R. Wang, S. Shan, and X. Chen, "Local feature enhancement network for set-based face recognition," in *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pp. 1–8, IEEE, Jodhpur, India, 2021.
[136] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," ArXiv:14117923, pp. 1–10, 2014.
[137] C. Whitelam, E. Taborsky, A. Blanton et al., "IARPA janus benchmark-B face dataset," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 592–600, IEEE, Honolulu, HI, USA, 2017.
[138] B. Maze, J. Adams, J. A. Duncan et al., "IARPA janus benchmark-C: face dataset and protocol," in *2018 International Conference on Biometrics (ICB)*, pp. 158–165, IEEE, Gold Coast, QLD, Australia, 2018.
[139] D. Pal, C. Bhagavatula, Y. Zheng, R. Tao, and M. Savvides, "Is pose really solved? a frontalization study on off-angle face matching," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 2058–2067, IEEE, Waikoloa, HI, USA, 2019.