LARNeXt: End-to-End Lie Algebra Residual Network for Face Recognition

Xiaolong Yang, Xiaohong Jia, Dihong Gong, Dong-Ming Yan, *Member, IEEE*, Zhifeng Li, *Senior Member, IEEE*, and Wei Liu, *Fellow, IEEE*

***Abstract*— Face recognition has always been courted in computer vision and is especially amenable to situations with significant variations between frontal and profile faces. Traditional techniques make great strides either by synthesizing frontal faces from sizable datasets or by empirical pose invariant learning. In this paper, we propose a completely integrated embedded end-to-end Lie algebra residual architecture (LARNeXt) to achieve pose robust face recognition. First, we explore how the face rotation in the 3D space affects the deep feature generation process of convolutional neural networks (CNNs), and prove that face rotation in the image space is equivalent to an additive residual component in the feature space of CNNs, which is determined solely by the rotation. Second, on the basis of this theoretical finding, we further design three critical subnets to leverage a soft regression subnet with novel multi-fusion attention feature aggregation for efficient pose estimation, a residual subnet for decoding rotation information from input face images, and a gating subnet to learn rotation magnitude for controlling the strength of the residual component that contributes to the feature learning process. Finally, we conduct a large number of ablation experiments, and our quantitative and visualization results both corroborate the credibility of our theory and corresponding network designs. Our comprehensive experimental evaluations on frontal-profile face datasets, general unconstrained face recognition datasets, and industrial-grade tasks demonstrate that our method consistently outperforms the state-of-the-art ones.**

***Index Terms*—Face recognition, lie algebra, pose estimation, profile face.**

I. Introduction

RECENT face recognition technologies have benefited from various datasets and have been extensively developed by further polishing deep learning models [1], [2], [3]. Although many existing technologies are strong and robust to face recognition in unconstrained environments, there remain quite a lot of challenges for recognizing faces varying from different age levels [4], [5], [6], [7], [8], different modalities [9], [10], [11], [12], [13], different poses [14], [15], [16], [17], and occlusions [18], [19]. In this paper, we develop a robust recognition algorithm to address the challenges in general face recognition with a particular effect on matching faces across different poses (e.g., frontal versus profile). Datasets play an essential role in tackling this problem because the generalization ability of a certain deep model is closely related to the size of the training data. Therefore, given an uneven and insufficient distribution of frontal and profile face images, the in-depth features tend to focus on frontal faces, and the learning results are exclusively biased incomplete statistics. Some pioneering work has examined this problem and reconstructed more datasets using different data augmentation methods. A typical approach is to enrich input sources either by synthesizing profile faces with appearance variations [20] or by treating a set of images as one image input [21], to eliminate the need for profile data. Another method combines more information, including multi-task learning [16], [17], [22] and template adaptation [23], [24]. Specifically, multi-task learning focuses on pose-aware targets combined with richer information, such as illumination, expression, gender, and age, to comprehensively boost recognition performance. Meanwhile, the methods based on template adaptation learning usually create a mean 3D model face. Employing migration and mapping, this method avoids processing the 3D transformation at the image level. Nevertheless, these strategies tend to increase the unnecessary computational burden. Some other approaches use profile faces to synthesize frontal faces to avoid large pose variations [25], [26], [27], [28]. However, these methods suffer from artifacts caused by occlusions and non-rigid expressions.

The studies mentioned above primarily rely on additional data sources or labels. We seek to address this problem fundamentally and efficiently by clarifying the inner relationship among different poses of samples. A recent approach called Deep Residual Equivalent Mapping (DREAM) [15] has further discussed the gap between the features of frontal-profile pairs simply by approximating the difference using a deep learning model. To a certain extent, this approach explores the gap in

Manuscript received 12 November 2021; revised 25 April 2023; accepted 14 May 2023. Date of publication 2 June 2023; date of current version 5 September 2023. This work was supported in part by National Key Research and Development Program under Grant 2019YFB2204104, in part by the CAS Project for Young Scientists in Basic Research under Grant YSBR-034, in part by the National Natural Science Foundation of China under Grants 12022117 and 62172415, and in part by Tencent AI Lab Rhino-Bird Focused Research Program under Grant JR202127. Recommended for acceptance by H. Ling. *(Corresponding authors: Xiaohong Jia, Zhifeng Li; Wei Liu.)*

Xiaolong Yang and Xiaohong Jia are with the Academy of Mathematics and Systems Science of the Chinese Academy of Sciences, University of Chinese Academy of Sciences, Huairou 101408, China (e-mail: yangxiaolong17@mails.ucas.ac.cn; xhjia@amss.ac.cn).

Dihong Gong, Zhifeng Li, and Wei Liu are with Tencent Data Platform, ShenZhen 518054, China (e-mail: gongdihong@gmail.com; michaelzfli@tencent.com; wl2223@columbia.edu).

Dong-Ming Yan is with the State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS) and NLPR, Institute of Automation of the Chinese Academy of Sciences, and School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 101408, China (e-mail: yandongming@gmail.com).

Our code and model are made publicly available at https://github.com/paradocx/LARNet.

Digital Object Identifier 10.1109/TPAMI.2023.3279378

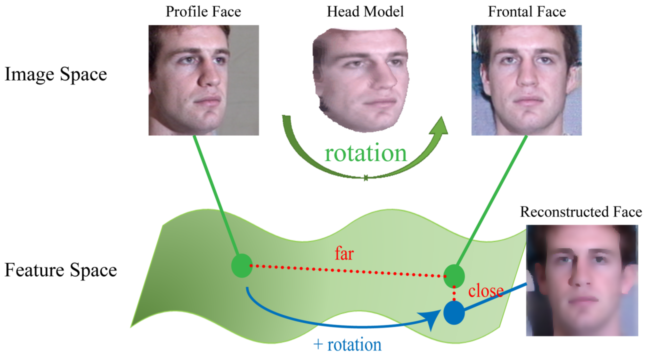


Fig. 1. *Frontalization* or *face rotation* in the feature space. Given that the frontal-profile pair is generated by head rotation, we prove that the face rotation in the image space is equivalent to an additive residual component in the deep feature space. To show the equivalence, we reconstruct the image corresponding to the modified feature (blue dot) and provide the visual result of the frontal face.

a way similar to the generative adversarial network (GAN), which makes the target sample (frontal face feature) and generated sample (profile face feature) as close as possible through encoding and decoding. Although mostly following empirical observations, this approach offers incentives to delve into the essence of deep features.

Frontal-profile pairs are generated by head rotations, which should not be ignored in profile face recognition. However, rotation matrices cannot be easily embedded in CNNs because the group of rotation matrices is closed under multiplication but not under addition. In contrast, the addition operation frequently appears in all gradient descent calculations. Benefiting from the pose estimation work [29] in the field of *simultaneous localization and mapping* (SLAM), we introduce Lie algebra to update the rotation matrices in CNNs.

In this paper, we prove that for each frontal-profile pair linked by a rotation, their corresponding deep features also preserve a complementary rotation relationship by Lie algebra. To the best of our knowledge, this study makes the first attempt to theoretically explore and explain the potential connection between the features of a frontal face and its profile counterpart. Moreover, to facilitate the numerical calculation, we prove that the face rotation in the image space is equivalent to an additive residual component in the feature space in CNNs. Based on this theoretical result, we propose the *end-to-end Lie algebra residual network* (LARNeXt), which achieves face frontalization or rotation-and-render in the deep feature space, as shown in Fig. 1. LARNeXt has three critical subnets for leveraging what we dig out, namely, a soft regression subnet with novel multi-fusion attention feature aggregation for efficient pose estimation, a residual subnet for decoding rotation information from input face images, and a gating subnet to learn rotation magnitude for controlling the strength of the residual component contributing to the feature learning process. We have performed comparative experiments with more than 30 solutions in the recent five years under various evaluation criteria and metrics and found that our method outperforms representative state-of-the-art competitors. Massive ablation experiments and abundant visualization results can also corroborate the credibility of our theory and corresponding network designs. This paper extends our recent conference publication LARNet [30] of ICML 2021 and addresses unsolved issues both theoretically and experimentally. The significant improvements we made are listed as follows:

1) We polish the network architecture design so that our LARNeXt follows an end-to-end mechanism, by adding a head estimation subnet rather than requiring prior pose labels, in which a novel attention feature aggregation strategy is proposed for efficient and high-accuracy pose estimation.
2) We refine our mathematical formulation model in interpretability and comprehensibility, including the embedding of Lie Algebra, a multi-fusion attention map, and the geometrically appropriate ℓ_2 norm instead of the rough ℓ_∞ approximation.
3) We implement the visualization results of face reconstruction to further elaborate on the advantages of our method's feature representation, explore the proposed model's scalability on the 16 million industrial-grade dataset, and conduct a detailed failure case analysis with a future potential improvement.

II. Related Work

We briefly discuss previous work on profile face and large pose face recognition. Besides, we also introduce some mathematical background knowledge related to Lie algebra for interested readers.

Insufficient Dataset: Many methods try to solve the profile face recognition problem by avoiding the unevenness of datasets. Masi et al. [20] proposed domain-specific data augmentation, which is a more accessible method of increasing the size of training data for face recognition systems and focused on crucial facial appearance variations. Meanwhile, the multicolumn network [21] and neural aggregation network (NAN) [31] try to use additional information, such as a set of images or videos, as input to address the potential shortcomings of a single image. Despite showing some progress, these methods still have their limitations. Specifically, they tend to falsely match the profile faces of different identities and miss the frontal and profile faces of the same identity.

Pose Variation: Many methods have been developed for large poses. For instance, template-adaptation-based studies [23], [24] have mainly conducted transfer learning using a constructed classifier and synthesizer and performed pooling based on image quality and head pose. As opposed to those techniques which expect to learn pose invariance, pose-aware deep learning methods [16], [17] use multiple pose-specific models and render face images to reduce sensitivity to pose variations. Other studies use more additional labels instead of only poses. Multi-task learning (MTL) is a widely used method that involves pose, illumination, and expression estimations. Yin et al. [27] proposed a pose-directed multi-task CNN and balanced between different tasks. DebFace [28] (de-biasing adversarial network) takes gender, age, and race into consideration and minimizes the correlation among feature factors to reduce the bias influence from other factors. Although effective, these methods yield high computational costs due to their use of multiple models and tasks, and the accuracy of their results cannot meet higher requirements.

Frontalization: Given the challenges associated with profiles and large poses, some methods directly use existing datasets to synthesize the frontal face and perform face recognition. Due to the widespread use of GAN, FF-GAN [22] and DR-GAN [25] have been applied and have outperformed many of their competitors through their disentangled encoder-decoder structure for learning a generative and discriminative representation. With the rapid progress of 3D face reconstruction technology, research interest in the projecting rendering of the frontal face after reconstruction has also increased. Rotate-and-Render [28] is a representative method for single-view images, which leverages the recent advances in 3D face modeling and high-resolution GAN to constitute building blocks given that the 3D rotation-and-render of faces can be applied to arbitrary angles without losing details. While reconstruction and synthesis only improve visualization performance, they yield a relatively poor feature representation as validated based on their performance on face recognition tasks.

Feature Space: Some studies have considered the latent features than the image itself. Shi et al. [32] proposed *probabilistic face embeddings* (PFEs), which represent each face image as a Gaussian distribution in the latent space. Meanwhile, feature transfer learning [33] makes the under-represented distribution closer to the regular distribution. These approaches attempt to make the sample distribution tend to a Gaussian prior. Despite paying attention to features, specific datasets or face recognition techniques in real-world applications cannot guarantee that the samples follow a Gaussian distribution; therefore, their intuition and persuasiveness are generally below users' expectations. Another representative work, DREAM [15], uses the residual network to directly modify the features of the profile face to the frontal one, which is similar to our proposed method. DREAM roughly bridges these features through mapping from deep learning, and due to the lack of in-depth analysis of the potential physical relationship between the frontal and profile faces, the ad-hoc designed results of DREAM reach a bottleneck of feature-representation-based methods. While the design of DREAM mostly follows empirical observations, our work is the first to explore how the face rotation in the 3D space affects the deep feature generation process of CNNs. It mathematically reveals the fundamental and accurate relationship between the rotation and resulting features. Based on these theoretical findings, we propose LARNeXt, which not only significantly outperforms DREAM but also pushes the boundary of the state of the arts.

Prior Art of Lie Algebra: A Lie group includes the structure of a differentiable manifold such that the addition, multiplication, and inverse are differentiable maps, a property not possessed by rotation in $SO(3)$. When considering the derivative, the tangent space of the group forms a Lie algebra ϕ. A simple diagram is shown in Fig. 2. For a more detailed description and proof, refer to the following text and supporting material. Lie algebra theory can adapt to visual tasks with 3D transformations. Tuzel et al. [29] used Lie algebra theory to define a new geodesic distance and design the loss function of the network to optimize the pose estimation training. The *Lie Algebra Residual Network* (LARNet) [30] represents the first attempt to introduce Lie algebra to a face recognition task for improving face recognition performance. However, LARNet also faces limitations, such as its approximations, empirical observations, and additional prior labels. We then propose LARNeXt to compensate for these shortcomings by achieving a more precise derivation, integrating a soft stagewise regression subnet with multi-fusion attention feature aggregation for efficient pose estimation from a single RGB image and providing more ablation experiments and visualization results for corroborating the credibility of our theory and corresponding network designs.

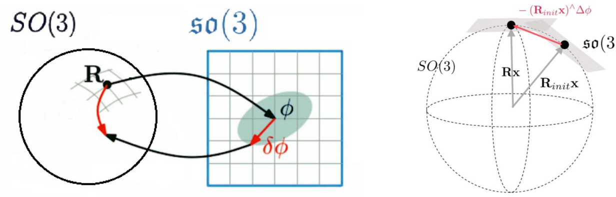


Fig. 2. A simple diagram of Lie Algebra and rotation. Left: Mutual representation of Lie algebra and rotation under the perturbation scheme. Right: Gradient descent form.

III. Methodology

In this section, we assume that a frontal face and its profile face have a corresponding rotation relationship in the original 3D space. For ease of understanding, only the rotation with the orthogonal transformation relationship is discussed here. The derivation of more complex euclidean transformation relationships, including translation and zooming, can be found in our supplementary material.

A. Problem Formulation

Our goals are to find a transformation between the features of an input profile face image and the expected frontal face image, to realize *frontalization* in the deep feature space, and to achieve a powerful feature representation that is robust to pose variations as shown in Fig. 1.

We denote $\mathcal{F}(\mathbf{x})$ as a feature extraction function in CNNs for an input image $\mathbf{x}$. For each pixel (u, v) in image $\mathbf{x}$, we adopt its homogeneous coordinate representation $(u, v, 1)^\top$, and for convenience, we denote the collection of these 3D homogeneous coordinates by $\mathbf{x}$.

Let d be the dimension of layers, and the extracted feature be $\mathcal{F}(\mathbf{x}) \in \mathbb{R}^d$, respectively. We shall prove that there exists a map $\mathcal{R}_{map}(\cdot) : \mathbb{R}^d \to \mathbb{R}^d$ that plays a similar role to rotation in the deep feature space corresponding to the rotation $\mathbf{R} \in SO(3)$ of the (homogenized) image $\mathbf{x}$:

$$\mathcal{F}(\mathbf{R} \cdot \mathbf{x}) = \mathcal{R}_{map}(\mathcal{F}(\mathbf{x})). \tag{1}$$

For the frontal face image $\mathbf{x}_f$ and its profile face image $\mathbf{x}_p$, the homography transformation matrix of these two images degenerates into the rotation matrix: $\mathbf{x}_f = \mathbf{R} \cdot \mathbf{x}_p$. Therefore, we have:

$$\mathcal{F}(\mathbf{x}_f) = \mathcal{F}(\mathbf{R} \cdot \mathbf{x}_p) = \mathcal{R}_{map}(\mathcal{F}(\mathbf{x}_p)). \tag{2}$$

We then use Lie group theory [34] and prove that the mapping $\mathcal{R}_{map}(\cdot)$ can be decomposed into an additive residual component that is solely determined by the rotation as

$$\mathcal{F}(\mathbf{x}_f) = \mathcal{F}(\mathbf{x}_p) + \omega(\mathbf{R}) \cdot \mathbf{C}_{res}(\mathbf{R}, \mathbf{x}_p). \tag{3}$$

Therefore, we only need a residual subnet $\mathbf{C}_{res}$ for decoding pose variant information from the input face image, a robust head rotation estimation subnet for obtaining rotation information, and a gating subnet ω to learn rotation magnitude for controlling the strength of the residual component contributing to the feature learning process. (3) is the core principle of our proposed method. The detailed derivations and experimental design are presented in the following sections.

B. Rotation in Networks and Lie Algebra

To find $\mathcal{R}_{map}$, we directly explore and analyze the role of rotation $\mathbf{R}$ in networks from (2). The authors in ResNet [35] proposed the novel *shortcut*, which not only retains the depth of deep networks, but also has the advantages of shallow networks in avoiding the overfitting issue. The feature learning from the shallow layer l to the deep layer L is described as

$$\mathbf{x}_L = \mathbf{x}_l + \sum_{i=l}^{L-1} H\left(\mathbf{x}_i, w_i\right), \tag{4}$$

$$\begin{aligned} \frac{\partial \mathbf{Loss}}{\partial \mathbf{x}_l} &= \frac{\partial \mathbf{Loss}}{\partial \mathbf{x}_L} \cdot \frac{\partial \mathbf{x}_L}{\partial \mathbf{x}_l} \\ &= \frac{\partial \mathbf{Loss}}{\partial \mathbf{x}_L}\left(1 + \frac{\partial}{\partial \mathbf{x}_l}\sum_{i=l}^{L-1} H\left(\mathbf{x}_i, w_i\right)\right), \end{aligned} \tag{5}$$

where $\mathbf{x}_l$ represents the input of the lth residual block, and $H(\cdot)$ is the residual function with weights w. Given that the second term enclosed in big brackets in (5) quickly drops to 0, we focus on the first principal term $\partial \mathbf{Loss}/\partial \mathbf{x}_L$.

Note that the rotation matrix $\mathbf{R} \in SO(3)$ is not closed under matrix additions. Therefore, in nonlinear optimization of CNNs, updating $\mathbf{R}$ using derivations does not yield a new rotation matrix [36]. Therefore, directly using $\mathbf{R}$ is not appropriate, and we need to find a new approach for embedding $\mathbf{R}$ in the network.

Inspired by prior work [29], we adopt Lie algebra with its own addition, multiplication, and derivative to replace the rotation matrix $\mathbf{R}$ in CNNs. First, each rotation matrix $\mathbf{R} \in \mathbb{R}^{3\times 3}$ corresponds to a vector $\boldsymbol{\phi}$ through the exponential mapping [37]:

$$\mathbf{R} = \exp(\boldsymbol{\phi}^\wedge), \tag{6}$$

where $^\wedge$ is the skew-symmetric operator. The detailed definition of operator $^\wedge$ and a proof for (6) can be found in the supplementary material.

Meanwhile, the vector $\boldsymbol{\phi}$ can be obtained from $\mathbf{R}$ by the following Rodriguez' rotation formula [38] and Taylor expansion:

$$\begin{aligned} \mathbf{R} &= \exp(\theta\psi^\wedge) = \sum_{n=0}^{\infty} \frac{1}{n!}\left(\theta\psi^\wedge\right)^n \\ &= \cos\theta \mathbf{I} + (1-\cos\theta)\psi\psi^T + \sin\theta\psi^\wedge, \end{aligned} \tag{7}$$

where $\boldsymbol{\phi} = \theta\psi$ is in the *Axis-Angle representation* form, with the unit vector $\psi \in \mathbb{R}^3$ being the direction of the rotation axis and θ being the rotation angle according to the right hand rule, respectively. Given that $\mathbf{R}\psi = \psi$, ψ is the eigenvector of matrix $\mathbf{R}$ for eigenvalue $\lambda_{\mathbf{R}} = 1$. (7) leads to

$$tr(\mathbf{R}) = 2\cos\theta + 1. \tag{8}$$

Therefore, we can solve $\boldsymbol{\phi}$ as

$$\boldsymbol{\phi} = \theta\psi = \arccos\left(\frac{tr(\mathbf{R}) - 1}{2}\right)\psi. \tag{9}$$

We then show the addition and multiplication in Lie algebra using *Baker-Campbell-Hausdorff* (BCH) formula [39], [40] and *Friedrichs'* theorem [41], [42] as follows:

$$\begin{aligned} &\exp\left(\Delta\boldsymbol{\phi}^\wedge\right)\exp\left(\boldsymbol{\phi}^\wedge\right) = \exp\left(\left(\boldsymbol{\phi} + \mathbf{J}_l(\boldsymbol{\phi})^{-1}\Delta\boldsymbol{\phi}\right)^\wedge\right), \\ &\exp\left((\boldsymbol{\phi} + \Delta\boldsymbol{\phi})^\wedge\right) = \exp\left((\mathbf{J}_l\Delta\boldsymbol{\phi})^\wedge\right)\exp\left(\boldsymbol{\phi}^\wedge\right), \end{aligned} \tag{10}$$

where $\mathbf{J}_\ell$ is the $left\ Jacobian$ of $SO(3)$. For point $\mathbf{p} \in \mathbb{R}^3$, the derivative of $\mathbf{Rp}$ with respect to a perturbed rotation is

$$\begin{aligned} \frac{\partial(\mathbf{Rp})}{\partial(\Delta\boldsymbol{\phi})} &= \lim_{\Delta\boldsymbol{\phi}\to 0} \frac{\exp\left(\Delta\boldsymbol{\phi}^\wedge\right)\exp\left(\boldsymbol{\phi}^\wedge\right)\mathbf{p} - \exp\left(\boldsymbol{\phi}^\wedge\right)\mathbf{p}}{\Delta\boldsymbol{\phi}} \\ &= -(\mathbf{Rp})^\wedge. \end{aligned} \tag{11}$$

For a current $\mathbf{R}_i$, we choose perturbation $\Delta\phi^\wedge$, such that $\mathbf{R}_{i+1} = \exp(\Delta\phi^\wedge)\mathbf{R}_i$. Then, for point $\mathbf{p}$, (11) leads to

$$\mathbf{R}_{i+1}\mathbf{p} = \exp(\Delta\phi^\wedge)\mathbf{R}_i\mathbf{p} \approx \mathbf{R}_i\mathbf{p} - (\mathbf{R}_i\mathbf{p})^\wedge\Delta\boldsymbol{\phi}. \tag{12}$$

Meanwhile, for the target function to be optimized (denoted by u), we use Taylor expansion to derive

$$\begin{aligned} u(\mathbf{R}_{i+1}\mathbf{p}) &= u\left(\exp\left(\Delta\phi^\wedge\right)\mathbf{R}_i\mathbf{p}\right) \approx u\left(\left(\mathbf{1} + \Delta\phi^\wedge\right)\mathbf{R}_i\mathbf{p}\right) \\ &\approx u\left(\mathbf{R}_i\mathbf{p}\right) - \underbrace{\left.\frac{\partial u}{\partial \mathbf{d}}\right|_{\mathbf{d}=\mathbf{R}_i\mathbf{p}} (\mathbf{R}_i\mathbf{p})^\wedge}_{\boldsymbol{\delta}^T} \Delta\boldsymbol{\phi} \\ &= u\left(\mathbf{R}_i\mathbf{p}\right) + \boldsymbol{\delta}^T\Delta\boldsymbol{\phi}. \end{aligned} \tag{13}$$

We need to determine $\Delta\phi$ such that the value of u decreases. A possible choice is to select $\Delta\phi = -\alpha D\delta$, where $\alpha > 0$ is a small step size, and D is an arbitrary positive-definite matrix.

By applying this perturbation within the scheme, we can update the rotation matrix by $\mathbf{R}_{i+1} \leftarrow \exp(-\alpha \mathrm{D}\delta^{\wedge})\mathbf{R}_i$.

Back to the original problem, given (10)–(13), we can rewrite the first principal term of (5) as

$$\begin{aligned}\frac{\partial \mathbf{Loss}}{\partial \mathbf{x}_L^f} &\approx \lim_{\Delta\boldsymbol{\phi}\to 0} \frac{\partial \mathbf{Loss}}{\exp\left((\boldsymbol{\phi}+\Delta\boldsymbol{\phi})^{\wedge}\right)\cdot \mathbf{x}_L^p - \exp\left(\boldsymbol{\phi}^{\wedge}\right)\mathbf{x}_L^p} \\ &= \frac{\partial \mathbf{Loss}}{-(\mathbf{R}\cdot\mathbf{x}_L^p)^{\wedge}\cdot\partial\Delta\boldsymbol{\phi}} \\ &= \frac{\partial \mathbf{Loss}}{\partial(\mathbf{R}\cdot\mathbf{x}_L^p)}. \qquad (14)\end{aligned}$$

Note that in (2), the homography relationship between the two original images $\mathbf{x}_p$ and $\mathbf{x}_f$ is connected by a rotation, but this relationship generally cannot be guaranteed in CNNs. However, (14) suggests that this relationship is inherited in another way during the gradient descent at each layer. In fact, given that $\mathbf{R} \in SO(3)$, $\mathbf{R}\cdot\mathbf{x}_L^p$ and $\mathbf{x}_L^f$ are asymptotically stable according to Lyapunov's second method [43], [44]. With the gradual training progress of ResNet, the feature vectors of $\mathbf{R}\cdot\mathbf{x}_L^p$ and $\mathbf{x}_L^f$ have the same convergent representation, that is, $\mathcal{F}(\mathbf{x}_f) = \mathcal{F}(\mathbf{R}\cdot\mathbf{x}_p)$. Furthermore, we decouple the rotation relation from face features into (9) and (12). Let $V_{res} = \mathcal{F}(\mathbf{R}\cdot\mathbf{x}_p) - \mathcal{R}_{map}(\mathcal{F}(\mathbf{x}_p)) \in \mathbb{R}^d$ be the residual vector, and we have:

$$\begin{aligned}&\mathcal{R}_{map}^{-1}(\mathcal{F}(\mathbf{x}_f)) = \mathcal{F}(\mathbf{x}_p) + \mathcal{R}_{map}^{-1}(V_{res}), \\ &\mathcal{F}(\mathbf{x}_f) = \mathcal{F}(\mathbf{x}_p) + \mathcal{R}_{map}^{-1}(V_{res} + \mathcal{R}_{map}(\mathcal{F}(\mathbf{x}_f)) - \mathcal{F}(\mathbf{x}_f)).\end{aligned} \qquad (15)$$

Given that the feature $\mathcal{F}(\mathbf{x}_p)$ is approaching to $\mathcal{R}_{map}(\mathcal{F}(\mathbf{x}_f))$ in the training stage (see the corresponding analysis in (17 in Section III-C), (15) leads to

$$\mathcal{F}(\mathbf{x}_f) \approx \mathcal{F}(\mathbf{x}_p) + \mathcal{R}_{map}^{-1}(\mathcal{F}(\mathbf{x}_p) - \mathcal{R}_{map}(\mathcal{F}(\mathbf{x}_p))), \qquad (16)$$

which agrees exactly with (3). Therefore, we can design the gating control function $\omega(\mathbf{R})$ as $\mathcal{R}_{map}^{-1}$ to filter the feature flow and maintain geometric constraints, and build the head pose estimation subnet to efficiently obtain accurate rotation information $\mathcal{R}$. Meanwhile, the component $\mathbf{C}_{res}(\mathbf{R}, \mathbf{x}_p) = \mathcal{F}(\mathbf{x}_p) - \mathcal{R}_{map}(\mathcal{F}(\mathbf{x}_p))$ is solved through residual subnet training.

C. *Subnet Mechanism*

As previously stated, we theoretically prove the feasibility and provide a novel solution to the face recognition problem. Following our theory, we design a succinct architecture for our LARNeXt, and present an intuitive understanding of relatively complex mathematical formulas, shown in Fig. 3. We also explain the important compositions in detail and carefully demonstrate their validity and rationality in this section.

1) Backbone: First, we propose a mutual representation relationship between the features of frontal and profile faces to guide our face recognition. However, before doing so, we need a feature extraction backbone with superior performance. Inspired by previous network theory researches, such as Saxe et al.' [45], Highway Networks [46], and Balduzzi et al.' [47], we choose ResNet-50 as the network with the best layers after weighing efficiency and accuracy. This network is also prevalent in the literature and is very convenient for comparison. We select ResNet-50 with a combined margin loss (CM(1, 0.3, 0.2)) of ArcFace [3], whose loss function combines all margin penalties of SphereFace [1] and CosFace [2]. Given that we only need its feature extraction and hardly output the final classification results, we cut off the final fully-connected (FC) layer commonly used for classification in the original network. For any input face image $\mathbf{x}$, this backbone generates the corresponding deep feature $\mathcal{F}(\mathbf{x})$.

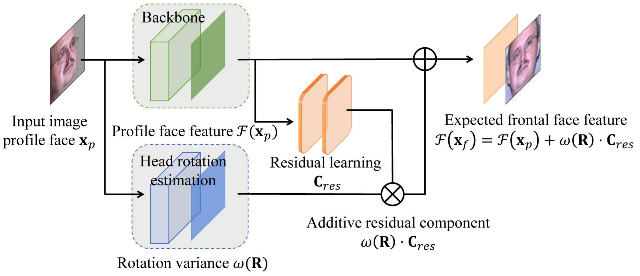

Backbone
Input image
profile face $\mathbf{x}_p$
Profile face feature $\mathcal{F}(\mathbf{x}_p)$
Head rotation
estimation
Rotation variance $\omega(\mathbf{R})$
Residual learning
$\mathbf{C}_{res}$
Additive residual component
$\omega(\mathbf{R})\cdot\mathbf{C}_{res}$
Expected frontal face feature
$\mathcal{F}(\mathbf{x}_f) = \mathcal{F}(\mathbf{x}_p) + \omega(\mathbf{R})\cdot\mathbf{C}_{res}$

Fig. 3. The architecture of the proposed LARNeXt. Adding a succinct residual subnet, an efficient rotation estimation subnet, and the gating control function to the clean feature learned from the existing state-of-the-art backbone would suffice.

2) Residual Learning Subnet: From (3) and (16), we expect to design a residual subnet $\mathbf{C}_{res}(\mathbf{R}, \mathbf{x}_p) = \mathcal{F}(\mathbf{x}_p) - \mathcal{R}_{map}(\mathcal{F}(\mathbf{x}_p))$ for decoding pose variant information from profile face features. The residual formulations allow us to use a succinct enough network structure for learning the residual compensation from the clean deep features, which is a relatively easy task. Residual learning can be arranged behind in the backbone without revising any learned parameters of the original model. Our residual learning has two fully-connected layers, with Parametric Rectified Linear Unit (PReLU) [48] as the activation function. We train this subnet by minimizing ℓ_2 norm of the difference between the profile features $\mathcal{F}(\mathbf{x}_p)$ and frontal features under the rotation $\mathcal{R}_{map}(\mathcal{F}(\mathbf{x}_f))$ using stochastic gradient descent.

$$\min_{\Omega_p} \Sigma||\mathcal{F}(\mathbf{x}_p) - \mathcal{R}_{map}(\Omega_p, \mathcal{F}(\mathbf{x}_f))||_2^2, \qquad (17)$$

where Ω_p denotes the learnable parameters. We train this subnet on frontal-profile pairs sampled from the MS-Celeb-1 M dataset (mentioned in Section IV-B), and fix these parameters for the testing. Applying a subnet with a complicated structure may increase the risk of overfitting, and the design with two FC layers considers both the task difficulty and the risk of model robustness. We demonstrate the superiority of our design in the following ablation experiment(Section IV-C1), and find that such a succinct structure is enough.

3) Head Rotation Estimation Subnet: This section mainly studies how to perform robust and efficient face pose estimation on a single image to obtain the angle θ. Considering the trade-off between accuracy and efficiency, as well as the constraints of the single view from a single image, we finally believe that the

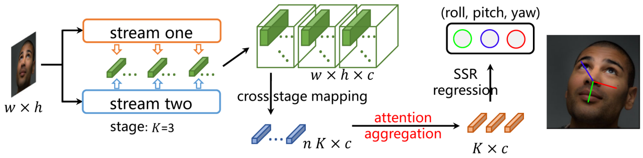


Fig. 4. The architecture of the head rotation estimation subnet. We propose a novel multi-fusion attention mechanism to explore the importance of feature aggregation and use a soft regression SSRNet to achieve a high precision rotation estimation.

design of the age estimation work SSRNet [49] is an attractive solution, whose ideas of stagewise regression and dynamic range significantly reduce the model size (about 0.32 M) yet maintain high precision. However, for the pose estimation problem we are concerned, it is necessary to transform its feature aggregation strategy, which is a constant estimation of age, into a vector estimation of the rotation angle. In conclusion, we model the pose estimation problem as follows:

$$\theta = \sum_{k=1}^{K} p^{(k)} \cdot V_{ra}^{(k)}, \tag{18}$$

where K is the number of stages ($K = 3$ in our problem context), and $V_{ra}^{(k)}$ is a representative vector of rotation angle θ groups with the corresponding probability distribution $p^{(k)}$ at the k stage. In this dynamic range case, a full-space classification problem becomes a hierarchical classification solution, and we need to aggregate features into representative vectors at each stage.

For the differences between our situation (three different angle estimations) and age estimation (a single prediction), we need a suitable feature aggregation method to polish the original regression backbone further. We investigate and analyze many extant methods: NetVLAD [50] and Capsule [51] try to extract features from a large whole but sacrifice the potential spatial information of the feature map. FSA-Net [52] proposes a fine-grained mapping strategy, but its ad-hoc designed feature-scoring function still has shortcomings that limit empirical performance. To address this issue, we adopt the fine-grained mapping strategy along with our novel multi-fusion attention map and replace this scoring function with the widespread attention mechanism to assign the weights of the features. The architecture of the head rotation estimation subnet is shown in Fig. 4. Note that SSRNet proposes these specific structures, e.g., stream one/two and SSR regression, and refers to Section VI in the supporting material for more details. Below we will solely explain our novel feature aggregation strategy.

For each feature map $M(w \times h \times c)$ exacted from SSRNet, we have the attention map $\Psi(M(i,j))$, where $\Psi(\cdot)$ is our designed pixel-level attention function and $M(i,j) \in \mathbb{R}^c$. Originally, we tried two solutions to $\Psi(\cdot)$, namely, 1×1 convolution and variance attention. In the first solution, $\Psi_{1\times1}(M(i,j)) = sigmoid(\sigma_{ker} * M(i,j))$, where σ_{ker} is a learnable convolution kernel, and the 1×1 convolution plays a role in weighting features from input data sources. Another solution takes variance into consideration: $\Psi_{var}(M(i,j)) = \sum_{n=1}^{c}(M_n(i,j) - \hat{M}(i,j))^2$, where $\hat{M}(i,j) = 1/c \sum_{n=1}^{c} M_n(i,j)$. Nevertheless, we find the shortcomings of these methods via an in-depth analysis. $\Psi_{1\times1}$ may lead to a risk of model overfitting, whereas Ψ_{var} is not learnable and is very sensitive to the difference between the training and testing data. Meanwhile, we draw inspiration from the "short-cut" mechanism of ResNet, and the identical map: $\Psi_{id}(M(i,j)) = M(i,j)$ can provide more potential information. We build a multi-fusion attention map, which averages all of the above predictions similar to ensemble learning as follows:

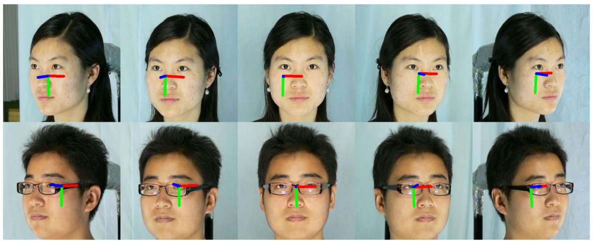


Fig. 5. Visualization results of our rotation estimation subnet on lab-standard labeled samples from the OFD dataset. The input only needs a single RGB image, and the colored lines (green, blue, and red) indicate the directions of head rotation angle (roll, pitch, and yaw).

$$\Psi_{fus} = \frac{\Psi_{1\times1} + \Psi_{var} + \Psi_{id}}{3}. \tag{19}$$

Simple average guarantees a robust low bound when facing unknown training/test data. To prove the superiority and rationality of our attention map, we conduct the corresponding ablation experiments in Section IV-C2. After feature aggregation, we use SSRNet to obtain the robust pose estimation $\theta = (pitch, yaw, roll)$. The head rotation estimation subnet is an embedded complete pose estimation module. We provide the visualization results of our rotation estimation subnet on lab-standard labeled samples from the OFD dataset in Fig. 5. Users can also use ground-truth prior labels or other models. To verify the superiority of our proposed head rotation estimation subnet, we also conduct comparison experiments in Section IV-C2 and provide extensive experimental results.

4) Gating Control Function: After obtaining the accurate rotation angle, we need to design a gating control function $\omega(\mathbf{R})$ to analyze the rotation magnitude for controlling the strength of the residual component contributing to the feature learning process. Our attempt can be viewed as a correction mechanism that adopts top-down information to influence the feed-forward process and as an activation function to filter information flow. In our problem context, ω needs to satisfy the following geometric constraints:

- $\omega \in [0, 1]$. Intuitively, when the input is frontal face input $\mathbf{x}_0$, almost no difference can be detected in the feature representation in the same network, and $\mathbf{C}_{res}$ of residual learning introduces errors and compromises the classification ability. Therefore, the gating control function is expected to be 0 at this time. Ideally, the magnitude of the residual is the largest at the complete profile pose: $\mathcal{F}(\mathbf{x}_0) - \mathcal{F}(\mathbf{x}_{\pi/2})$. In this case, when the maximum value of the gating control function is 1, we have :

$$\mathcal{F}(\mathbf{x}_0) = \mathcal{F}(\mathbf{x}_{\pi/2}) + 1 * (\mathcal{F}(\mathbf{x}_0) - \mathcal{F}(\mathbf{x}_{\pi/2})) = \mathcal{F}(\mathbf{x}_0).$$

• ω has symmetric weights. A gating control function tries to learn the rotation magnitude for controlling the strength of the residual component that contributes to the feature learning process, and the same deflection angle should have the same influence (e.g., yaw angle turning left or right). We also use data flipping augmentation to improve the symmetry of the model during training.

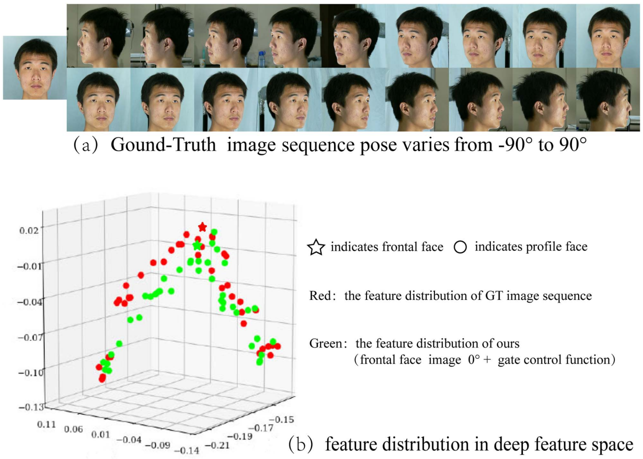
The roll, yaw, and pitch angles contribute differently to the final face recognition performance. Face alignment eliminates the roll's effect (Section IV-A). While face images with large pitch angles are relatively rare, we cannot ignore their small yet essential contributions when pursuing further improvements in recognition performance. Therefore, we abandon the rough ℓ_∞ approximation of the conference version and study how to represent rotation angles more geometrically appropriately. From (9), under *axis-angle representation*, the axis vector ψ is a unit vector that indicates the direction of the rotational offset, and rotation angle θ represents the modulus length of the Lie algebra $\boldsymbol{\phi}$. Therefore, the ℓ_2 norm presents a suitable and precise representation for Lie algebra in LARNeXt.

After combining all the above constraints and solving (8) via Chebyshev polynomial approximation, we obtain $\omega = |\sin\theta|$ with $\sin\theta = \sin(||(\theta_{pitch}, \theta_{yaw}, \theta_{roll})||_2)$, for all angles $\in [-\pi/2, \pi/2]$, which ensures a one-to-one correspondence between the elements in Lie algebra ϕ and the rotation $\mathbf{R}$ and guarantees the completeness of the proposed theory. The existing work SIREN [54] has also shown that periodic activation functions such as our used $\sin$ will achieve great performances. Similarly, we conduct another corresponding ablation experiment of different gating control functions in Section IV-C to prove that our design outperforms its competitors. We only discuss here the visual feature distribution results that improve the acceptance and understanding of our theory in a qualitative level.

To demonstrate the effectiveness of our proposed subnet, Fig. 6 illustrates its application for the same identity. When image sequences of the same individual with different yaw angles are used as inputs, ResNet can extract features and display the distribution of vectors as red dots. We use our gating control function with only the frontal face image to simulate pose variations, and green dots denote the distribution of our result. This visualization clearly shows that our model can accurately simulate the feature vector distribution of different faces varying from yaw angles, thereby proving that our gating control function has an improved feature representation capability and is especially amenable to pose variations.

Fig. 7 shows the effect of our subnet for different identities. This figure presents a challenging example even for the almost blameless face recognition model [55]. We collect more frontal and profile face data of the same two individuals and visualize the feature vectors of all images corresponding to this sample. Our subnet with the gating control function is instrumental in achieving better classification and clustering performance.

To intuitively understand our theory further, we present more rendered image results after reconstructing faces with the corresponding features. We use the encoding-decoding mechanism of an advanced TP-GAN model in [56] that can map deep features back to the reconstructed images. This mechanism is used to visualize the original and mapped features generated by our model. Some representative results are shown in Fig. 8. The rendered reconstruction images are only used for visualization purposes. The superiority of our LARNeXt can be fully validated by quantitatively examining its performance in various face recognition tasks. We also show many convincing experimental results in the following section.


(a) Gound-Truth image sequence pose varies from -90° to 90°

(b) feature distribution in deep feature space

Fig. 6. The effect of our gating control function on the same identity. (a) The top presents a sequence of images taken in real life, with the pose variant ranging from $-90°$ to $90°$ for the same individual. (b) The bottom shows the feature distribution in the deep feature space. The dots represent the profile faces whereas the stars denote the frontal faces. The red dots are the feature vectors generated by the image sequence, and the green dots are the feature vectors of the frontal face image (0°) with different yaw angle variants simulated by our gating control function. Their similar distributions indicate that our gating control function closely maps the features of the frontal and profile faces, thereby enhancing the feature representation ability to accommodate pose variations.

IV. Experimental Results

We initially describe the implementation details of LARNeXt (Section IV-A) and then briefly describe all the datasets we used in the experiments along with their characteristics (Section IV-B). We present many ablation studies and explain the contribution of our experimental design to recognition performance (Section IV-C). We also compare LARNeXt with existing methods and some findings on profile face representations and then conduct detailed experiments on frontal-profile face verification-identification tasks, general face recognition tasks and a industrial-grade mega dataset (Section IV-D).

A. Implementation Details

Data Preprocessing: As shown in Fig. 9, we use MTCNN [57] to detect the face areas and facial landmarks on both the training and testing sets. We use flipping to achieve data enhancement and strengthen the ability of our model to learn symmetry. We also apply face alignment and scaling (224×224) to reduce the impact of translation and zooming when we only consider $SO(3)$ instead of $SE(3)$.

Training Details: The model is trained in 180 K iterations with an initial learning rate of 0.1, and the learning rate is divided by 10 at 100 K and 160 K iterations. The SGD optimizer has

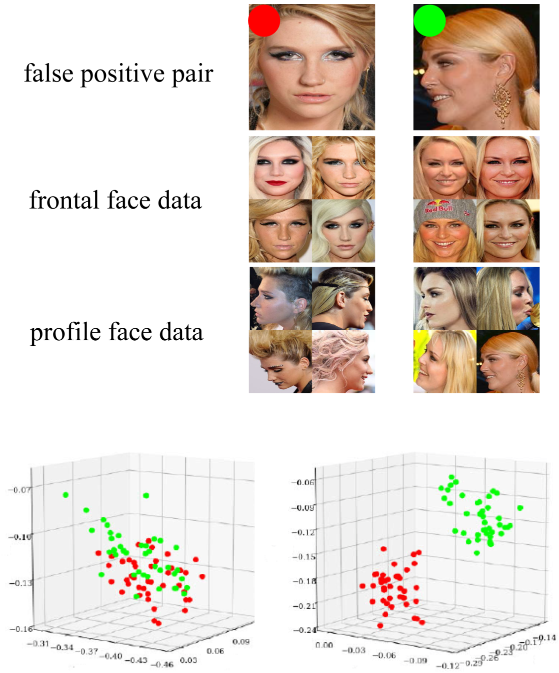
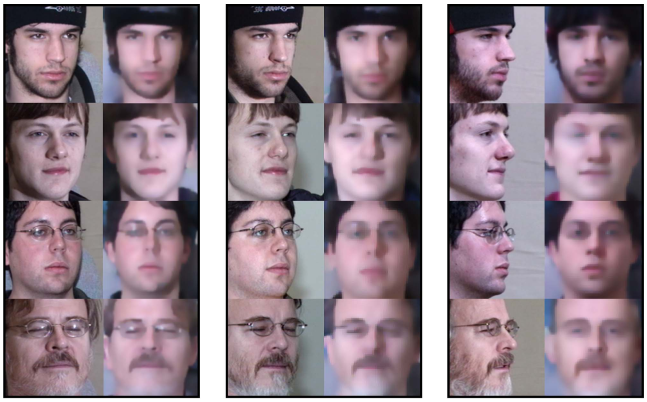


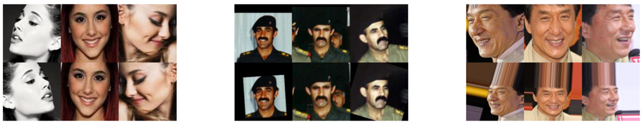
Fig. 7. The effect of our gating control function on different identities. This is a challenging false positive example for a general face recognition model [2]. We collect more frontal and profile face data of two individuals from the Celebrities Frontal-Profile dataset [53] and visualize the feature distributions of all images. Our model (on the right) with the gating control function obviously has a superior classification and clustering ability.



15° profile faces 45° profile faces 60° profile faces

Fig. 8. Rendered face reconstruction images of 15°, 45°, and 60°. The odd columns show the original profile faces, whereas the even columns depict the reconstructed visualization results after deep feature mapping using our method. The feature representation has not diminished regardless of the influencing factors, including gender, face decoration (glasses and beard), and head decoration (hat).

a momentum of 0.9, and weight decay of $5e{-}4$. We train the ResNet and residual learning together and then train the residual learning separately with pose variant frontal-profile face pairs and dropout$= 0.7$. As for the rotation angles, we follow the settings of SSRNet and FSA-Net, whose architectures are shown in the supporting material with parameters $(w, h, c) = (8, 8, 64)$ for the feature map and $m = 5$, $n = 7$, $c' = 16$ for the feature aggregation.



(a) face flipping (b) face alignment (c) face scaling

Fig. 9. Prepossessing data on frontal and profile faces. (a) Face flipping: Data enhancement for strengthening the ability of our model to learn symmetry; (b) face alignment: reduce the impact of translation and rotation in a plane such that the eyes lie along a horizontal line; (c) face scaling: reduce the impact caused by zooming because focal length differs across all images such that they are approximately identical in size.

Efficiency: The inference time of our model is around 5 ms per image, whereas the time for pose estimation is around 0.17 ms per image. Our lightweight LAR block only adds 1.3˜M FLOPs based on ResNet-50 (4˜G FLOPs) with head rotation estimation subnet 1.01˜G FLOPs.

B. Datasets Exhibition

Training Data: We separately employ the two most widely used face datasets, namely, the cleaned MS-Celeb-1 M database (*MS1MV2*) [58] and *CASIA-WebFace* [59], as training data to achieve a fair comparison with other methods. MS1MV2 is a clean version of the original MS-Celeb-1 M face dataset that has too many mislabeled images, and contains 5.8 M images of 85,742 celebrities. Meanwhile, CASIA-WebFace, which uses tag-similarity clustering to remove noise from the data source, contains 500 K images of 100 K celebrities from IMDb.

Testing Data: We explore many efficient face verification datasets for testing. In accordance with the order of the different face recognition task requirements in the following experiments, we briefly introduce the scales and characteristics of each dataset. (a) When studying the distribution of depth features of input images, we need accurate pose labels and select the Oriental Facial Database (*OFD*) [60], which sorts out 33,669 face images of 1,247 volunteers, with each volunteer taking 19 viewpoint images from -90 to 90 degrees at intervals of 10 degrees. The first subfigure in Fig. 6 presents a simple example. (b) Our comparative experiments with competitors are mainly conducted on profile datasets with large poses. Celebrities in Frontal-Profile (*CFP*) [53] is a challenging frontal to profile face verification dataset that contains 500 celebrities, each of which has 10 frontal and 4 profile face images. We extensively test another challenging dataset, the IARPA Janus Benchmark A (*IJB-A*) [61], which covers extreme poses and illuminations and contains 500 identities with 5,712 images and 20,414 frames extracted from videos. (c) Apart from focusing on frontal-profile face verification, we also conduct experiments on general face recognition datasets to verify that our method can reach the state-of-the-art for general face recognition tasks. By including the most widely used *LFW* [62] dataset (13,233 face images from 5749 identities) and *YTF* [63] dataset (3,425 videos of 1,595 different people), we also report the performance of Cross-Pose LFW (*CPLFW*) [64], which deliberately searches and selects 3,000 positive face pairs with pose difference to add pose variation to intra-class variance and to fully justify

TABLE I
ABLATION STUDY ON THE ARCHITECTURES OF RESIDUAL SUBNET

Architecture of residual subnet	EER
one FC	9.96
one FC + one Conv.	8.84
one Conv. + global average pooling	8.61
one Conv. + max pooling	8.73
Ours: two FC	**7.92**

The Training inputs include frontal-profile face pairs from the MS1MV2 dataset. Evaluation is conducted on the CFP-FP dataset With a metric equal error rate.

TABLE II
ABLATION STUDY ON THE FEATURE AGGREGATION OF THE HEAD ROTATION ESTIMATION SUBNET

Method	MAE
w./o. aggregation	6.95
NetVLAD [50]	5.97
Capsule [51]	4.24
FSA-Net [52]	3.75
Ours: $\Psi_{1\times1}$	3.77
Ours: Ψ_{var}	3.68
Ours: Ψ_{fus}	**3.60**

The training settings are 300W-LP dataset and SSRNet. Evaluation is conducted on the BIWI dataset with a metric mean absolute error (MAE).

the effectiveness of several face verification methods. (d) We also extensively conduct an in-depth ablation experiment on the large-scale CelebFaces Attributes (*CelebA*) Dataset [65], which contains 10,177 celebrities and 202,599 face images covering large pose variations.

Pose Data: We use three popular datasets for the training and testing of rotation pose estimation. *300W-LP* dataset [66] is a 3D dataset based on the 300 W dataset and 3DMM model simulation. This dataset contains 68 key points and camera parameters, and adopts the proposed face profiling to generate 61,225 samples across large poses (1,786 samples from IBUG, 5,207 from AFW, 16,556 from LFPW and, 37,676 from HELEN). *AFLW2000* [66] is a dataset containing 2000 images that have been annotated with image-level 68-point 3D facial landmarks. This dataset is used for evaluating 3D facial landmark detection models. The head poses are very diverse and often difficult to detect by using a CNN-based face detector. The *BIWI* dataset [67] contains over 15 K images of 20 people (6 females and 14 males). The head pose range covers about ±75 degrees yaw and ±60 degrees pitch. The ground truth is provided in the form of the 3D location of the head and its rotation.

C. Ablation Studies

To prove that the proposed LARNeXt improves profile face recognition performance, we conduct many ablation experiments for the architectures with the gating control function, for the form of the gating control function, for the multi-fusion attention feature aggregation strategy and head pose estimation performance, and for the feature distributions of other state-of-the-art face recognition models with or without our subnet design.

1) Residual Subnet Architecture: In this subsection, we study the effectiveness of architectures with different components. We train our results with the same backbone ResNet-50 and frontal-profile face pairs from MS1MV2 dataset. We perform the evaluation on the CFP-FP dataset with a metric Equal Error Rate (EER).

Table I compares the experimental performance of the two-layer FC with other commonly used succinct architectures and provides some quantitative and reliable results. The one-level FC obtains an unsatisfactory result because its linear structure is too simple for addressing complex problems. The one-level FC with a 1D convolution layer preforms slightly better in learning input patterns for the model. However, its local weight sharing mechanism presents an obstacle in performance improvement. GoogLeNet [68] uses the global average pooling (GAP) method to reduce the number of parameters and the risk of overfitting. This deep feature fusion method has excellent prediction performance and can be used in different tasks, such as semantic segmentation. However, for our required feature representation, the model capacity of GAP is slightly inferior. Max pooling allows for a more rapid convergence due to the larger number of gradients returned during back-propagation (LeCun [69]). However, this approach only has a slight effect on performance improvement. Many studies have discussed the role of two-layer FC adopted by our residual network design. For example, the classic deep learning network AlexNet [70] proves that two-layer FC is a reasonable approximation that facilitates the learning of input patterns for the model and shows that removing any FC will lead to a drop in performance by around 2%. We believe that the two-layer FC can map the abstract information in the receptive fields of different sizes to a larger space, thereby improving the nonlinear expression ability of the model. The experimental results show that the adopted two-layer FC structure achieves the leading performance.

2) Head Rotation Subnet: We then examine the effects of different feature aggregation strategies. On the basis of the soft regression of SSRNet, we explore the performances of (1) w/o aggregation (without aggregation), (2) NetVLAD [50], (3) Capsule [51], (4) the scoring function of FSA-Net [52], and (5) Ours: $\Psi_{1\times1}$, Ψ_{var}, and Ψ_{fus}. We conduct fair testing experiments on the BIWI dataset, and the results are using the same backbone SSRNet and trained on the 300W-LP dataset with the metric Mean Absolute Error (MAE).

Table II shows that the w/o aggregation strategy, where the stage of SSRNet is $K = 1$, only outputs the most important orientation of the rotation angle. Therefore, this strategy only yields a rough numerical result. Meanwhile, NetVLAD and Capsule both consider the reduced feature aggregation method from large to small ($K = 3$), and improve the performance by adopting different reduction methods. FSA-Net proposes a novel fine-grained feature mapping for weight calculation and further enhances the improvement resulting from feature aggregation. We introduce an attention mechanism into our problem that targets the importance of features, and all our variants clearly outperform the other methods. In particular, the result of our designed multi-fusion method Ψ_{fus} is almost half of that w/o aggregation, which represents a huge performance improvement.

TABLE III
COMPARISONS WITH POSE ESTIMATION METHODS

Method	Yaw(°)	Pitch(°)	Roll(°)	MAE
ERT(68 points) [71]	23.53	13.18	10.61	17.48
FAN(12 points) [72]	12.40	6.71	8.35	10.17
KEPLER(GoogLeNet) [73]	5.86	11.27	8.92	9.65
3DDFA(standard model) [66]	5.53	8.25	8.40	8.21
Hopenet(best $\alpha = 2$) [74]	6.56	6.44	5.47	6.84
FSA-Net(best caps) [52]	4.08	6.64	4.50	5.63
CTFIO(best refinement) [75]	3.53	4.12	3.11	3.99
3DDFA-v2(M+R+S)	-	-	-	3.51
Img2pose(best)	3.43	5.03	3.28	3.91
Ours: Ψ_{fus}	**3.06**	**3.18**	**2.53**	**2.74**

The 300w-LP dataset is used for the training. Evaluation is conducted on the AFLW2000 dataset with the differences between the euler angles (roll, pitch, and yaw) of estimation and ground truths as criteria. The Quantitative evaluation results under the metric MAE are also provided.

Given that the head rotation estimation subnet is a complete pose estimation structure, we also compare our subnet with other advanced pose estimation methods, such as ERT [71], FAN [72], KEPLER [73], 3DDFA [66], Hopenet [74], FSA-Net [52], CTFIO [75], 3DDFA-v2 [76] and Img2pose [77]. The results use the same training 300W-LP dataset. To display these results intuitively and conveniently, we provide two evaluation criteria, namely, 1) the differences between Euler angles (roll, pitch, and yaw) of the estimation and ground truth, and 2) the metric MAE.

As shown in Table III, our multi-fusion attention Ψ_{fus} brings a significant improvement in all evaluation indicators. As mentioned earlier, many methods ignore spatial information, and the existing preprocessing procedure can solve rotation in the plane. Therefore, these methods suffer from the sensitive *roll* item, which should have been a relatively simple task but still remains a problem. Our results achieve an average 2.53° error of the roll angle, and the performance on the yaw and pitch angle is also very eye-catching. Under the more comprehensive MAE metric, our result (2.74) is nearly one-third ahead of the recent CTFIO (best refinement: 3.99) [75] proposed in 2020. Our results are also superior over the recent well-known methods 3DDFA-v2(M+R+S[all modules]:3.51) and Img2pose (best refinement: 3.91). Our attention feature aggregation mechanism also avoids a complex network design and achieves efficiency with precision. The time cost for pose estimation is about $0.17\tilde{}ms$ per image.

3) The Gating Control Function: We further study which gating control function has the greatest contribution to performance. To achieve a fair comparison with existing methods, the CASIA-WebFace dataset and ResNet-50 are used for the training, and an evaluation is conducted on Cele-A dataset with a metric Equal Error Rate (EER).

In Table IV, identity mapping is denoted by $\omega \equiv 1$, which represents some GAN-based works yet ignores the internal connection of the frontal-profile face, and relies only on the generator and discriminator to produce results. The linear mapping $\omega = 2\theta/\pi$ represents a natural attempt and meets the geometric constraints mentioned previously. To some extent, our gating control function acts as a filtering activation function, so we compare that with two widely used activation functions PReLU [48] and cReLU with OW [47]. The nolinear mapping $\omega = sigm(4\theta/\pi - 1)$ is reported by DREAM [15], and is also taken into consideration. LARNet [30] $\omega = |\sin\theta|$ achieves an outstanding EER of 6.26. This observation ascertains its high degree of correction to a profile face. Furthermore, our LARNeXt addresses the approximation problem left by LARNet and offers a better representation of the rotation angle. We fully consider the contribution of yaw, pitch, and roll to the final result theoretically, which cannot be ignored when pursuing further performance improvement. As a result, we surpass the frontier of LARNet (6.26) and obtain a higher score of 6.03.

TABLE IV
ABLATION STUDY ON THE GATING CONTROL FUNCTION

Gating Control Function	EER
Identity mapping: $\omega \equiv 1$	15.35
Linear mapping: $\omega = 2\theta/\pi$	9.68
Nolinear mapping: $\omega = sigmoid(4\theta/\pi - 1)$	8.45
PReLU	9.72
cReLU with OW	7.92
LARNet: $\omega = \lvert\sin(\theta_\infty)\rvert$	6.26
Our LARNeXt: $\omega = \lvert\sin(\theta_2)\rvert$	**6.03**

The training settings are the CASIA-webface dataset and ResNet-50. Evaluation is conducted on the celeba dataset with a metric equal error rate (EER).

TABLE V
ABLATION STUDY ON OUR LAR SUBNET FOR THE FEATURE DISTRIBUTION

Method		MSE
Baseline	ISM	0.56
Ours	ISM + LAR	**0.26**
Baseline	ArcFace	0.23
Ours	ArcFace+LAR	**0.09**

The training settings are MS1MV2 dataset and ResNet-50. Evaluation is conducted on the OFD dataset with a metric mean square error (MSE).

4) Feature Distribution: We study from a quantitative level whether our subnet design based on Lie algebra theory actually strengthens the feature representation and clustering capabilities of the original backbone, which is qualitatively described in Section III-C. We choose two excellent representative face recognition methods: independent softmax (ISM) [55] and ArcFace [3] as baselines, and will respectively explore the performance of these models on the feature distribution before and after our LAR subnet is added, which also helps to alleviate the gap in face verification based deep feature. To achieve a fair comparison, the MS1MV2 dataset and ResNet-50 are used for the training. The evaluation is conducted on OFD dataset with a metric Mean Square Error (MSE) between the deep features of each pair, which consists of a ground-truth face image with pose label and a profile face feature generated by our models. An individual visualization example is shown in Fig. 6, and our experiment will give quantitative results on the whole dataset.

As shown in Table V, in the comparisons between the baselines and ours, we find that LAR subnet can reduce the MSE even for strong and robust face recognition models. Our LAR subnet also shows a more than 50% improvement for each baseline ($0.56 \rightarrow 0.26$ and $0.23 \rightarrow 0.09$), thereby proving our theory about feature representation and distribution.

D. Quantitative Evaluation Results

We compare our method with more than 30 excellent methods with different loss functions that are proposed from 2016 to

TABLE VI
QUANTITATIVE EVALUATION ON THE IJB-A DATASET. AND O.S. DENOTES THE OPTIMAL SETTING, WHEREAS F Denotes fine tuning/refinement. The symbol '-' indicates that the metric is not available for that protocol. The MS1MV2 dataset is used for the training.

Method	TAR@FAR=0.01	TAR@FAR=0.001	Rank-1	Rank-5
Wang *et al.* [78]	0.729	0.510	0.822	0.931
Pooling Faces [23]	0.819	0.631	0.846	0.933
Multi Pose-Aware [17]	0.787	—	0.846	0.927
DCNN Fusion (f.) [79]	0.838	—	0.903	0.965
PAMs [16]	0.826	0.652	0.840	0.925
Augmentation+Rendered [20]	0.886	0.725	0.906	0.962
Multi-task learning [22]	0.787	—	0.858	0.938
TPE(f.) [80]	0.900	0.813	0.932	—
DR-GAN [25]	0.831	0.699	0.901	0.953
FF-GAN [27]	0.852	0.663	0.902	0.954
NAN [31]	0.921	0.861	0.938	0.960
Multicolumn [21]	0.920	—	—	—
VGGFace2 [81]	0.904	—	—	—
Template Adaptation(f.) [24]	0.939	—	0.928	—
DREAM [15]	0.872	0.712	0.915	0.962
DREAM(E2E+retrain,f.) [15]	0.934	0.836	0.939	0.960
FTL with 60K parameters (o.s.) [33]	0.864	0.744	0.893	0.947
PFEs [32]	0.944	—	—	—
DebFace [82]	0.902	—	—	—
Rotate-and-Render [28]	0.920	0.825	—	—
HPDA [83]	0.876	0.803	0.84	0.88
CDA [84]	0.911	0.823	0.936	0.957
LARNet [30]	0.951	0.874	0.949	0.971
Ours:LARNeXt	**0.955**	**0.891**	**0.965**	**0.979**

2021. These technologies cover template based, GAN, residual learning, 3D reconstruction, and other method systems. They aim at various tasks, including face search, face recognition, face verification, and large pose recognition. All numerical statistics are the best results obtained from original quotation, cross-reference, and experimental reproduction.

1) IJBA Dataset: Verification and Identification Tasks With State of the Arts: In this experiment, we evaluate our method on the challenging benchmark IJBA that covers full pose variation and complies with the original standard protocol. The evaluation metrics include the popular True Acceptance Rate at False Acceptance Rate (TAR@FAR) of 0.01 and 0.001 on the verification task and the Rank-1/Rank-5 recognition accuracy on the identification task. The MS1MV2 dataset and ResNet-50 are used for the training.

Table VI compares our model with various state-of-the-art techniques. LARNet reaches 0.951 (TAR@FAR=0.01) with refinement and end-to-end retrain, whereas our LARNeXt achieves a better performance with 0.955. Both of these two models outperform other methods by a large margin. Our methods also show a significant improvement over the more challenging TAR@FAR=0.001 (0.874 and 0.891 respectively). For face identification, LARNeXt shows an advantage in both Rank-1 (0.965) and Rank-5 (0.979), and also achieves an advanced performance in both recognition and verification.

TABLE VII
QUANTITATIVE EVALUATION ON THE IJB-B/IJB-C DATASET. AND O.S. DENOTES THE OPTIMAL SETTING, WHEREAS F Denotes fine-tuning/refinement. All results are the best-reported ones in original papers or from comparative experiments in other published papers.

Method	TAR@FAR=0.0001	
	IJB-B(%)	IJB-C(%)
CosFace (o.s.) [2]	94.80	96.37
ArcFace (o.s.+f.) [3]	94.25	96.03
CircleLoss [85]	-	93.95
Sub-center Arcface [86]	94.94	96.28
MV-Softmax Loss [87]	93.6	95.2
Curricularface [88]	94.8	96.1
Broadface [89]	94.97	96.38
URface [90]	-	96.6
Groupface [91]	94.93	96.26
DUL [92]	-	94.61
Magface [93]	94.51	95.97
UPL Arcface [94]	95.56	96.76
Our:LARNeXt	**95.72**	**97.26**

To fully reflect the advantages of LARNeXt on the face verification task, we fairly perform comparative experiments on the more challenging IJB-B/IJB-C datasets with the metric TAR@FAR=0.0001. We also provide many competitors' results over the past three years, including CosFace [2], ArcFace [3], CircleLoss [85], Sub-center Arcface [86], MV-Softmax Loss [87], Curricularface [88], Broadface [89], URface [90], Groupface [91], DUL [92], Magface [93], and UPL Arcface [94], which are the best-reported ones in original papers or from comparative experiments in other published papers. Table VII

TABLE VIII
QUANTITATIVE EVALUATION ON THE CFP-FP DATASET. AND O.S

Method	Verification(%)
SphereFace (o.s.+f.)	94.17
CosFace (o.s.)	94.40
ArcFace (o.s.+f.)	94.04
URFace (all modules, o.s.)	98.64
Human-level	98.92
LARNet	99.01
LARNeXt	**99.19**

Denotes the optimal setting, whereas f. Denotes fine tuning/refinement. The MS1MV2 dataset and ResNet-50 are used for the training.

shows that our proposed LARNeXt (95.72/97.26) consistently outperforms these state-of-the-art methods.

2) CFP-FP Dataset: Profile Face Verification Challenge: We employ CFP-FP as our frontal profile face verification dataset with the protocol that the whole dataset is divided into 10 folds, each containing 350 same and 350 not-same pairs of 50 individuals. The MS1MV2 dataset and ResNet-50 are used for the training.

Table VIII shows that the face verification results of state-of-the-art face recognition models are around 94%+. In 2020, the latest universal representation learning face work (URFace) [90], has achieved an astonishing improvement of 98.64% under the auxiliary learning of a large number of modules, such as variation augmentation, confidence-aware identification loss, and multiple embeddings. However, LARNet achieves a 99.01% improvement in 2021, outperforming all of its competitors. As for further advanced LARNeXt, our model achieves a 99.19% performance. To the best of our knowledge, these two are the first to surpass the reported human-level performance (98.92%) on the CFP-FP dataset. LARNeXt reduces the error by approximately 20% under high-precision result of LARNet.

3) LFW, YTF and CPLFW Datasets: General Face Recognition: To further highlight the superiority of our LARNeXt, we conduct an in-depth comparison on general face recognition. The LFW and YTF datasets are the most widely used benchmarks for unconstrained face verification on images and videos. We follow the unrestricted with labelled outside data protocol to report the performance. CPLFW emphasizes pose difference to further enlarge intra-class variance. The CASIA-WebFace dataset and ResNet-50 are used for the training. To achieve a fair comparison with existing methods, we do not report the results of more complex networks (e.g., ArcFace[ResNet-100]).

TABLE IX
QUANTITATIVE EVALUATION ON THE GENERAL FACE RECOGNITION DATASETS: LFW, YTF, AND CPLFW. AND O.S

Method	LFW	YTF	CPLFW
HUMAN-Individual	97.27	-	81.21
HUMAN-Fusion	**99.85**	-	85.24
DeepID [95]	99.47	93.20	-
Deep Face [96]	97.35	91.4	-
VGG Face [97]	98.95	97.30	90.57
FaceNet [98]	99.63	95.10	-
Baidu [99]	99.13	–	-
Center Loss [100]	99.28	94.9	85.48
Range Loss [101]	99.52	93.70	-
Marginal Loss [102]	99.48	94.98	-
SphereFace+(o.s.) [1]	99.47	95.0	90.30
CosFace(o.s.) [2]	99.51	95.1	-
ArcFace(o.s.) [3]	99.53	–	92.08
LARNet [30]	99.61	95.63	93.23
Ours: LARNeXt	99.69	**95.91**	**93.77**

Denotes the optimal setting, whereas f. Denotes fine-tuning/refinement. Symbol '-' indicates that the metric is not available for that protocol. For fairness, the casia-webface dataset and ResNet-50 are used for the training.

TABLE X
QUANTITATIVE RESULTS OF THE MEGA TRAINING DATASET (MTD)

Training Dataset	MS1MV2	MTD
Size (img & id)	5.8M & 86K	16M & 622K
Evaluation		
YTF	0.9591	**0.9614**
CFP-FP	0.9919	**0.9929**
AgeDB-30	0.9773	**0.9830**

Evaluation is conducted on three representative datasets with the Metric ACC.

Table IX shows that because of the very small size of the LFW dataset, almost all methods can achieve a performance of 99+. Although the meaning behind this result is very weak, our method achieves a improvement of 99.69, which is also at the forefront and is only inferior to the HUMAN-Fusion performance of 99.85. We use this comparative result for an extensive study. The same observation is made for the video-sampled dataset YTF, and the results of our LARNeXt remain superior compared over those shown in ResNet-50-based face recognition work. We also introduce a more challenging CPLFW dataset with a large number of poses, and has highly realistic pose intra-class variation considerations. Our LARNeXt also demonstrates its superiority with the improvement of 93.77.

4) Mega Training Dataset: Industrial-Grade Face Recognition Tasks: In this study, we leverage an industrial-grade training dataset called mega training dataset (MTD) to further improve the performance of our approach. The mega training dataset (MTD) that we use is composed of several public datasets and a private face dataset, containing 16,565,811 images from 621,587 identities. To achieve a fair comparison, in our experiments we remove face images belonging to identities that appear in the testing datasets. We select three popular and representative datasets for evaluations, including general face recognition *YTF* dataset, profile face recognition *CFP-FP* dataset and challenging *AgeDB-30* dataset [103], while LFW and its variants cannot be considered due to many overlapping identities with MS1MV2.

The comparative results are reported in Table X, from which we can see that the performance of our approach can be further improved across *ALL* three different testing datasets. Especially for the CFP-FP dataset, the error rates can be reduced by approximately 14% under high-precision result (99.19% → 99.29%).

E. Failure Analysis and Future Improvement

To investigate the limitations of the proposed method, we have traced the original images of the failure cases and found an

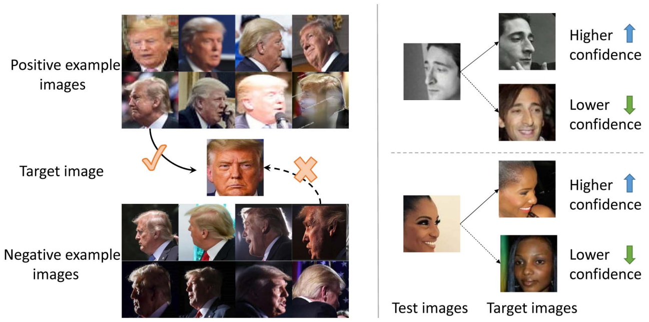
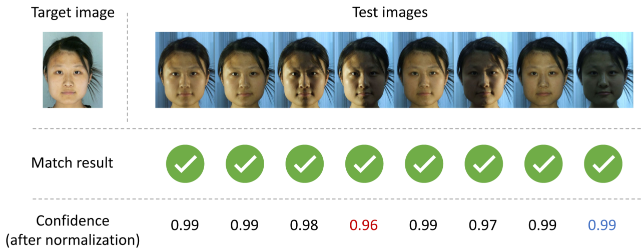


Fig. 10. Case analysis. Left: Highly blurred images with consistent hue are positive examples, while high-resolution images with hue corrupted by pose or lighting are negative examples. Right: Face matches tend to share the same hue, pose, and lighting.



Dataset	#Identities*Images	Verification(%)
Oriental Facial Database	1247 * 8	100

Fig. 11. Quantitative experiments of the hue theory with results for the illumination variation samples on the OFD dataset.

interesting observation: the intrinsic properties of the images can aid face recognition, such as hue - the distribution relationship of pixel RGB values. As shown in Fig. 10, the test images on the top are highly blurred faces and do not look like good quality samples, but it is overjoyed that each of them can accurately match the target image with a high confidence score. But the score is drastically low for the test images on the bottom, which are high-definition profile face images with a large pose or terrible lighting. Moreover, when performing 1:N face verification with the seed library on more samples, we found that at the highest confidence successfully matched samples always tended to share the same hue, pose, or lighting. We speculate that the occlusion caused by the large angle and the shadow brought by the lighting destroy the harmony of the entire facial hue, which is inconsistent with the target image, resulting in a face mismatch.

It is worth noting that the intrinsic hue theory discussed above does not imply that two images with a match must have exactly the same lighting conditions or poses. we have verified that different lighting will slightly change the confidence score rather than have a decisive impact with a simple quantitative experiment on the OFD dataset [60]. It shows that from the illustration (Fig. 11) where the artifacts brought by one-sided light will cause a little interference. Still, the verification results of faces under natural light and in shadow are not different. Therefore, we conclude that the hue distribution of the whole face will influence the accuracy of the face-matching, but the change of lighting or pose will only affect the hue distribution.

Therefore, we believe that the pixel RGB values of the image have some intrinsic properties, which are not affected by uniform transformations externally (e.g., the frontal face image becomes blurrier overall) but are heavily influenced internally (e.g., too much occlusion and shadow from one-sided lighting can destroy this intrinsic nature). By exploring the distribution of intrinsic properties affecting latent space features, we can recover the inner relationship for challenging test samples and alleviate the impact of the large pose angle and terrible lighting. The theoretical research about intrinsic properties may be a new direction for future improvement.

V. Conclusion

We proposed LARNeXt for an enhanced large pose or profile face recognition performance. First, we presented a novel method with Lie algebra theory to explore how face rotation in the 3D space affects the deep feature generation process, and proved that the face rotation in the image space is equivalent to an additive residual component in the deep feature space, which is only determined by the rotation. We also designed three important components, including a residual learning subnet for decoding rotation information from input face images, a soft regression subnet with multi-fusion attention feature aggregation for efficient pose estimation, and a gating control function derived using Lie algebra that learns the rotation magnitude and controls the strength of the residual component contributing to the feature learning process. We present the results of ablation studies to verify the effectiveness of our theory. Compared against various state-of-the-art techniques on the benchmark datasets, our extensive experiments demonstrate the superior performance of our models on frontal-profile face verification, face identification, general face recognition, and industrial-grade tasks.

References

[1] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, “SphereFace: Deep hypersphere embedding for face recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 212–220.

[2] H. Wang et al., “CosFace: Large margin cosine loss for deep face recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 5265–5274.

[3] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, “ArcFace: Additive angular margin loss for deep face recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4690–4699.

[4] D. Gong, Z. Li, D. Lin, J. Liu, and X. Tang, “Hidden factor analysis for age invariant face recognition,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 2872–2879.

[5] H. Wang, D. Gong, Z. Li, and W. Liu, “Decorrelated adversarial learning for age-invariant face recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3527–3536.

[6] Y. Wang et al., “Orthogonal deep features decomposition for age-invariant face recognition,” in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 738–753.

[7] D. Gong, Z. Li, D. Tao, J. Liu, and X. Li, “A maximum entropy feature descriptor for age invariant face recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 5289–5297.

[8] Z. Li, D. Gong, X. Li, and D. Tao, “Aging face recognition: A hierarchical learning model based on local patterns selection,” *IEEE Trans. Image Process.*, vol. 25, no. 5, pp. 2146–2154, May 2016.

[9] Z. Li, D. Gong, Y. Qiao, and D. Tao, “Common feature discriminant analysis for matching infrared face images to optical face images,” *IEEE Trans. Image Process.*, vol. 23, no. 6, pp. 2436–2445, Jun. 2014.

[10] D. Gong, Z. Li, W. Huang, X. Li, and D. Tao, "Heterogeneous face recognition: A common encoding feature discriminant approach," *IEEE Trans. Image Process.*, vol. 26, no. 5, pp. 2079–2089, May 2017.

[11] Z. Li, D. Gong, Q. Li, D. Tao, and X. Li, "Mutual component analysis for heterogeneous face recognition," *ACM Trans. Intell. Syst. Technol.*, vol. 7, no. 3, pp. 1–23, 2016.

[12] D. Gong, Z. Li, J. Liu, and Y. Qiao, "Multi-feature canonical correlation analysis for face photo-sketch image retrieval," in *Proc. 21th ACM Int. Conf. Multimedia*, 2013, pp. 617–620.

[13] Y. Luo, Y. Zhang, J. Yan, and W. Liu, "Generalizing face forgery detection with high-frequency features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 16312–16321.

[14] F. J. Huang, Z. Zhou, H.-J. Zhang, and T. Chen, "Pose invariant face recognition," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2000, pp. 245–250.

[15] K. Cao, Y. Rong, C. Li, X. Tang, and C. C. Loy, "Pose-robust face recognition via deep residual equivariant mapping," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 5187–5196.

[16] I. Masi, S. Rawls, G. Medioni, and P. Natarajan, "Pose-aware face recognition in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 4838–4846.

[17] W. AbdAlmageed et al., "Face recognition using deep multi-pose representations," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2016, pp. 1–9.

[18] L. Song, D. Gong, Z. Li, C. Liu, and W. Liu, "Occlusion robust face recognition based on mask learning with pairwise differential siamese network," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 773–782.

[19] W. Zhang, S. Shan, X. Chen, and W. Gao, "Local Gabor binary patterns based on Kullback - Leibler divergence for partially occluded face recognition," *IEEE Signal Process. Lett.*, vol. 14, no. 11, pp. 875–878, Nov. 2007.

[20] I. Masi, A. T. Tran, T. Hassner, J. T. Leksut, and G. Medioni, "Do we really need to collect millions of faces for effective face recognition?," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 579–596.

[21] W. Xie and A. Zisserma, "Multicolumn networks for face recognition," 2018, *arXiv: 1807.09192*.

[22] X. Yin and X. Liu, "Multi-task convolutional neural network for pose-invariant face recognition," *IEEE Trans. Image Process.*, vol. 27, no. 2, pp. 964–975, Feb. 2018.

[23] T. Hassner et al., "Pooling faces: Template based face recognition with pooled face images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2016, pp. 59–67.

[24] N. Crosswhite, J. Byrne, C. Stauffer, O. Parkhi, Q. Cao, and A. Zisserman, "Template adaptation for face verification and identification," *Image Vis. Comput.*, vol. 79, pp. 35–48, 2018.

[25] L. Tran, X. Yin, and X. Liu, "Disentangled representation learning GAN for pose-invariant face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1415–1424.

[26] J. Wang, J. Zhang, C. Luo, and F. Chen, "Joint head pose and facial landmark regression from depth images," *Comput. Vis. Media*, vol. 3, no. 3, pp. 229–241, 2017.

[27] X. Yin, X. Yu, K. Sohn, X. Liu, and M. Chandraker, "Towards large-pose face frontalization in the wild," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 3990–3999.

[28] H. Zhou, J. Liu, Z. Liu, Y. Liu, and X. Wang, "Rotate-and-render: Unsupervised photorealistic face rotation from single-view images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 5911–5920.

[29] O. Tuzel, F. Porikli, and P. Meer, "Learning on lie groups for invariant detection and tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2008, pp. 1–8.

[30] X. Yang, X. Jia, D. Gong, D.-M. Yan, Z. Li, and W. Liu, "LARNet: Lie algebra residual network for face recognition," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 11738–11750.

[31] J. Yang et al., "Neural aggregation network for video face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4362–4371.

[32] Y. Shi and A. K. Jain, "Probabilistic face embeddings," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 6902–6911.

[33] X. Yin, X. Yu, K. Sohn, X. Liu, and M. Chandraker, "Feature transfer learning for face recognition with under-represented data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5704–5713.

[34] W. Rossmann, *Lie Groups: An Introduction Through Linear Groups*. London, U.K.: Oxford Press, 2002.

[35] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[36] D. W. McKenzie, *An Elementary Introduction to Lie Algebras for Physicists*. Ithaca, NY, USA: Cornell University, 2015.

[37] M. P. D. Carmo, *Riemannian Geometry*. Cambridge, MA, USA: Birkhäuser, 1992.

[38] O. Rodriguez, "Des lois geometriques qui regissent les desplacements d'un systeme solide dans l'espace et de la variation des coordonnees provenant de deplacements consideres independamment des causes qui peuvent les produire," *J Mathematiques Pures Appliquees*, vol. 5, pp. 380–440, 1840.

[39] R. Wulf, *Lie Groups–An Introduction Through Linear Groups.*Oxford Graduate Texts in Mathematics, Oxford, U.K.: Oxford Science, 2002.

[40] H. Brian, *Lie Groups, Lie Algebras, and Representations: An Elementary Introduction*. Berlin, Germany: Springer, 2015.

[41] M. Wilhelm, "On the exponential solution of differential equations for a linear operator," *Commun. Pure Appl. Math.*, vol. 7, no. 4, pp. 649–673, 1954.

[42] N. Jacobson, *Lie Algebras*. Hoboken, NJ, USA: Wiley, 1966.

[43] A. M. Lyapunov, "The general problem of the stability of motion," *Int. J. Control*, vol. 55, no. 3, pp. 531–534, 1992.

[44] N. P. Bhatia and G. P. Szegö, *Stability Theory of Dynamical Systems*. Berlin, Germany: Springer, 2002.

[45] A. M. Saxe, J. L. McClelland, and S. Ganguli, "Exact solutions to the nonlinear dynamics of learning in deep linear neural networks," in *Proc. Int. Conf. Learn. Representations*, 2014, pp. 1–14.

[46] R.K. Srivastava, K. Greff, and J. Schmidhuber, "Highway networks," in *Proc. Int. Conf. Mach. Learn. Deep Learn. Workshop*, 2015, pp. 1–6.

[47] D. Balduzzi, M. Frean, L. Leary, J. Lewis, K. W.-D. Ma, and B. McWilliams, "The shattered gradients problem: If ResNets are the answer, then what is the question?," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 342–350.

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

[49] T.-Y. Yang, Y.-H. Huang, Y.-Y. Lin, P.-C. Hsiu, and Y.-Y. Chuang, "SSR-Net: A compact soft stagewise regression network for age estimation," in *Proc. Int. Joint Conf. Artif. Intell.*, 2018, pp. 1078–1084.

[50] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: CNN architecture for weakly supervised place recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 5297–5307.

[51] S. Sabour and N. F. G. E. Hinton, "Dynamic routing between capsules," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 3859–3869.

[52] T.-Y. Yang, Y.-T. Chen, Y.-Y. Lin, and Y.-Y. Chuang, "FSA-Net: Learning fine-grained structure aggregation for head pose estimation from a single image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 1087–1096.

[53] 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 *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2016, pp. 1–9.

[54] V. Sitzmann, J. N. P. Martel, A. W. Bergman, D. B. Lindell, and G. Wetzstein, "Implicit neural representations with periodic activation functions," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2020, Art. no. 33.

[55] Y. Wu, J. Li, Y. Kong, and Y. Fu, "Deep convolutional neural network with independent softmax for large scale face recognition," in *Proc. 24th ACM Int. Conf. Multimedia*, 2016, pp. 1063–1067.

[56] 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 *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2439–2448.

[57] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.

[58] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "MS-Celeb-1 M: A dataset and benchmark for large-scale face recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 87–102.

[59] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," 2014, *arXiv:1411.7923*.

[60] "Oriental facial database institution of artificial intelligence," XJTU. 2016.

[61] B. F. Klare et al., "Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus Benchmark A," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1931–1939.

[62] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Univ. Massachusetts, Amherst, Tech. Rep. 07-49, Oct., 2007.

[63] L. Wolf, T. Hassner, and I. Maoz, “Face recognition in unconstrained videos with matched background similarity.,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 529–534.

[64] T. Zheng and W. Deng, “Cross-pose LFW: A database for studying cross-pose face recognition in unconstrained environments,” Beijing Univ. Posts, Tech. Rep. 18–01, Feb. 2018.

[65] Z. Liu, P. Luo, X. Wang, and X. Tang, “Deep learning face attributes in the wild,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 3730–3738.

[66] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li, “Face alignment across large poses: A 3D solution,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 146–155.

[67] G. Fanelli, T. Weise, J. Gall, and L. V. Gool, “Real time head pose estimation from consumer depth cameras,” in *Proc. Joint Pattern Recognit. Symp.*, 2011, pp. 101–110.

[68] C. Szegedy et al., “Going deeper with convolutions,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.

[69] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[70] K. Alex, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[71] V. Kazemi and J. Sullivan, “One millisecond face alignment with an ensemble of regression trees,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1867–1874.

[72] A. Bulat and G. Tzimiropoulos, “How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks),” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 1021–1030.

[73] A. Kumar, A. Alavi, and R. Chellappa, “KEPLER: Keypoint and pose estimation of unconstrained faces by learning efficient H-CNN regressors,” in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2017, pp. 258–265.

[74] N. Ruiz, E. Chong, and J. M. Rehg, “Fine-Grained head pose estimation without keypoints,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2018, pp. 2074–2083.

[75] X. Yang, X. Jia, M. Yuan, and D.-M. Yan, “Real-time facial pose estimation and tracking by coarse-to-fine iterative optimization,” *Tsinghua Sci. Technol.*, vol. 25, no. 5, pp. 690–700, 2020.

[76] J. Guo, X. Zhu, Y. Yang, F. Yang, Z. Lei, and S. Z. Li, “Towards fast, accurate and stable 3D dense face alignment,” in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 152–168.

[77] V. Albiero, X. Chen, X. Yin, G. Pang, and T. Hassner, “img2pose: Face alignment and detection via 6DoF, face pose estimation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 7617–7627.

[78] D. Wang, O. Charles, and K. J. Anil, “Face search at scale,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1122–1136, Jun. 2017.

[79] J.-C. Chen, V. M. Patel, and R. Chellappa, “Unconstrained face verification using deep CNN features,” in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2016, pp. 1–9.

[80] S. Sankaranarayanan, A. Alavi, C. D. Castillo, and R. Chellappa, “Triplet probabilistic embedding for face verification and clustering,” in *Proc. IEEE Int. Conf. Biometrics Theory Appl. Syst.*, 2017, pp. 1–8.

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

[82] S. Gong, X. Liu, and A. K. Jain, “Jointly De-biasing face recognition and demographic attribute estimation,” in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 330–347.

[83] Q. Wang, T. Wu, H. Zheng, and G. Guo, “Hierarchical pyramid diverse attention networks for face recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 8326–8335.

[84] M. Wang and W. Deng, “Deep face recognition with clustering based domain adaptation,” *Neurocomputing*, vol. 393, no. 1, pp. 1–14, 2020.

[85] Y. Sun et al., “Circle loss: A unified perspective of pair similarity optimization,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 6398–6407.

[86] J. Deng, J. Guo, T. Liu, M. Gong, and S. Zafeiriou, “Sub-center arcface: Boosting face recognition by large-scale noisy web faces,” in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 741–757.

[87] X. Wang, S. Zhang, S. Wang, T. Fu, H. Shi, and T. Mei, “Mis-classified vector guided softmax loss for face recognition,” in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 12241–12248.

[88] Y. Huang et al., “Curricularface: Adaptive curriculum learning loss for deep face recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 5901–5910.

[89] Y. Kim, W. Park, and J. Shin, “BroadFace: Looking at tens of thousands of people at once for face recognition,” in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 536–552.

[90] Y. Shi, X. Yu, K. Sohn, M. Chandraker, and A. K. Jain, “Towards universal representation learning for deep face recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 6817–6826.

[91] Y. Kim, W. Park, M.-C. Roh, and J. Shin, “GroupFace: Learning latent groups and constructing group-based representations for face recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 5621–5630.

[92] J. Chang, Z. Lan, C. Cheng, and Y. Wei, “Data uncertainty learning in face recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 5710–5719.

[93] Q. Meng, S. Zhao, Z. Huang, and F. Zhou, “MagFace: A universal representation for face recognition and quality assessment,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 14225–14234.

[94] J. Deng, J. Guo, J. Yang, A. Lattas, and S. Zafeiriou, “Variational prototype learning for deep face recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 11906–11915.

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

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

[97] O. M. Parkhi, A. Vedaldi, and A. Zisserman, “Deep face recognition,” in *Proc. Brit. Mach. Vis. Assoc.*, 2015, pp. 1–12.

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

[99] J. Liu, Y. Deng, T. Bai, Z. Wei, and C. Huang, “Targeting ultimate accuracy: Face recognition via deep embedding,” in 2015, *arXiv:1506.07310*.

[100] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, “A discriminative feature learning approach for deep face recognition,” in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 499–515.

[101] X. Zhang, Z. Fang, Y. Wen, Z. Li, and Y. Qiao, “Range loss for deep face recognition with long-tailed training data,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 5409–5418.

[102] J. Deng, Y. Zhou, and S. Zafeiriou, “Marginal loss for deep face recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2017, pp. 60–68.

[103] S. Moschoglou, A. Papaioannou, C. Sagonas, J. Deng, I. Kotsia, and S. Zafeiriou, “AgeDB: The first manually collected, in-the-wild age database,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2017, pp. 51–59.

Xiaolong Yang received the BS degree in information and computing science from Northwestern Polytechnical University, in 2017. He is currently working toward the PhD degree with the Key Laboratory of Mathematics Mechanization, Academy of Mathematics and Systems Sciences, Chinese Academy of Sciences(CAS). His research interests is on computer graphics and computer vision.

Xiaohong Jia received the bachelor’s and PhD degrees from the University of Science and Technology of China, in 2004 and 2009, respectively. She is currently a professor with the Key Laboratory of Mathematics Mechanization, Academy of Mathematics and Systems Science, Chinese Academy of Sciences (CAS). Her research interests include computer graphics, computer aided geometric design, and computational algebraic geometry.

Dihong Gong received the PhD degree of computer science from the University of Florida, in 2018. He then joined the Tencent AI Lab as a senior research scientist. His research interests primarily focused on face related technologies, including face detection, recognition, and liveness examination.

Dong-Ming Yan (Member, IEEE) received the bachelor's and master's degrees in computer science and technology from Tsinghua University, in 2002 and 2005, respectively, and the PhD degree in computer science from Hong Kong University, in 2010. He is a professor in State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS) and the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences. His research interests include computer graphics, computer vision, and pattern recognition.

Zhifeng Li (Senior Member, IEEE) is currently a top-tier principal research scientist with Tencent. Before joining Tencent, he was a full professor with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences. He was one of the most cited Chinese researchers (Elsevier-Scopus) in computer science and technology for the years 2020, 2021, and 2022. He is currently serving on the editorial boards of Pattern Recognition, *IEEE Transactions on Circuits and Systems for Video Technology, and Neurocomputing*. He is a fellow of British Computer Society (FBCS).

Wei Liu (Fellow, IEEE) received the PhD degree in electrical engineering and computer science from Columbia University, in 2012. He is currently a distinguished scientist of Tencent and the director of Ads Multimedia AI with Tencent Data Platform. Prior to that, he has been a research staff member of IBM T. J. Watson Research Center, USA from 2012 to 2015. He has long been devoted to fundamental research and technological development in core fields of AI, including deep learning, machine learning, reinforcement learning, computer vision, information retrieval, Big Data, etc. To date, he has published extensively in these fields with more than 280 peer-reviewed technical papers, and also issued more than 30 US patents. He currently serves on the editorial boards of internationally leading AI journals like *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, *IEEE Transactions on Neural Networks and Learning Systems (TNNLS)*, and *IEEE Intelligent Systems*. He is an area chair of top-tier computer science and AI conferences, e.g., NeurIPS, ICML, IEEE CVPR, IEEE ICCV, IJCAI, and AAAI. He is a Fellow of the IAPR and IMA, and an Elected Member of the ISI.