Hierarchical Pyramid Diverse Attention Networks for Face Recognition

Qiangchang Wang[1], Tianyi Wu[2,3], He Zheng[2,3], Guodong Guo[1,2,3,*]
[1]West Virginia University, Morgantown, USA, [2]Institute of Deep Learning, Baidu Research, Beijing, China
[3]National Engineering Laboratory for Deep Learning Technology and Application, Beijing, China
qiangchang.wang666@gmail.com, {wutianyi01,zhenghe01}@baidu.com,guodong.guo@mail.wvu.edu

Abstract

Deep learning has achieved a great success in face recognition (FR), however, few existing models take hierarchical multi-scale local features into consideration. In this work, we propose a hierarchical pyramid diverse attention (HPDA) network. First, it is observed that local patches would play important roles in FR when the global face appearance changes dramatically. Some recent works apply attention modules to locate local patches automatically without relying on face landmarks. Unfortunately, without considering diversity, some learned attentions tend to have redundant responses around some similar local patches, while neglecting other potential discriminative facial parts. Meanwhile, local patches may appear at different scales due to pose variations or large expression changes. To alleviate these challenges, we propose a pyramid diverse attention (PDA) to learn multi-scale diverse local representations automatically and adaptively. More specifically, a pyramid attention is developed to capture multi-scale features. Meanwhile, a diverse learning is developed to encourage models to focus on different local patches and generate diverse local features. Second, almost all existing models focus on extracting features from the last convolutional layer, lacking of local details or small-scale face parts in lower layers. Instead of simple concatenation or addition, we propose to use a hierarchical bilinear pooling (HBP) to fuse information from multiple layers effectively. Thus, the HPDA is developed by integrating the PDA into the HBP. Experimental results on several datasets show the effectiveness of the HPDA, compared to the state-of-the-art methods.

1. Introduction

CNN representations achieve the state-of-the-art in face recognition (FR). However, most existing models learn global representations where whole faces are regarded as CNN inputs [22, 34, 19, 3]. Few works take hierarchical multi-scale local representations into account.

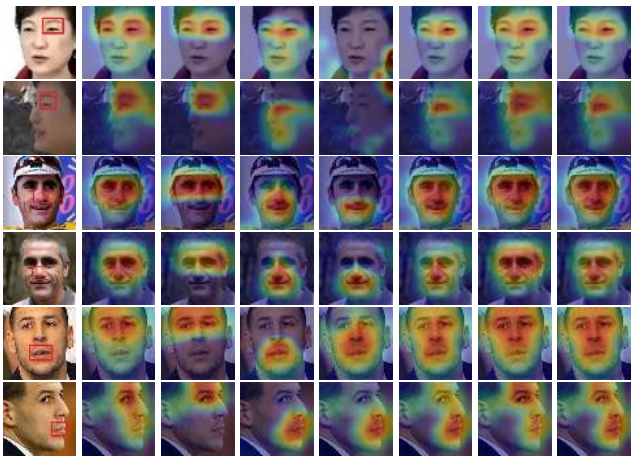

Figure 1. **Illustration on the effects of local CNNs, the diverse learning and global CNN**. Local CNNs learn diverse local representations using multiple local branches at various scales from different hierarchical layers. For good illustration, only one scale and the last convolutional layer are used. The number of local branches is 3. The diverse learning guides multiple local branches to locate diverse local patches. Global CNN extracts holistic representations. **Column 1**: faces with varying challenging factors (e.g. resolution, pose, occlusion, aging and expression). **Column 2**: a model which has a single local branch. **Columns 3, 4 and 5**: 1st, 2nd and 3rd local branch of local CNNs, which are guided by the diverse learning. **Column 6**: local CNNs. **Column 7**: global CNN. **Column 8**: fused global and local CNNs.

It is observed that global face geometry and appearances may change dramatically under pose, age, or large quality variations. In contrast, some facial parts remain similar in these cases, which would play important roles in FR. For instance, as shown in Fig. 1, it is difficult to consider the global face representation to match the frontal face (Row 1) with the profile (Row 2) which is influenced by blur, pose and background distraction. However, we notice that similar eyes in these two faces can contribute to verification.

*Corresponding author.
‡Part of this work was done when Qiangchang Wang was a research intern with Baidu.

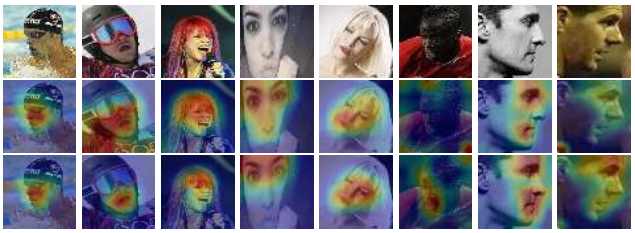

Figure 2. **Qualitative comparison between prior local methods and the proposed HPDA model**. Landmark-based methods suffer from landmark detection failure (**Row 1**) and attention-based methods locate uninformative or even noisy regions (**Row 2**). The proposed HPDA can emphasize discriminative information and inhibit less important (**Row 3**).

Similar observations about pointy noses (Rows 3, 4) and big lips (Rows 5, 6) can be made. Thus, representing similar facial parts become especially important. Previous works mainly depend on face landmarks to incorporate local information [23, 18, 4, 14, 38, 13, 12]. However, landmark detection may be inaccurate or even fail due to occlusions, large head poses, extreme illuminations, or dramatic expression changes. As shown in Fig. 2, Row 1, MTCNN [42] failes to detect landmarks for faces which are influenced by occlusions (Columns 1, 2), expressions (Columns 3, 4), illumination (Columns 5, 6), and poses (Columns 7, 8).

Without relying on face landmarks, discriminative local patches are located automatically in [30, 12]. As shown in Fig. 1, compared to the model without an attention module (Column 7), important facial parts are enhanced and some useless ones are suppressed when the attention is applied (Column 2). However, it is observed that only specific facial regions are located, while neglecting some potential discriminative regions. For instance, the attention map has strong responses around eyes (Column 2), but ignores some discriminative regions, such as similar big lips (Column 2, Rows 5, 6). From the above analysis, we expect that emphasized facial parts should be well distributed over face images to extract more useful features. To achieve this goal, a diverse learning is developed to guide multiple attention modules to accurately locate diverse discriminative facial parts as well as reduce the background distraction.

Learning multi-scale representations is beneficial to various tasks [25, 43, 26, 31, 35]. As for FR, local patches may have various sizes or shapes under pose or expression changes, making it necessary to learn multi-scale features. Take faces in Fig. 2 as examples, eyes with different expressions in Columns 3, 4 and 5 appear at varying sizes and shapes; due to pose variations, mouths have different sizes in Columns 6, 7 and 8. [23] fuses multi-scale features from the last two layers. [30] extracts rich multi-scale features from two harmonious perspectives: different convolutional sizes in a single layer and hierarchical concatenation of feature maps from varying layers. However, both approaches ignore the fact that features in a layer may cover a large range of scales. This is especially important for layers which concatenate feature maps from prior layers and generate feature maps by multi-scale convolution kernels. To address this challenge, we propose a pyramid attention which scales feature maps within a layer under different scales, exploring multi-grained information.

Most previous works only use the last convolutional layer, lacking of low-level information. This is because units in high layers have large receptive fields, and hence respond around large-scale facial parts and represent high-level semantic information, but inevitably lack of locally detailed information or small-scale face parts in low layers. [30] can alleviate the above problem by combining hierarchical information from different layers within a block. [13] incorporates low-level features with high-level features to capture discriminative representations. Both methods combine hierarchical information simply by concatenation, which leads to sub-optimal cross-layer information fusion.

In this work, we propose a hierarchical pyramid diverse attention (HPDA) network which can describe diverse local patches at various scales adaptively and automatically from varying hierarchical layers. Fig. 3 illustrates the framework. First, we propose a pyramid diverse attention, as shown in Fig. 4. Specifically, feature maps are pooled to various scales, allowing for exploiting features at different scales. Meanwhile, since attention modules tend to have redundant responses around some similar face regions, a diverse learning is proposed to guide multiple local branches in each pyramid scale to focus on diverse facial regions automatically, instead of relying on face landmarks. As shown in Fig. 1, the diverse learning leads to localization of different discriminative local patches in Columns 3, 4 and 5. Second, a hierarchical bilinear pooling is proposed to combine complementary information from different hierarchical layers. Exactly, it uses different cross-layer bilinear modules to integrate both the high-level abstraction and the low-level detained information. The major contributions of our work are three-fold:

1. The proposed pyramid diverse attention introduces multiple attention-based local branches at different scales to emphasize different discriminative facial regions at various scales automatically, avoiding the need of face landmark detection. To our knowledge, this is the first attempt of automatic locating multiple complementary facial regions in general face recognition.

2. Instead of simple concatenation or addition, a hierarchical bilinear pooling is presented to combine features from different hierarchical layers, covering both local details or small-scale face regions in low layers to high-level abstraction and large-scale parts in high layers.

3. The proposed model achieves the state-of-the-art performance on several challenging face recognition tasks.

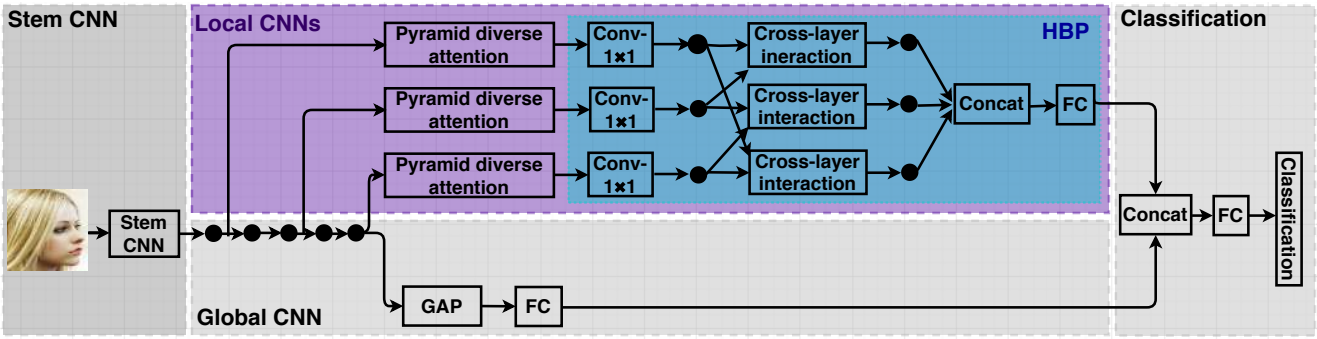

Figure 3. **Framework of the proposed hierarchical pyramid diverse attention (HPDA) model**. GAP and FC layers mean global average pooling and fully connected layers. Local CNNs learn rich local representations which mainly consist of a pyramid diverse attention (PDA) and a hierarchical bilinear pooling (HBP). The PDA aims at learning multi-scale diverse local features. The HBP fuses complementary local information from hierarchical layers.

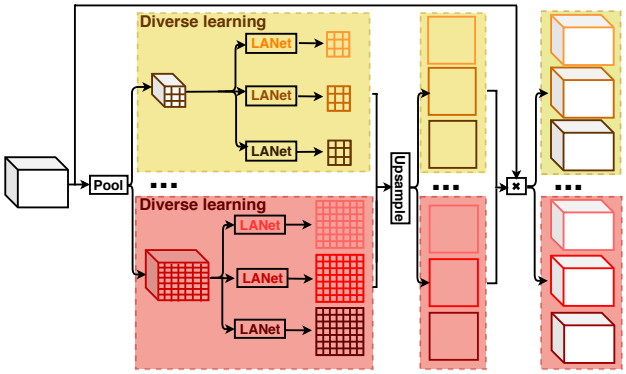

Figure 4. **Framework of the proposed pyramid diverse attention (PDA)**. We set the number of local branches to 3 as an example. It consists of a pyramid attention and a diverse learning. The former allows the network to weigh face parts at various scales automatically. The latter guides multiple local branches to extract diverse local representations.

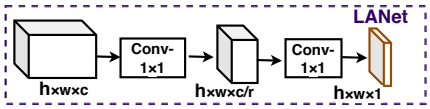

Figure 5. **Framework of the LANet**, where h, w and c represent height, width and number of feature maps, respectively. r is the reduction ratio.

2. Related Work

Related face recognition and attention modules are reviewed.

2.1. Face Recognition

Deep learning learns representations from global faces or local patches for face recognition. For the latter, there are landmark-based and attention-based methods.

Global faces based models usually accept whole faces as inputs [22, 34, 19, 28, 3]. However, local patches are not taken into consideration, resulting in sub-optimal performances when global face appearances change dramatically. To address this problem, several works train multiple CNNs separately on different local patches cropped around face landmarks [23, 18]. Since networks are not trained jointly, correlations of multiple face regions are not well explored. To overcome this issue, some methods are proposed to jointly train models [4, 14, 38, 13]. However, these landmark-based methods rely on face landmarks which may not be reliable in some cases. Besides, different facial parts should have different discriminative abilities [30], which are not considered in most existing works.

Attention-based works use attention mechanisms to weigh local patches automatically without using face landmarks. [30] is an early attempt to achieve this for the general face recognition, where a spatial attention (i.e. LANet) is proposed to capture important local regions and weigh adaptively on different local regions. In [12], importance of each local pair is modeled by a low-rank bilinear pooling. However, only the last convolutional layer is used, resulting in loss of local details or small-scale objects in low layers because of information loss in CNN propagation.

2.2. Attention Modules

Attention modules have been widely used in computer vision [9, 6, 27, 37]. [5, 16, 17] localize flexible parts by spatial transformer networks with localization constraints. [40] generates local features on feature space by the ROI pooling. [44, 39] locate parts with prior knowledge about their approximate locations. [20, 36] use RNN or LSTM models for sequentially selecting attention regions and learning features for each part. To our knowledge, our work is the first effort to automatically learn diverse local representations for general face recognition without prior spatial constraints. Besides, compared with many approaches which learn local features from original images, our proposed model extracts features from feature maps, which is computationally cheap.

Orthogonolity can guide the diverse feature learning. The singular value decomposition (SVD) is applied in [21, 24] to constrain the solution on a Stiefel manifold. [11] uses SVD on weights of the last layer to reduce feature correlations. However, SVD-based constraints are computationally expensive, limiting the learning flexibility. In contrast, our approach encourages diverse feature learning by a simple yet effective divergence loss on feature maps with very limited additional computation overhead.

3. Proposed Method

3.1. Overall Framework

As shown in Fig. 3, our framework consists of four parts: stem CNN, local CNNs, global CNN and classification.

Due to its good performance, HSNet-61 model [30] is used. SENet [9] can enhance important feature maps and inhibit less informative. Since layers in HSNet model contain feature maps captured at various sizes, we fuse SENet with HSNet model to improve the model capacity, namely SENet-HSNet. There are three blocks. The first two blocks are used as the stem CNN, as illustrated in Fig. 3. In the third block, every two layers of five layers are used as inputs of local CNNs which are designed to extract hierarchical multi-scale diverse local features. More details are discussed in Sec. 3.2. SENet-HSNet model is the default model if not specified. We also consider LS-CNN model [30] as the stem CNN due to its powerful generalization ability. In theory, our proposed hierarchical pyramid diverse attention module can be applied to any networks. Here, we investigate the above two networks as representatives.

Finally, the learned global and local information is combined to further boost the accuracy. Global CNN extracts the holistic face representation by consecutive global average pooling (GAP) and fully connected (FC) layers with 512 units. Local CNNs output 512 units. We can obtain joint local and global representations with 1024 units, followed by a FC layer with 512 units and a classification layer.

3.2. Local CNNs

Local CNNs are developed to extract multi-scale diverse local features hierarchically, which consist of a pyramid attention and a diverse learning. Specifically, due to pose variations or large expression changes, facial parts may have varying sizes. Therefore, it is necessary to represent local patches at various sizes. To this aim, a pyramid attention is proposed to locate multi-scale discriminative facial regions adaptively. However, it is possible that multiple local branches have redundant responses around the similar regions, like similar eye responses in Fig. 1, Column 2. To address this issue, a diverse learning is proposed to guide multiple branches to locate diverse facial regions automatically. The pyramid diverse attention (PDA) is formed by combining the pyramid attention with the diverse learning.

Besides, few works consider hierarchical features from different layers in face recognition, resulting in loss of discriminative features from low layers. To alleviate this issue, the hierarchical bilinear pooling is used to explore good inter-layer interactions.

3.2.1 Pyramid Attention

The framework is shown in Fig. 4. There are multi-scale local representations encoded in a single layer, making it necessary to calibrate features at different scales.

Let $X \in R^{\mathrm{h} \times \mathrm{w} \times \mathrm{c}}$ denote the input of the pyramid attention, where h, w and c represent height, width and number of feature maps, respectively. First, feature maps are splited into outputs with different scales $\left[X_1, X_2, ..., X_S\right]$, where S is the number of scales. $X_i \in R^{\mathrm{h_i} \times \mathrm{w_i} \times \mathrm{c}}$ is the output of the i^{th} level, where $h_i \times \mathrm{w}_i$ represent the spatial size. The finest level has the same size as input X. The other levels split feature maps into different sub-regions, and then pool features in corresponding sub-regions.

Second, for the i^{th} scale, the target is to output diverse attention masks $\left[M_i^1, M_i^2, ..., M_i^B\right]$, where B is the number of local branches. The j^{th} output attention mask in the i^{th} scale has the same spatial size as the input X_i, i.e. $M_i^j \in R^{\mathrm{h_i} \times \mathrm{w_i}}$. To model the feature map spatially, the Local Attention Network (LANet) from [30] is used, as shown in Fig. 5. Weights of different spatial locations are regressed by two consecutive convolutional layers. The first layer has $\frac{c}{r}$ feature maps, followed by a ReLU layer to increase the non-linearity. r means the channel reduction ratio. The second one generates a feature map (i.e. M_i^j) with a sigmoid function.

Third, we upsample different attention masks M_i^j across multiple local branches ($j \in \left[1, 2, ..., B\right]$) in different pyramid scales ($i \in \left[1, 2, ..., S\right]$) to have the same size as input $X \in R^{\mathrm{h} \times \mathrm{w} \times \mathrm{c}}$ using a bilinear interpolation.

Then, refined feature maps R_i^j for the j^{th} local branch in the i^{th} scale is aggregated by the product of the attention mask M_i^j and input X:

$$R_i^j = X \circ M_i^j, \tag{1}$$

where $\circ$ denotes Hadamard product.

Finally, output of i^{th} scale is obtained by first concatenating B local branches and followed by a 1×1 convolutional layer to output c feature maps. Then, feature maps across different scales are concatenated, regarding as outputs of the pyramid attention.

However, it may be difficult for multiple local branches in the same scale to find different discriminative regions simultaneously. To address this problem, a diverse learning is proposed to guide the learning of complementary information across different local branches.

3.2.2 Diverse Learning

A divergence loss L_{D} is proposed to guide multiple local branches to learn diverse attention masks (i.e. $M_i^1, M_i^2, ..., M_i^B$), locating diverse face regions and achieving a robust recognition from diversified parts (e.g. the left eye and mouth if the right eye is occluded in Fig. 1, Row 6, Column 8). The formulation is defined as following:

$$L_{\mathrm{D}} = \frac{2}{SB(B-1)} \sum_{i=1}^{S} \sum_{j=1}^{B} \sum_{\substack{\mathrm{k=1,} \\ k \neq j}}^{B} max(0, t - (M_i^j - M_i^k)^2), \tag{2}$$

where t is a hyper-parameter margin. M_i^j and M_i^k represent attention masks learned by local branches j and k in the i^{th} scale, respectively.

The diverse learning encourages each local branch to learn different attention masks by increasing their distances. Since each attention mask is applied on the same feature maps, diverse attention masks can locate different local patches. Consequently, the diverse learning can alleviate the problem of redundant responses in Fig. 1, Column 2, and extract diverse local patches, as shown in Fig. 1, Columns 3, 4 and 5. It is possible that some less important facial parts and even noisy background noise are located, resulting in sub-optimal facial representations. To overcome this issue, the classification loss is used with the divergence loss to select discriminative local patches. Only discriminative local patches are emphasized. This explains why attention masks may have overlaps among different local branches where some attention masks emphasize small regions while some focus on larger facial parts.

3.2.3 Hierarchical Bilinear Pooling

Most existing works learn features from the last convolutional layer. However, representations from the individual layer are not comprehensive. We consider to fuse features from multiple hierarchical layers.

There are five layers in the last block of HSNet-61 model, shown in Fig. 3. The pyramid diverse attention is applied in every two layers, extracting complementary local information across different layers. Each single layer contains three parallel paths and the deepest path has three convolutional layers. Therefore, different features can be extracted hierarchically in every two layers. Instead of simple concatenation or addition, we adopt the approach in [41] to aggregate information from different layers.

Suppose $X \in R^{\mathrm{h} \times \mathrm{w} \times \mathrm{c}_1}$ and $Y \in R^{\mathrm{h} \times \mathrm{w} \times \mathrm{c}_2}$ are outputs of two different layers, where h and w represent height and width of feature maps, respectively. c_1 and c_2 mean the number of feature maps in two different layers. We denote a c_1 dimensional feature at a spatial location on X as $x = \left[X_1, X_2, ..., X_{\mathrm{c}_1}\right]$. Similarly, a c_2 dimensional feature from Y is $y = \left[Y_1, Y_2, ..., Y_{\mathrm{c}_2}\right]$.

To capture more comprehensive local features, the cross-layer interaction is used, which is defined as following:

$$z_{\mathrm{i}} = x^\top W_{\mathrm{i}} y = x^\top U_{\mathrm{i}} V_{\mathrm{i}}{}^\top y = U_{\mathrm{i}}{}^\top x \circ V_{\mathrm{i}}{}^\top y, \tag{3}$$

where $W_{\mathrm{i}} \in R^{\mathrm{c} \times \mathrm{c}}$ is the projection matrix and $\mathrm{z_i}$ is the projected output. $\circ$ means Hadamard product. In [41], the projection matrix is factorized into two one-rank vectors $U_{\mathrm{i}}, V_{\mathrm{i}} \in R^{\mathrm{c}}$.

To encode local information, features should be expanded to a high dimensional space by linear mappings. Therefore, a weight matrix $w = \left[w_1, w_2, ..., w_{\mathrm{d}}\right]$ is defined to obtain a d dimensional feature z.

$$z = U^\top x \circ V^\top y, \tag{4}$$

where $U, V \in R^{\mathrm{c} \times \mathrm{d}}$ and d is the dimension of the projected feature.

We consider to aggregate features from more layers, capturing more discriminative local features. Let X^1, X^2, X^3 represent outputs from three different layers. Eq. (4) is extended to concatenate multiple cross-layer representations:

$$z = Concat(U^\top x^1 \circ V^\top x^2, U^\top x^1 \circ S^\top x^3, V^\top x^2 \circ S^\top x^3). \tag{5}$$

Finally, a FC layer is used to reduce dimension to 512.

4. Experiments

In this section, we conduct experimental validation of the proposed HPDA model on several datasets.

4.1. Data Sets

4.1.1 Training Data

VGGFace2. VGGFace2 [1] has 3.14 million faces from 8,631 subjects, which covers a large range of poses, ages, ethnicities and professions. It is the default training dataset if unspecified.

MS-Celeb-1M. The original MS-Celeb-1M dataset [8] contains too much noise. To get a high-quality dataset, [3] refined the dataset and made it publicly available. There are about 85,000 subjects with 5.8 million aligned images.

4.1.2 Test Data

Experiments are conducted on several datasets, including IJB-A [15] quality, CALFW [46], CACD-VS [2], CPLFW [45], VGGFace2-FP [1] and LFW [10] datasets. For the IJB-A quality dataset, following the work [7], there are 1,543 high-quality faces from 500 subjects and 6,196 low-quality images from 489 identities. The protocol is that each image is compared with every other image, making the task very challenging. There are 9.56 million pairs. For more details about other datasets, please refer to their references.

Model	LFW	CPLFW	CALFW
W/o diverse learning	99.15	85.82	90.52
W/o hierarchical	99.07	85.63	90.35
Global CNN	99.10	79.30	89.90
Local CNNs	99.22	85.73	90.20
HPDA	99.33	86.07	90.93

Table 1. Ablation analysis (%) of the HPDA. Global CNN or local CNNs refer to the framework in Fig. 3.

S	B	t	Channel fusion	LFW	CPLFW	CALFW
1	3	1.00	HBP	99.30	85.93	90.57
2	3	1.00	HBP	99.17	85.77	90.58
3	3	1.00	HBP	99.33	86.07	90.93
4	3	1.00	HBP	99.15	85.53	91.10
3	1	1.00	HBP	99.08	85.63	90.47
3	3	1.00	HBP	99.33	86.07	90.93
3	5	1.00	HBP	99.07	86.14	90.87
3	3	0.25	HBP	99.20	85.98	90.30
3	3	0.50	HBP	99.15	85.73	90.75
3	3	1.00	HBP	99.33	86.07	90.93
3	3	2.00	HBP	98.97	85.18	90.65
3	3	4.00	HBP	99.20	85.97	90.43
3	3	1.00	Concat	99.17	85.82	90.88
3	3	1.00	Add	99.18	85.67	90.55
3	3	1.00	HBP	99.33	86.07	90.93

Table 2. Comparison of varying number of **optimal scale levels** S and **local branches** B, values of **hyper-parameter** t and **methods to fuse channels from different layers**.

4.2. Implementation Details

CPLFW and CALFW datasets provide cropped faces. For other datasets, faces are cropped by MTCNN [42].

The t in Eq. (2) is set to 1 empirically. After comparative experiments, the number of local branches is 3 (i.e. $B = 3$) and the pyramid diverse attention in Fig. 4 has three pyramid scales (i.e. $S = 3$).

4.3. Ablation Analysis

We conduct several experiments to analyze the proposed model on LFW, CPLFW and CALFW datasets. Tables 1 and 2 give detailed investigations.

Importance of Diverse Learning. To locate diverse facial regions automatically in multiple local branches, a diverse learning is proposed. Competitive results demonstrate its superiority in Table 1.

Besides, this is a face landmark free approach. For example, MTCNN [42] fails to detect landmarks for faces in Fig. 2, Row 1. For such challenging cases, our model can still learn discriminative local patches regardless of loss of face organs and noisy background (Row 3). Besides, it is also observed that in Fig. 1, Columns 3, 4 and 5, diverse attentions are generated by three local branches when the diverse learning is used, compared with the model without the diverse learning in Fig. 1, Column 2. This is because the diverse learning can emphasize informative face regions and suppress less important ones or background distraction. On one hand, discriminative local patches are enlarged. For example, as demonstrated in Fig. 2, Columns 6, 8, without diverse learning, discriminative face regions are lost in Row 2 under dramatic illumination changes or pose variations. However, our method locates rich discriminative face regions in Row 3. This benefits from the diverse learning which pushes models to learn more discriminative regions. On the other hand, our method suppresses noisy regions, like goggles in Fig. 2, Columns 1, 2, Row 3, compared with the model without the diverse learning in Row 2.

Importance of Hierarchical Features. Compared with the last convolutional layer which captures high-level face abstractions, lower layers preserve more local details or small-scale face parts, exhibiting complementary components.

Our model fuses features from three different layers hierarchically, as shown in Fig. 3. To investigate the necessity, we compare the HPDA model with the model without hierarchical information (i.e. w/o hierarchical) which only extracts features based on the last convolutional layer. As demonstrated in Table 1, our HPDA has slightly better overall performances. This shows the complementary information contained in low layers.

Effects of Hierarchical Bilinear Pooling. It is intuitive to directly concatenate or add feature maps from different layers. However, simply concatenation or addition fails to capture rich inter-layer feature relations. To overcome this issue, we use hierarchical bilinear pooling (HBP) to incorporate multiple cross-layer interaction modules to learn complementary information from different layers.

Experiments are conducted to compare HBP with concatenation and addition. For fair comparison, channels in each layer is transformed to a high dimension by 1×1 convolution as the HBP, encoding rich local information. Table 2 shows that HBP achieves slightly better performances than concatenation or addition, demonstrating that its superiority in capturing inter-layer feature relations.

Effects of Number of Local Branches B. We set B to 1, 3 and 5. The performance increases when B is changed from 1 to 3. This is because more discriminative facial regions are located by varying local branches which are guided by the diverse learning, boosting the performance. However, the performance drops slightly when B is changed from 3 to 5. This can be explained by the fact that the diverse learning encourages diverse information extracted by different branches, resulting in useless or noisy information in some branches when B is too large.

Importance of Local CNNs and Global CNN. First, as shown in Fig. 1, the global CNN tends to characterize some less informative regions (e.g. cheek in Column 7, Rows 3, 4) and model more background information, compared with the local CNNs (Column 7 to Column 6, Rows 3, 4). On the other hand, the local CNNs focus on informative areas (Column 6, Rows 3, 4). This explains why the local CNNs offer performance improvement compared with the global CNN, as shown in Table 1.

Second, global features describe general information of whole faces (Fig. 1, Column 7). Differently, the local CNNs encode more detailed characteristics within different local patches (Fig. 1, Column 6). Take faces in Fig. 1, Rows 3, 4 as examples, the noses and the shape of whole faces remain similar in these two faces despite changes of the global face appearances. The global CNN can represent holistic face shapes (Column 7) and the local CNNs can describe noses (Column 6). It is intuitive to combine the global CNN with the local CNNs to boost the performance. It is observed that the combined global and local representation (Column 8) is more descriptive compared with the global CNN (Column 7) and the local CNNs (Column 6).

Effects of the Scale Level Number S. We set the scale number S in the pyramid attention to 1, 2, 3 and 4, respectively, where scale 7 is used for $S = 1$ and scale 7, 5, 3 and 1 are used for $S = 4$. However, setting proper scale levels is necessary to boost the performance. The HPDA achieves the best overall performance when $S = 3$ in Table 2.

Discriminative face regions have varying sizes due to expression changes (e.g. eyes in Fig. 2, Row 1, Columns 3, 4, 5) and pose changes (e.g. mouths in Fig. 2, Row 1, Columns 6, 7, 8). Although HSNet model extracts rich multi-scale features in a single layer by utilization of multi-scale convolutional kernels and concatenation of feature maps from previous layers, it can just weigh parts with fixed scales, while lacking the capacity to emphasize regions with flexible scales. In contrast, when the pyramid attention is used, attention masks are able to capture information at various scales.

Effects of Hyper-parameter t. Values of t (0.25, 0.5, 1, 2 and 4) are compared. It can be seen that performance increases when t increases from 0.25 to 1, but drops when t increases from 1 to 4.

4.4. Experiments on Cross-quality Face Matching

Notice that this task is very close to real-world scenarios where the match is between faces from access control, video surveillance, or public safety (low-quality faces) and enrolled photos (high-quality faces). Following the work [7], several public models are compared. As shown in Table 3, our method HPDA increases the accuracies.

There are many factors that can influence the quality of face images, such as resolution, pose, expression, aging, and illumination. A positive pair of high-quality and low-quality faces are shown in Fig. 1, Rows 1, 2. In such cases, the local CNNs can learn discriminative and diverse local regions. As illustrated in Column 6, it is observed that the local CNNs filter out useless background information and enhance important regions. In contrast, without the diverse learning, these compared models cannot filter out background, which would inevitably decrease the performance.

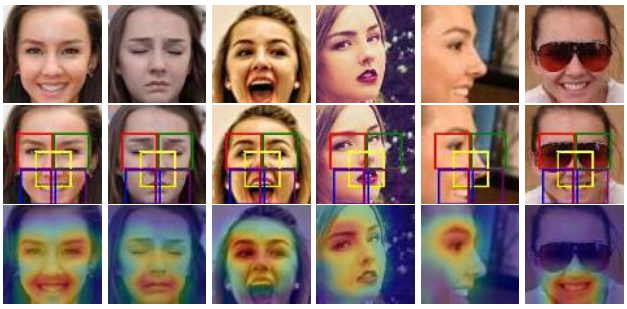

Figure 6. **Qualitative comparison between landmark-based methods and the proposed HPDA model**. **Row 1**: some challenging faces. **Row 2**: even face landmarks are detected, predefined crops may not be flexible and robust under pose, expression and occlusion variations. **Row 3**: the HPDA model can locate discriminative facial parts flexibly.

Previous landmark-based and attention-based local methods are qualitatively compared with the proposed HPDA model. Landmark-based methods crop face parts around landmarks, suffering from noise from adjacent parts or background. For example, as shown in Fig. 6, Columns 1, 2, 3, mouths have varying shapes or sizes (Row 1) due to expression changes. As a consequence, some crops contain noise from noses and background (Row 2). Meanwhile, because of pose variations (Row 1, Columns 4, 5) and occlusions (Row 1, Column 6), some parts are partly or completely invisible. In such cases, background (Row 2, Columns 4, 5) or sunglasses (Row 2, Column 6) are cropped, leading to noisy representations. Moreover, it is noticed that landmark detection may be not accurate, resulting in inaccurate crops (Row 2, Columns 4, 5). For the attention-based models without diverse learning, many unimportant or noisy facial regions are emphasized in Fig. 2, Row 2. In contrast, our proposed HPDA model can locate discriminative parts and suppress less informative, as shown in Row 3, demonstrating its superiority.

4.5. Experiments on Cross-age Face Matching

We compare performances on CALFW and CACD-VS datasets in Table 3. It can be seen that HPDA model can slightly outperform others. For this task, as shown in Fig. 1, Rows 3, 4, even if these faces have undergone dramatic appearance changes, some local facial regions (e.g. the pointy

Methods	IJB-A		CALFW	CACD -VS	CPLFW	VGGFace2 -FP	LFW
	FAR= 0.01	FAR= 0.001					
VGGFace [22]	60.5	36.7	86.50	96.00	-	-	99.13
Center loss [34]	52.1	31.3	85.48	97.48	77.48	75.10	99.28
SphereFace [19]	54.8	39.6	90.30	-	81.40	20.10	99.42
VGGFace2 [1]	-	-	90.57	-	84.00	62.22	99.43
LS-CNN [30] (VGGFace2)	87.5	75.5	92.00	99.50	88.03	69.92	99.52
LS-CNN [30] (MS-Celeb-1M)	87.2	77.5	94.40	99.10	-	-	-
ArcFace [3]	68.6	65.7	95.45	-	92.08	46.20	99.83
Co-mining [32]	-	-	93.28	-	87.31	-	-
MV-Softmax [33]	-	-	95.63	-	89.69	-	99.79
HPDA	87.6	80.3	95.90	99.55	92.35	95.32	99.80

Table 3. Performance comparison (%) of the HPDA model with state-of-the-art methods. LS-CNN model and MS-Celeb-1M dataset are used as the stem CNN and training dataset, respectively. As the baseline, LS-CNN model is run on both VGGFace2 and MS-Celeb-1M datasets.

nose) remain to be very characteristic and the face shape (i.e. the oval face) is still similar. Accordingly, the global CNN (Column 7) and local CNNs (Column 6) focus on facial shapes and pointy noses, respectively.

4.6. Experiments on Cross-pose Face Matching

As shown in Table 3, our HPDA model outperforms the state-of-the-art on CPLFW and VGGFace2-FP datasets. Results on VGGFace2-FP dataset of Center loss, SphereFace and ArcFace models are from [29].

There are several reasons that can explain the results. First, local patches appear at different sizes due to pose changes. Especially, profile faces are self-occlusion where only partial organs are visible, like the small-scale mouths in Fig. 2, Columns 7, 8. The problem above motivates us to develop the pyramid attention to capture multi-scale features. Second, background distraction is a common problem for faces with large pose variations. On one hand, there are some transition regions between discriminative local patches (e.g. noses, eyes, mouths) and background for frontal faces. On the other hand, discriminative local patches are near to noisy background for profile faces. As a result, noisy background tends to be cropped with discriminative local patches, leading to inferior representations. This is especially serious for landmark-based local methods because face landmarks may lie in the background, as shown in Fig. 6, Row 2, Columns 4, 5. Meanwhile, because previous attention-based local methods do not consider the attention diversity, some useful regions are missed, as shown Fig. 2, Columns 7, 8, Row 2. The proposed diverse learning can alleviate this issue by emphasizing discriminative face regions and suppressing less important and background distraction, as illustrated in Row 3.

Besides, the local CNNs extract complementary information to the global CNN. One positive pair with pose changes are shown in Fig. 1, Rows 5, 6, which have similar eyes and big lips. The global CNN describes eyes in Column 7. Meanwhile, the local CNNs focus on big lips in Column 6. The combined local and global CNNs (Column 8) show that the complementarity between the local CNNs and the global CNN is learned.

4.7. Experiments on LFW Dataset

We also report the accuracy on the popular LFW dataset where most faces are frontal or near-frontal. As shown in Table 3, our model is only slightly worse than ArcFace, but outperforms all other models. Although our proposed HPDA model is designed to solve challenging face matching tasks, it still has an excellent generalization ability on LFW dataset.

5. Conclusion

We have proposed the HPDA model for face recognition, which adaptively extracts hierarchical multi-scale local representations. A pyramid attention has been applied to locate multi-scale disiminative face regions automatically and weigh adaptively for multiple facial parts. To capture diverse local facial representations, a diverse learning has been introduced to guide multiple attentions to locate diverse facial parts. The hierarchical bilinear pooling has been used to fuse complementary features from different layers. Experimental results on several very challenging face recognition tasks have validated the effectiveness and importance of our proposed HPDA model.

References

[1] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *FG*, pages 67–74. IEEE, 2018.

[2] Bor-Chun Chen, Chu-Song Chen, and Winston H Hsu. Cross-age reference coding for age-invariant face recognition and retrieval. In *Proceedings of the European Conference on Computer Vision*, pages 768–783. Springer, 2014.

[3] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019.

[4] Changxing Ding and Dacheng Tao. Trunk-branch ensemble convolutional neural networks for video-based face recognition. *TPAMI*, 2017.

[5] Guodong Ding, Salman Khan, Zhenmin Tang, and Fatih Porikli. Let features decide for themselves: Feature mask network for person re-identification. *arXiv preprint arXiv:1711.07155*, 2017.

[6] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3146–3154, 2019.

[7] Guodong Guo and Na Zhang. What is the challenge for deep learning in unconstrained face recognition? In *FG*, pages 436–442. IEEE, 2018.

[8] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *Proceedings of the European Conference on Computer Vision*, pages 87–102. Springer, 2016.

[9] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. *arXiv preprint arXiv:1709.01507*, 7, 2017.

[10] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008.

[11] Lei Huang, Xianglong Liu, Bo Lang, Adams Wei Yu, Yongliang Wang, and Bo Li. Orthogonal weight normalization: Solution to optimization over multiple dependent stiefel manifolds in deep neural networks. In *AAAI*, 2018.

[12] Bong-Nam Kang, Yonghyun Kim, Bongjin Jun, and Daijin Kim. Attentional feature-pair relation networks for accurate face recognition. *Proceedings of the IEEE International Conference on Computer Vision*, 2019.

[13] Bong-Nam Kang, Yonghyun Kim, Bongjin Jun, and Daijin Kim. Hierarchical feature-pair relation networks for face recognition. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[14] Bong-Nam Kang, Yonghyun Kim, and Daijin Kim. Pairwise relational networks for face recognition. *arXiv preprint arXiv:1808.04976*, 2018.

[15] Brendan F Klare, Ben Klein, Emma Taborsky, Austin Blanton, Jordan Cheney, Kristen Allen, Patrick Grother, Alan Mah, and Anil K Jain. Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1931–1939, 2015.

[16] Dangwei Li, Xiaotang Chen, Zhang Zhang, and Kaiqi Huang. Learning deep context-aware features over body and latent parts for person re-identification. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 384–393, 2017.

[17] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2285–2294, 2018.

[18] Jingtuo Liu, Yafeng Deng, Tao Bai, Zhengping Wei, and Chang Huang. Targeting ultimate accuracy: Face recognition via deep embedding. *arXiv preprint arXiv:1506.07310*, 2015.

[19] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. *arXiv preprint arXiv:1704.08063*, 2017.

[20] Volodymyr Mnih, Nicolas Heess, Alex Graves, et al. Recurrent models of visual attention. In *Advances in neural information processing systems*, pages 2204–2212, 2014.

[21] Mete Ozay and Takayuki Okatani. Optimization on submanifolds of convolution kernels in cnns. *arXiv preprint arXiv:1610.07008*, 2016.

[22] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, et al. Deep face recognition. In *BMVC*, volume 1, page 6, 2015.

[23] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation from predicting 10,000 classes. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1891–1898, 2014.

[24] Yifan Sun, Liang Zheng, Weijian Deng, and Shengjin Wang. Svdnet for pedestrian retrieval. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3800–3808, 2017.

[25] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015.

[26] Zichang Tan, Yang Yang, Jun Wan, Guodong Guo, and Stan Z. Li. Deeply-learned hybrid representations for facial age estimation. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 3548–3554. International Joint Conferences on Artificial Intelligence Organization, 7 2019.

[27] Zichang Tan, Yang Yang, Jun Wan, Hanyuan Hang, Guodong Guo, and Stan Z. Li. Attention-based pedestrian attribute analysis. *IEEE TIP*, 28(12):6126–6140, 2019.

[28] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Zhifeng Li, Dihong Gong, Jingchao Zhou, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. *arXiv preprint arXiv:1801.09414*, 2018.

[29] Qiangchang Wang and Guodong Guo. Benchmarking deep learning techniques for face recognition. *Journal of Visual Communication and Image Representation*, 65:102663, 2019.

[30] Q. Wang and G. Guo. Ls-cnn: Characterizing local patches at multiple scales for face recognition. *IEEE Transactions on Information Forensics and Security*, 15:1640–1653, 2020.

[31] Qiangchang Wang, Yuanjie Zheng, Gongping Yang, Weidong Jin, Xinjian Chen, and Yilong Yin. Multiscale rotation-invariant convolutional neural networks for lung texture classification. *IEEE journal of biomedical and health informatics*, 22(1):184–195, 2017.

[32] Xiaobo Wang, Shuo Wang, Jun Wang, Hailin Shi, and Tao Mei. Co-mining: Deep face recognition with noisy labels. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9358–9367, 2019.

[33] Xiaobo Wang, Shifeng Zhang, Shuo Wang, Tianyu Fu, Hailin Shi, and Tao Mei. Mis-classified vector guided softmax loss for face recognition. *arXiv:1912.00833*, 2019.

[34] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *Proceedings of the European Conference on Computer Vision*, pages 499–515. Springer, 2016.

[35] Tianyi Wu, Sheng Tang, Rui Zhang, Juan Cao, and Jintao Li. Tree-structured kronecker convolutional network for semantic segmentation. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 940–945. IEEE, 2019.

[36] Tianyi Wu, Sheng Tang, Rui Zhang, Guodong Guo, and Yongdong Zhang. Consensus feature network for scene parsing. *arXiv preprint arXiv:1907.12411*, 2019.

[37] Tianyi Wu, Sheng Tang, Rui Zhang, and Yongdong Zhang. Cgnet: A light-weight context guided network for semantic segmentation. *arXiv preprint arXiv:1811.08201*, 2018.

[38] Weidi Xie, Li Shen, and Andrew Zisserman. Comparator networks. In *Proceedings of the European Conference on Computer Vision*, pages 782–797, 2018.

[39] Jiwei Yang, Xu Shen, Xinmei Tian, Houqiang Li, Jianqiang Huang, and Xian-Sheng Hua. Local convolutional neural networks for person re-identification. In *ACM Multimedia Conference on Multimedia Conference*, pages 1074–1082. ACM, 2018.

[40] Hantao Yao, Shiliang Zhang, Richang Hong, Yongdong Zhang, Changsheng Xu, and Qi Tian. Deep representation learning with part loss for person re-identification. *TIP*, 28(6):2860–2871, 2019.

[41] Chaojian Yu, Xinyi Zhao, Qi Zheng, Peng Zhang, and Xinge You. Hierarchical bilinear pooling for fine-grained visual recognition. In *Proceedings of the European Conference on Computer Vision*, pages 574–589, 2018.

[42] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.

[43] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.

[44] Heliang Zheng, Jianlong Fu, Tao Mei, and Jiebo Luo. Learning multi-attention convolutional neural network for fine-grained image recognition. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5209–5217, 2017.

[45] Tianyue Zheng and Weihong Deng. Cross-pose lfw: A database for studying crosspose face recognition in unconstrained environments. *Beijing University of Posts and Telecommunications, Tech. Rep*, pages 18–01, 2018.

[46] Tianyue Zheng, Weihong Deng, and Jiani Hu. Cross-age lfw: A database for studying cross-age face recognition in unconstrained environments. *arXiv preprint arXiv:1708.08197*, 2017.