# Local Feature Enhancement Network for Set-based Face Recognition

Ziyi Bai[1,2], Ruiping Wang[1,2,3], Shiguang Shan[1,2], Xilin Chen[1,2]
[1] Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing, 100190, China
[2] University of Chinese Academy of Sciences, Beijing, 100049, China
[3] Beijing Academy of Artificial Intelligence, Beijing, 100084, China
ziyi.bai@vipl.ict.ac.cn, {wangruiping, sgshan, xlchen}@ict.ac.cn

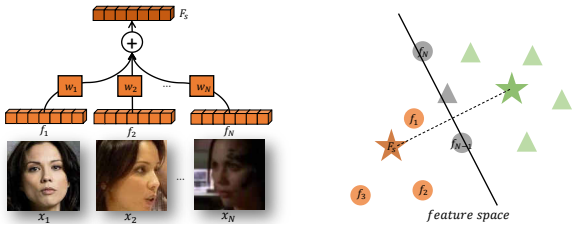
2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021) | 978-1-6654-3176-7/21/$31.00 ©2021 IEEE | DOI: 10.1109/FG52635.2021.9667020

***Abstract*— Set-based Face Recognition is widely applied in scenarios like law enforcement and online media data management. Compared with face recognition using a single image, the faces in the set often contain abundant appearance changes. Therefore, how to make full use of the rich information from the set and integrate them into a unified set representation become the key to set-based face recognition. Inspired by the fact that humans usually complete this fine-grained task through integrating the information from the congruent local regions (*e.g.* an eye to an eye) of multiple faces in a set, we propose a novel method called Local Feature Enhancement Network (LFENet), which can automatically enhance the local feature through transferring the local information across the images. Specifically, we retain the spatial semantic information of the feature maps and apply different relational functions to establish the correlation among the local features. The contained local information will be transferred to the relevant local features to enhance their discriminability. By doing so, the valuable local information carried in some local features can complement those with incomplete information. Besides, the various local information is aligned across faces under different conditions to help the model learn intra-set-compact face representations. Our method achieves state-of-the-art performances on two mainstream set-based face recognition benchmarks: IJB-A and IJB-C, which fully reflects the rationality and effectiveness of our local feature enhancement mechanism.**

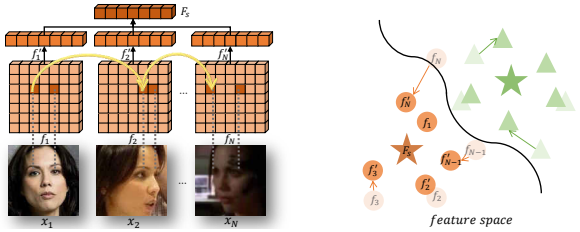
## I. INTRODUCTION

Face recognition based on single image [28], [31], [25], [30], [37], [15], [34], [33], [4], [29] has been well studied in these years. However, in some real world scenarios, a set of images gathered from various sources need to be compared at the same time. Although the diversity of data brings the richness of face information, it is exhausted to measure the similarities of every pair of images [45]. Therefore, integrating the discriminative information of each image within the set to get a unified set-level representation is necessary for the set-based face recognition.

Nevertheless, the image sets usually suffer from large intra-set variance for images collected from different sources have their own characteristics. Still images from the website include a variety of subject attributes changes, such as pose, expression, make-up, *etc*, while video data are often of lower quality due to the influence of illumination, compression and occlusion. Thus the set-based data with the mixture of still faces and video faces is more challenging for the face recognition model.




(a) set modeling based on global discriminability




(b) set modeling through local feature enhancement

Fig. 1: Comparison of two different set modeling methods. The left part of each subfigure shows the process of set modeling to get the unified set representation $F_s$ while the right side shows the corresponding discriminant feature space. (a) represents set modeling by quality score weighted sum; (b) represents set modeling by learning intra-set-compact image features from $f$ to $f'$ through transferring semantic consistent local information (indicated with the yellow lines). Best viewed in color.

Obviously, simply applying average pooling on image features obtained by the single image face recognition model will introduce noise to set representation [16], [43], [3]. To alleviate the effects of the large variations within the image set, existing set-based face recognition methods [40], [18], [17], [43] usually focus on exploiting those samples of high quality while suppress those of low quality for set modeling to reduce the impact of noise on the unified set representation. Specifically, the face representations are aggregated with the weight of the global image quality score indicating the sample discriminability which is shown in Fig.1(a). However, we cannot expect that the image set always contains samples with high quality (*i.e.* strong

This work is partially supported by National Key R&D Program of China (2020AAA0105200), Natural Science Foundation of China under contracts Nos. U19B2036, 61922080, 61772500, and Aerospace Information Research Institute of Chinese Academy of Sciences Project No. GXTC-A1-21630166.

978-1-6654-3176-7/21/$31.00 ©2021 IEEE

Authorized licensed use limited to: INSTITUTE OF COMPUTING TECHNOLOGY CAS. Downloaded on January 13,2022 at 06:28:43 UTC from IEEE Xplore. Restrictions apply.

discriminability). Although the images of low quality are often far away from the set center in the feature space, they can serve as hard samples to guide model learning more robust feature representation [25], [36]. Thus, how to fully utilize those low-quality images instead of just suppressing them is more appealing to the set-based face recognition in real scenes.

Intuitively, compared with focusing on certain samples in the image set, humans will continuously observe the congruent local regions of all faces and integrate the information in these regions to give an overall impression of the person. Motivated by this fact, we propose a novel method called *Local Feature Enhancement Network* (LFENet) to deeply exploit the interaction of the abundant information contained in local regions through local information transferring. As shown in Fig.1(b), we maintain the spatial structure of the feature maps to extract local features of faces instead of directly performing average pooling on them. We use relational learning to establish the correlation between local features. The face information distilled from each local feature will be transferred to its related local features to replenish and enrich their information. By transferring the local information we enhance the local features from two aspects: Firstly, the feature with local information missing (occlusion) or of low quality (blur) is improved. Secondly, various local information (different poses/expressions) in multiple images is aligned. Accordingly, the discriminability of all face representations (especially those of low quality) in the image set is improved, and these representations become compact to each other at the same time.

To conclude, our proposed LFENet can make full use of the rich information contained in the image set: 1) *Fine-grained.* We consider a more refined way of information integration, which keeps the spatial semantics of feature maps and aggregates the related local information from multiple faces. 2) *Complementary.* Through information transfer, the valuable information across the set can complement each other, and the various information is aligned across faces under different conditions. 3) *Integral.* Each sample in the set including the images with low quality plays the same role in set modeling, which enables the model to learn robust face representations based on comprehensive information. Extensive experiments on the benchmarks for image set-based face recognition have shown the effectiveness of our proposed method.

## II. RELATED WORK

### A. Deep Face Recognition

The ongoing research aiming at single image face recognition have made great achievements [4], [15], [25], [30], [34], [29] with the deep neural network. These works focus on how to extract robust face features to tackle the large intra-class variation through the deep metric learning. When directly applying these methods to set-based face recognition, *i.e.*, extracting features from all images and then aggregating them with the average/max pooling, the performance is often limited [31], [14], [3]. Treating all the images uniformly will introduce much noise into the overall set representation [40], [6], [16]. To tackle this problem, some set-based face recognition methods [8], [40], [18], [24], [17], [19], [38], [45], [6] are devoted to designing more elaborate set modeling modules.

### B. Set Modeling

Intuitively, humans often give an overall impression of a person by observing from different perspectives. Similarly, researchers propose to get a unified subject representation (set modeling) through the information aggregation process carried on image level and component level respectively.

*Image-level aggregation.* This type of works [8], [40], [18], [24], [17], [43] regard each image as a point in the feature space and conduct set modeling at the image level. [40], [18], [17], [43] consider that the images with stronger discriminability (higher quality) contain more key information. Therefore the high-quality samples are dominant in learning the set representation while the low-quality images are suppressed. [40], [18] learn quality scores for each element in the set and perform quality weighted sum on image representations. [17] uses reinforcement learning to determine the importance of the image. In [43], images that are close to the class center and far from others are considered as representative samples. Faces in the same bins which are divided based on the pose and quality are pooled before the feature extraction in [8]. [24] utilizes adversarial learning to generate synthesized images as the prototype of the set for faster matching.

*Component-wise aggregation.* To solve a fine-grained problem like face recognition, performing information aggregation among face components (local regions) is necessary. [19] encodes second-order statistics information of local features to further improve the performance of the model, but the exploited covariance matrix loses the original spatial semantic information. [38] assumes that comparing the features of the same local region (landmarks) is important for face verification, which can be obtained by landmark attention map from detection. [45] extends the local aggregation feature VLAD [1] to face recognition task. [6] refines the quality weighted sum from image-level to component-wise to further reduce noise information.

### C. Relational Learning

Although some methods perform set modeling at the component level, few works pay attention to the correlation of them across the set. Relational learning is proposed to model the relevance between entities. [35], [46] model the long-distance dependency across the video content through Relational Learning. In Natural Language Processing, the self-attention mechanism is proposed to learn contextual information and capture the internal structure of the sentence [32], [5], [23], [13]. Inspired by these works, we design a local information transfer module by performing relational learning on local features of multiple images to establish their correlations. Then the related information can be transferred across the set to complement each other.

Authorized licensed use limited to: INSTITUTE OF COMPUTING TECHNOLOGY CAS. Downloaded on January 13,2022 at 06:28:43 UTC from IEEE Xplore. Restrictions apply.

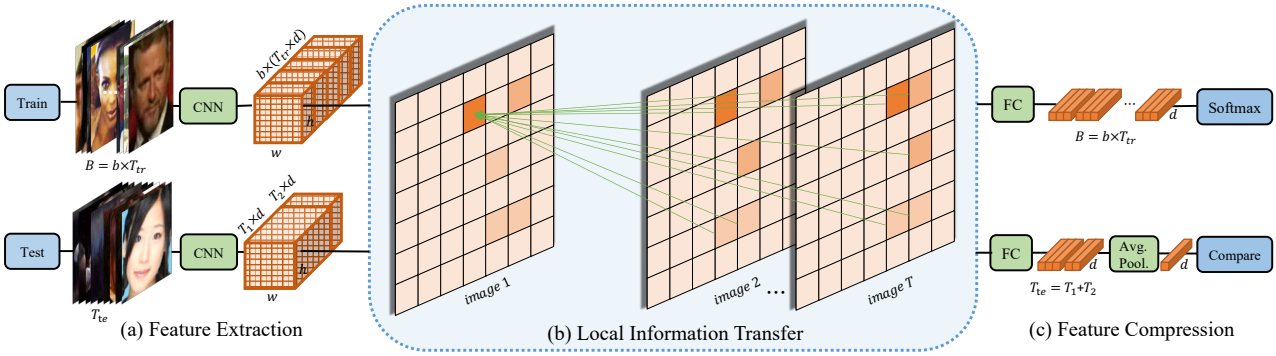



Fig. 2: The framework of our proposed Local Feature Enhancement Network (LFENet) for set-based face recognition. LFENet contains three stages: (a) At the feature extraction stage, $b$ image sets with $T_{tr}$ images in each (for training) or an image set with size $T_{te}$ (for testing) are input to the CNN to get their corresponding feature representations; (b) Local information transfer module takes the split-up feature blocks ($F \in \mathbb{R}^{T\times d\times h\times w}$ for each set/segment) as input and performs the information transfer (indicated with green lines) among the related local features with similar semantic information (highlighted in the feature maps) across the set; (c) The refined feature blocks will pass the feature compression layer to obtain the embedding of each image during the training phase and the unified set representation for testing.

## III. PROPOSED METHOD

In this section, we will first give a brief overview of our proposed LFENet and then introduce each part of it in detail.

### A. Overview

The set-based face recognition model takes a set of images $X = \{x_i\}_{i=1}^{T}$ of the same identity $y$ as input, where $T$ is the size of the set. And it is required to learn a unified fixed-size set representation $F_s$ of the set with any size for face identification or face verification. Our proposed *Local Feature Enhancement Network* (LFENet) mainly considers the face verification task which needs to compare the similarity of representations from two face sets and determine whether these two sets come from the same person (with the same set label $y$).

As shown in Fig.2, the network consists of three parts: 1) feature extraction, which extracts feature with local semantic information of each image in a batch/set by a Convolutional Neural Network (CNN), 2) local information transfer (LIT), which is the key module in the LFENet. The LIT module first builds the correlation among local features from different images in the set through relational function. Then the distilled local information from each local region is transferred to the related local features, and 3) feature compression, which will compress the high dimensional feature representation into a compact vector. Note that the average pooling across the set is not applied during the training stage.

### B. Feature Extraction

Any existing CNN can be used as the backbone network in our framework. To preserve the local semantic information of the feature map, we truncate the networks before the global average pooling layer.

We train the model in a mini-batch manner. $b$ image sets which contain $T_{tr}$ images in each are randomly chosen in a batch, thus each batch contains $B$ images ($B = b \times T_{tr}$). For each image, the output of the last convolution layer is an $h \times w \times d$ tensor with spatial height $h$, spatial width $w$, channel $d$. Before feeding a mini-batch of image features to the LIT module, they will be split into $b$ four-dimensional feature blocks $F \in \mathbb{R}^{T_{tr}\times h\times w\times d}$, each block for one set.

During testing, the input image sets can be of any size. Since set size can be quite large, we set a threshold $\theta$ to speed up calculation. The set whose size exceeds the threshold will be split into $K$ segments, where $K = \lceil T/\theta \rceil$, the size of each segment is $\theta$, except that the size of the last segment is $T-(K-1)\theta$. Note that the feature blocks of each segment are fed to the LIT module in turn.

### C. Local Information Transfer

Because of the huge appearance changes within the image set, the extracted features of the image set contain overwhelmed local information. To effectively integrate all discriminative information within the set, we propose a plug-and-play module called the LIT module which can lead the model to find the related local feature with similar semantic information from different images and transfer this information across the whole set.

Specifically, we first utilize relational learning to obtain the similarity between local features and then use the self-attention mechanism to transfer the local information across the related regions. As shown in Fig.3, there are altogether $N = T \times h \times w$ local features in the block. For each local feature $l$ with the dimensionality $d$, we regard it as the query and use $Q(l; W_Q)$ to distill the queried local feature $q$. Besides, a pair of functions $K(l, W_K)$ and $V(l; W_V)$ are used to separately extract the key $k$ and the value $v$ of all local regions in the feature block, where the key is used to calculate the relevance with query region and the value is the distilled information. $W_Q, W_K, W_V$ are all learnable parameters. The relevance of any two local regions $\alpha$ and $\beta$ across the whole set is reflected by the attention map $A \in \mathbb{R}^{N\times N}$, which is computed by performing the relational

Authorized licensed use limited to: INSTITUTE OF COMPUTING TECHNOLOGY CAS. Downloaded on January 13,2022 at 06:28:43 UTC from IEEE Xplore. Restrictions apply.

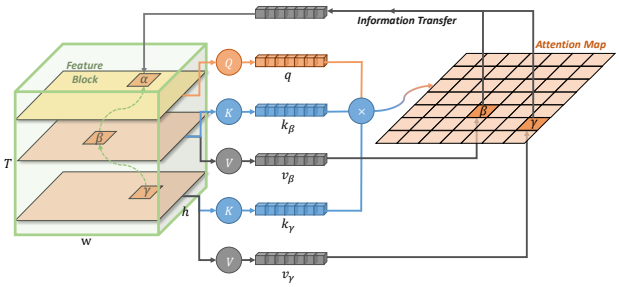



Fig. 3: Our proposed local information transfer module. Each layer of the feature block represents the image feature extracted by the backbone network. Given a query local region $\alpha$, we compute its corresponding attention map to find the related local regions: $\beta$ and $\gamma$, *etc*. Then the distilled information $v$ from these parts will be transferred to the query region.

function $G(.)$ on their local features:

$$A_{\alpha\beta} = G(Q(l_\alpha), K(l_\beta)). \tag{1}$$

The attention score in $A$ represents the relevance between the local features, only the most related features have the high attention score.

Inspired by previous relational learning works [32], [21], [35], we design three kinds of relational functions as follows to verify that LIT module is not sensitive to these choices: *SoftmaxA*: directly multiplying the original local feature of two regions and normalizing the result with softmax function,

$$G(l_\alpha, l_\beta) = softmax(l_\alpha^T l_\beta). \tag{2}$$

*SoftmaxB*: multiplying the distilled local features (extracted by $Q(l)$, $K(l)$) of two regions,

$$G(q, k) = softmax(q^T k). \tag{3}$$

*Scaled Dot*: multiplying the distilled local feature and scaling the result by dividing the number of regions,

$$G(q, k) = \frac{1}{N}(q^T k). \tag{4}$$

Then the distilled related information is superposed to the original local feature using the residual mechanism,

$$l'_\alpha = l_\alpha + \sum_{\beta=1}^{N} A_{\alpha\beta} V(l_\beta). \tag{5}$$

Hence, each local feature is enhanced with the information from the strongly related regions and those irrelevant noise information will not participate in the process of information transfer.

In this way, we can establish the correlation of each local feature across the image set, and let the distilled local information complement each other through information transfer. Specifically, the local features with valuable local information will replenish the ones who suffer from the loss of key information; The local features of congruent components from faces with different poses will be aligned through sharing their local information. Thus each local feature will be enhanced, which will further boost the discriminability of the image representation. Besides, the face representation within the set will become compact at the same time. Finally, we feed the refined feature block $F' \in \mathbb{R}^{T\times d\times h\times w}$ into the feature compression module.

### D. Feature Compression

The dimensionality of the set feature block is relatively high, which is not conducive to the subsequent calculation and storage. Therefore, we compress the feature block to obtain a more compact representation. Spatial dimensionality reduction of each image is performed to get the face embedding $f' \in \mathbb{R}^d$ via BN [11]-Dropout [27]-FC-BN following the previous work [4]. Besides, since the local features in the set have been already aligned in the LIT module, we simply use the average pooling on each face embedding to get the compact vector $F_s = \frac{1}{T_{te}} \sum_{i=1}^{T_{te}} f'_i$ as the set-level representation for testing.

What's more, most set-based learning methods [40], [18], [38], [16] directly supervise the set-level embedding in the feature space during the training phase. This unduly relaxed constraint can not force the model to learn robust feature representation for each image in the set, as shown in Fig.1(a). Furthermore, to ensure that the LIT module can transfer valuable local information of faces to improve the discriminability of each sample within the set, we calculate the classification loss on all face embeddings rather than at the set level. In this way, the model can learn intra-set-compact face representations at the same time.

## IV. EXPERIMENTS

In this section, we will evaluate the performance of the proposed LFENet on set-based face recognition task. We first introduce the datasets and evaluation protocols, and then the implementation details are presented. Next, we compare LFENet with the state-of-the-art methods on several mainstream set-based face recognition benchmarks. Finally, we explore the effectiveness of the LIT module through the ablation experiments.

### A. Datasets and Evaluation Protocols

Following the previous works [17], [16], [43], we employ the still image dataset—VGGFace2 [2] as the training data. It contains 3.3M images of 8,631 subjects with rich face variations within the same subject. To let the model learn from set-based data, we divide the images that belong to the same subject into multiple unordered image sets. Each set includes $T_{tr}$ randomly selected images. Note that requirement of the model's ability to establish correlation is higher when $T_{tr}$ becomes larger.

As for the test dataset, we use mainstream set-based face recognition evaluation datasets: IARPA Janus Benchmark A [12] (IJB-A) and IARPA Janus Benchmark C [20] (IJB-C). Both of them collect images captured under unconstrained environments which show large variations in image quality (e.g. low-resolution and motion-blur) and subject state (e.g. pose, expression, accessories).

Authorized licensed use limited to: INSTITUTE OF COMPUTING TECHNOLOGY CAS. Downloaded on January 13,2022 at 06:28:43 UTC from IEEE Xplore. Restrictions apply.

TABLE I: Evaluation of the 1:1 verification protocol on IJB-A dataset (higher is better).

| Method | Backbone | Training Data | 1:1 Verification TAR | | |
|---|---|---|---|---|---|
| | | | FAR=1e-3(%) | FAR=1e-2(%) | FAR=1e-1(%) |
| NAN [40] | GoogleNet | Crawled(3M) | 88.10±1.10 | 94.10±0.80 | 97.80±0.30 |
| QAN [18] | GoogleNet | Ext. VGGFace2(5M) | 89.31±3.92 | 94.20±1.53 | 98.02±0.55 |
| DAC [17] | GoogleNet | Crawled(3M) | - | 95.40±0.10 | 98.10±0.80 |
| Multicolumn [39] | ResNet-50 | VGGFace2(3.3M) | 92.00±1.30 | 96.20±0.50 | 98.90±0.20 |
| GhostVLAD [45] | SENet-50 | VGGFace2(3.3M) | 93.05±1.60 | 97.20±0.50 | 99.00±0.20 |
| C-FAN [6] | Face-ResNet | MS1M(10M) | 91.59±0.99 | 93.97±0.78 | - |
| PIFR [16] | ResNet-50 | VGGFace2(3.3M) | 95.50±1.00 | 98.30±0.40 | 99.30±0.30 |
| PFE [26] | 64CNN | Web.+MS1M(4.4M) | 95.25±0.89 | 97.50±0.43 | - |
| Baseline(Avg.) | ResNet-34 | VGGFace2(3.3M) | 80.68±4.27 | 92.91±1.40 | 96.76±0.02 |
| **LFENet(Ours)** | ResNet-34 | VGGFace2(3.3M) | 87.68±3.44 | 95.25±1.09 | 98.40±0.42 |
| Baseline(Avg.) | SENet-50 | VGGFace2(3.3M) | 93.62±0.84 | 95.32±0.76 | 97.45±0.44 |
| **LFENet(Ours)** | SENet-50 | VGGFace2(3.3M) | **96.83±0.65** | **98.93±0.39** | **99.57±0.15** |

IJB-A dataset proposes the concept of 'template matching'. Each template is composed of a mixture of still images and video frames from the same subject. We regard the template as an unordered image set. The entire dataset contains 5,712 still images and 20,414 video frames of 500 subjects. The benchmark provides two protocols: 1:1 face verification and 1:N face identification, which are all template-based. We only focus on the former protocol where 10-fold testing is conducted. Unlike the traditional closed-set classification problem, face verification is an open-world task, where face representations of high discriminability are required. We report the true acceptance rate (TAR) at different false acceptance rates (FARs). IJB-C dataset extends the IJB-A to 21,294 images and 11,779 videos of 3,531 subjects, which brings more challenges to the face recognition model.

### B. Implementation Details

*Pre-processing:* Before feeding the images to the network, we first detect and align faces by applying the MTCNN algorithm [42] in both training and testing datasets. For the robustness of the model, the bounding box is extended by a factor of 0.3. We follow the recent works [4], [43] to generate the normalized face crops with the size of $112 \times 112$, and faces are randomly horizontal flipped with the probability of 0.5 during training.

*Training:* All experiments in this paper are implemented by PyTorch [22] and are conducted on four GeForce GTX Titan X GPUs. We use ResNet-34 [9] and SENet-50 [10] respectively in this paper as the backbone network. To train LFENet based on ResNet-34, we first train the original network on the VGGFace2 dataset from scratch and then finetune the LFENet with the learning rate of 0.001 on the backbone, 0.05 on the LIT module. The learning rate is divided by 10 every 3 epochs. As for SENet-50, we directly use the pre-trained model provided by [4], which has been trained on the cleaned MS1M dataset [7], and then finetune on our LFENet, as this can considerably accelerate the training. During the finetuning, we set the learning rate to start from $1e-4$ on the backbone network, 0.01 on the LIT module and divided by 10 every epoch. For both backbones, we set the momentum to 0.9 and weight decay to $5e-4$. What's more, to let the LIT module learn from data composed of several complete image sets in batches at the training stage, the batch size $B$ should be an integral multiple of the image set size. Different image set sizes are selected to train the ability of the LIT module to extract the correlation of the local features across the set in the data preparation stage. Besides, we simply use the softmax function to compute cross-entropy loss for all samples within the set to force the model to learn intra-set-compact face representations. Thus the training process becomes equivalent to training on the single sample in a mini-batch manner.

*Testing:* In the test phase, the LFENet is applied to the testing datasets without finetuning. The images in each set will be sent to the network to calculate the corresponding face embeddings. To achieve efficient computing, we set the set size threshold $\theta$ to 12, as mentioned in Sec.III-B. The larger image sets will contain too much redundant information, so it is not necessary to transfer all such information across the set.

### C. Comparison with State-of-the-art Methods

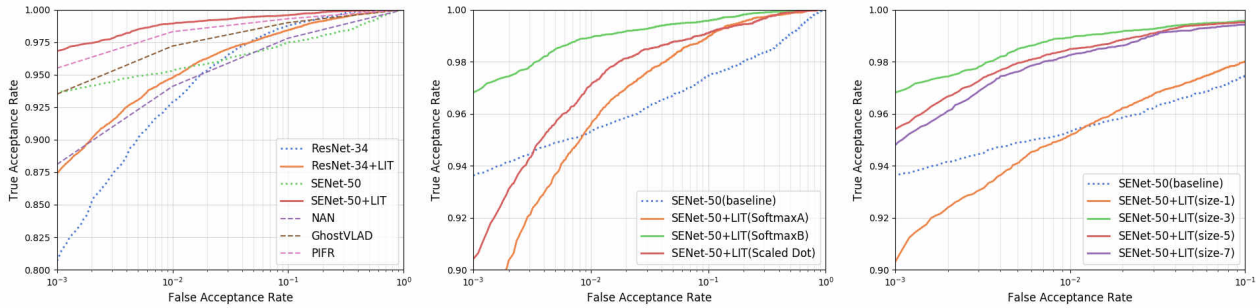
Our proposed LFENet based on two backbones—ResNet-34 and SENet-50 is used to evaluate the performance on IJB-A and IJB-C datasets. To verify the effectiveness of the LIT module, we compare the LFENet with baseline methods, which use the original backbone network structures to calculate the representations of each image and simply apply the average pooling (Avg.) on them to get the set-level feature. The experiment results show that LFENet outperforms the baseline by a large margin and is superior to other state-of-the-art methods no matter with the same or more powerful backbones.

*Experiments on IJB-A.* We compare LFENet with the existing state-of-the-art methods aiming to solve the set-based face recognition task, where the video face recognition methods are also included because they consider video clips as unordered image sets as well. As shown in the Tab.I, our method gains about 2∼7% performance improvement

Authorized licensed use limited to: INSTITUTE OF COMPUTING TECHNOLOGY CAS. Downloaded on January 13,2022 at 06:28:43 UTC from IEEE Xplore. Restrictions apply.

TABLE II: Evaluation of the 1:1 verification protocol on the IJB-C dataset (higher is better). The methods in the first four rows focus on static face recognition, and the set-level feature is obtained by naive average/max pooling. The other methods are proposed for set-based face recognition.

| Method | Backbone | Training Data | 1:1 Verification TAR | | | |
|---|---|---|---|---|---|---|
| | | | FAR=1e-5(%) | FAR=1e-4(%) | FAR=1e-3(%) | FAR=1e-2(%) |
| VGGFace2 [2] | SENet-50 | VGGFace2(3.3M) | 73.40 | 82.50 | 90.00 | 95.00 |
| VGGFace2 [2] | SENet-50 | VGGFace2(3.3M) | 74.70 | 84.00 | 91.00 | 96.00 |
| Yin et al. [41] | ResNet-50 | MS1M(10M) | - | - | 93.20 | 95.80 |
| Zhao et el. [44] | Light CNN-29 | - | 82.60 | 89.50 | 93.50 | 96.20 |
| Multicolumn [39] | ResNet-50 | VGGFace2(3.3M) | 77.10 | 86.20 | 92.70 | - |
| DCN [38] | ResNet-50 | VGGFace2(3.3M) | - | 88.00 | 94.40 | 98.10 |
| DCN [38] | SENet-50 | VGGFace2(3.3M) | - | 88.50 | 94.70 | 98.30 |
| PFE [26] | 64CNN | Web.+MS1M(4.4M) | 89.64 | 93.25 | 95.49 | 97.17 |
| **LFENet(Ours)** | SENet-50 | VGGFace2(3.3M) | 88.39 | **93.63** | **96.69** | 98.28 |




(a) Comparison with SOTA methods (b) Comparison about relational functions (c) Comparison about different set sizes

Fig. 4: Average ROC curves on the IJB-A dataset over 10 splits. (a) Comparison among our proposed LFENet (solid line), baselines and other methods (dotted line); (b) Comparison among LFENet trained with different relational functions; (c) Comparison among LFENet trained with different set sizes.

compared with the corresponding baseline, which indicates that our proposed LIT module can make full use of the rich local information for set modeling. Although we only use the naive average pooling to compute the set representation without representative sample selection, the local information is transferred across the set to complement each other to promote each local feature. In addition, our method can achieve better performance under the same scale backbone network and training data, and the performance of LFENet on SENet-50 is better than the method [26] trained with large-scale 64CNN.

In Fig.4(a), the receiver operating characteristics (ROC) curve is visualized for a more clear comparison. Note that the ROC curves of other methods are drawn by connecting discrete points from the reported TAR@FAR. The larger area under the curve (AUC) illustrates that the our method can learn better face representation robust to drastic appearance changes, owing to local feature enhancement especially samples suffering from the loss of local information.

*Experiments on IJB-C.* To further verify the robustness of our model, we do experiments on the more challenging IJB-C dataset. Here we compare the performance of our LFENet with the latest static face recognition methods and set-based face recognition methods. Note that we use SENet-50 as our backbone. It can be seen from Tab. II that LFENet can still outperform all the other methods.

### *D. Ablation Study*

Here we evaluate various design choices of our LFENet and compare it with the baseline on the IJB-A dataset to comprehensively verify the effectiveness of the LIT module. Note that the models are trained based on the SENet-50.

*Attention Map Computing.* Firstly, we use three different relational functions which are introduced in III-C to calculate the attention map and train the corresponding model respectively. Experiment results in Tab.III and Fig.4(b) show that the three different functions can all achieve the purpose of building the correlation between the local features. The SoftmaxB function is the calculation method we used in the main experiment, which performs better than other ways. Besides, compared with the SoftmaxB which distills the local feature through the functions $Q(.)$ and $V(.)$, the SoftmaxA function directly computes the similarities between the local features, therefore the performance is not as good as the former.

*Size of the Image Set.* We further evaluate the influence of choosing different image set sizes on training the model to capture local feature correlation. We separately select 3, 5, and 7 images and a single image to make up the image sets. It is noticed that when one image is selected, the LIT

Authorized licensed use limited to: INSTITUTE OF COMPUTING TECHNOLOGY CAS. Downloaded on January 13,2022 at 06:28:43 UTC from IEEE Xplore. Restrictions apply.

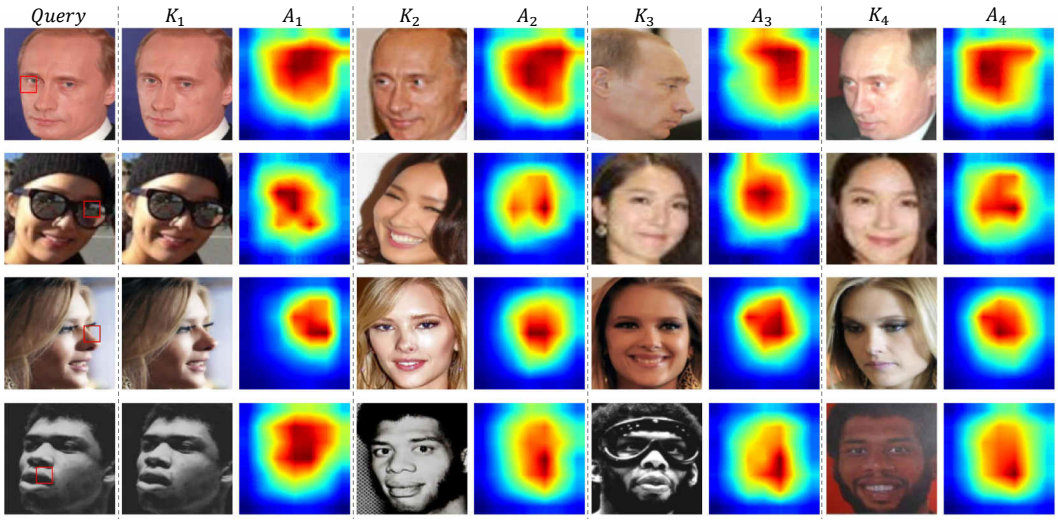



Fig. 5: Visualization results of attention maps on IJB-C dataset. The first column is the query image. We select some representative areas as the query region (bounded with red box). The next few columns represent the 'key' image and its corresponding attention map. **Best viewed in color.**

TABLE III: Evaluation of our proposed LFENet under different settings (*relational functions* for computing the attention map / *set size* chosen for training the model). TAR@FARs are reported under 1:1 verification protocol on IJB-A dataset.

| Method | 1:1 Verification TAR | | |
|---|---|---|---|
| | FAR=1e-3 | FAR=1e-2 | FAR=1e-1 |
| SENet-50 | 93.62±0.84 | 95.32±0.76 | 97.45±0.44 |
| LFENet(SoftmaxA/size-3) | 86.58±3.25 | 95.56±1.05 | 98.97±0.28 |
| LFENet(Scaled Dot/size-3) | 90.42±1.78 | 97.11±0.50 | 99.10±0.30 |
| **LFENet(SoftmaxB/size-3)** | **96.83±0.65** | **98.93±0.39** | **99.57±0.15** |
| LFENet(SoftmaxB/size-1) | 90.31±1.87 | 95.16±0.84 | 98.00±0.40 |
| LFENet(SoftmaxB/size-5) | 95.58±0.90 | 98.42±0.40 | 99.53±0.17 |
| LFENet(SoftmaxB/size-7) | 94.91±1.63 | 98.25±0.57 | 99.42±0.27 |

module is equivalent to aggregate the relevant local features within a single image. We do experiments on IJB-A and the results are shown in the Tab.III. The ROC curves are drawn in Fig.4(c).

According to the results, when the set size is 3, the model achieves the best performance. In addition, the model trained with image sets composed of multiple images performs better than that trained with a single image. It indicates that compared with only integrating spatial semantic information within the same image, the aggregation of relevant local information from multiple images is more helpful to improve the discriminability of the local feature.

*Attention Map Visualization.* In order to explore the ability of the LIT module on capturing the correlation among the local features, we visualize the attention maps on some samples from the IJB-C dataset. The visualization result is shown in Fig.5. The first column shows the query image, in which we select some representative local areas as the query regions. The following columns show the 'key' images and their corresponding visualized attention maps that are resized to the same size as the input image. Note that the response value is normalized by the sum of all response values of the attention map. The higher the response value, the stronger the correlation with the query area. It can be seen that LFENet can effectively find the same local region of the faces under different conditions. In addition, for queries with low quality (*e.g.* occlusion in the second example), the LIT module can still find their congruent regions from other images in the set. Thus the key information from these regions can transfer to the query regions to enhance its discriminability.

## V. CONCLUSIONS

To address the problem of face recognition based on the image set, we propose an advanced method called Local Feature Enhancement Network (LFENet). Specifically, we enhance the local feature of each image within the set by transferring the local information for two reasons: 1) The local features with serious information loss can obtain the key face information from the others; 2) The various local information can be aligned in multiple faces under different conditions. The extensive experiments reflect that our LFENet can learn robust face representation through deeply exploiting the rich information within the image set.

## REFERENCES

[1] R. Arandjelović, P. Gronát, A. Torii, T. Pajdla, and J. Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40:1437–1451, 2018.

[2] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age. *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, pages 67–74, 2018.

[3] A. R. Chowdhury, T.-Y. Lin, S. Maji, and E. Learned-Miller. One-to-many face recognition with bilinear cnns. *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9, 2016.

[4] J. Deng, J. Guo, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4685–4694, 2019.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019.

Authorized licensed use limited to: INSTITUTE OF COMPUTING TECHNOLOGY CAS. Downloaded on January 13,2022 at 06:28:43 UTC from IEEE Xplore. Restrictions apply.

[6] S. Gong, Y. Shi, N. D. Kalka, and A. K. Jain. Video face recognition: Component-wise feature aggregation network (c-fan). *2019 International Conference on Biometrics (ICB)*, pages 1–8, 2019.

[7] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *ECCV*, 2016.

[8] T. Hassner, I. Masi, J. Kim, J. Choi, S. Harel, P. Natarajan, and G. Medioni. Pooling faces: Template based face recognition with pooled face images. *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 127–135, 2016.

[9] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[10] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu. Squeeze-and-excitation networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42:2011–2023, 2020.

[11] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *ArXiv*, abs/1502.03167, 2015.

[12] B. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. C. Allen, P. Grother, A. Mah, M. Burge, and A. K. Jain. Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1931–1939, 2015.

[13] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut. Albert: A lite bert for self-supervised learning of language representations. *ArXiv*, abs/1909.11942, 2020.

[14] H. Li, G. Hua, X. Shen, Z. L. Lin, and J. Brandt. Eigen-pep for video face recognition. In *ACCV*, 2014.

[15] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song. Sphereface: Deep hypersphere embedding for face recognition. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6738–6746, 2017.

[16] X. Liu, Z. Guo, S. Li, L. Kong, P. Jia, J. You, and B. Kumar. Permutation-invariant feature restructuring for correlation-aware image set-based recognition. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4985–4995, 2019.

[17] X. Liu, B. V. Kumar, C. Yang, Q. Tang, and J. You. Dependency-aware attention control for unconstrained face recognition with image sets. In *ECCV*, 2018.

[18] Y. Liu, J. Yan, and W. Ouyang. Quality aware network for set to set recognition. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4694–4703, 2017.

[19] Y. Mao, R. Wang, S. Shan, and X. Chen. Cosonet: Compact second-order network for video face recognition. In *ACCV*, 2018.

[20] B. Maze, J. Adams, J. A. Duncan, N. D. Kalka, T. Miller, C. Otto, A. K. Jain, W. T. Niggel, J. Anderson, J. Cheney, and P. Grother. Iarpa janus benchmark - c: Face dataset and protocol. *2018 International Conference on Biometrics (ICB)*, pages 158–165, 2018.

[21] O. Oktay, J. Schlemper, L. L. Folgoc, M. J. Lee, M. Heinrich, K. Misawa, K. Mori, S. G. McDonagh, N. Hammerla, B. Kainz, B. Glocker, and D. Rueckert. Attention u-net: Learning where to look for the pancreas. *ArXiv*, abs/1804.03999, 2018.

[22] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. 2017.

[23] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners. 2019.

[24] Y. Rao, J. Lin, J. Lu, and J. Zhou. Learning discriminative aggregation network for video-based face recognition. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3801–3810, 2017.

[25] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, 2015.

[26] Y. Shi, A. K. Jain, and N. D. Kalka. Probabilistic face embeddings. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6901–6910, 2019.

[27] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15:1929–1958, 2014.

[28] Y. Sun, Y. Chen, X. Wang, and X. Tang. Deep learning face representation by joint identification-verification. In *NIPS*, 2014.

[29] Y. Sun, C. Cheng, Y. Zhang, C. Zhang, L. Zheng, Z. Wang, and Y. Wei. Circle loss: A unified perspective of pair similarity optimization. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6397–6406, 2020.

[30] Y. Sun, D. Liang, X. Wang, and X. Tang. Deepid3: Face recognition with very deep neural networks. *ArXiv*, abs/1502.00873, 2015.

[31] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1701–1708, 2014.

[32] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. *ArXiv*, abs/1706.03762, 2017.

[33] F. Wang, J. Cheng, W. Liu, and H. Liu. Additive margin softmax for face verification. *IEEE Signal Processing Letters*, 25:926–930, 2018.

[34] H. Wang, Y. Wang, Z. Zhou, X. Ji, Z. Li, D. Gong, J. Zhou, and W. Liu. Cosface: Large margin cosine loss for deep face recognition. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5265–5274, 2018.

[35] X. Wang, R. B. Girshick, A. Gupta, and K. He. Non-local neural networks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2018.

[36] X. Wang, S. Zhang, S. Wang, T. Fu, H. Shi, and T. Mei. Mis-classified vector guided softmax loss for face recognition. In *AAAI*, 2020.

[37] Y. Wen, K. Zhang, Z. Li, and Y. Qiao. A discriminative feature learning approach for deep face recognition. In *ECCV*, 2016.

[38] W. Xie, L. Shen, and A. Zisserman. Comparator networks. In *ECCV*, 2018.

[39] W. Xie and A. Zisserman. Multicolumn networks for face recognition. In *BMVC*, 2018.

[40] J. Yang, P. Ren, D. Zhang, D. Chen, F. Wen, H. Li, and G. Hua. Neural aggregation network for video face recognition. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5216–5225, 2017.

[41] B. Yin, L. Tran, H. Li, X. Shen, and X. Liu. Towards interpretable face recognition. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9347–9356, 2019.

[42] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23:1499–1503, 2016.

[43] M. Zhang, G. Song, H. Zhou, and Y. Liu. Discriminability distillation in group representation learning. In *ECCV*, 2020.

[44] J. Zhao, Y. Cheng, Y. Cheng, Y. Yang, H. Lan, F. Zhao, L. Xiong, Y. Xu, J. Li, S. Pranata, S. Shen, J. Xing, H. Liu, S. Yan, and J. Feng. Look across elapse: Disentangled representation learning and photorealistic cross-age face synthesis for age-invariant face recognition. In *AAAI*, 2019.

[45] Y. Zhong, R. Arandjelović, and A. Zisserman. Ghostvlad for set-based face recognition. *ArXiv*, abs/1810.09951, 2018.

[46] B. Zhou, A. Andonian, and A. Torralba. Temporal relational reasoning in videos. *ArXiv*, abs/1711.08496, 2018.


Authorized licensed use limited to: INSTITUTE OF COMPUTING TECHNOLOGY CAS. Downloaded on January 13,2022 at 06:28:43 UTC from IEEE Xplore. Restrictions apply.