# Deformable Face Net: Learning Pose Invariant Feature with Pose Aware Feature Alignment for Face Recognition

Mingjie He[1,2] Jie Zhang[1] Shiguang Shan[1,2] Meina Kan[1] Xilin Chen[1,2]
[1] Key Lab 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

***Abstract*— Face recognition plays an important role in computer vision. It still remains a challenging task due to pose, expression, illumination, partial occlusion, etc. In this work, we propose a novel Deformable Face Net (DFN) to handle the pose variations in face recognition. The Deformable Face Net introduces deformable convolution modules to simultaneously learn face recognition oriented alignment and feature extraction. Specifically, two loss functions, namely displacement consistency loss (DCL) and identity consistency loss (ICL) are designed to minimize the intra-class feature variation caused by different poses. These two loss functions jointly learn pose-aware displacement fields for deformable convolutions in the DFN. Different from the existing methods, the DFN focuses on aligning features across different poses rather than frontalizing the input faces. Extensive experiments show that the proposed DFN outperforms the state-of-the-art methods, especially on the datasets with large poses.**

## I. INTRODUCTION

Face recognition, as a key topic in computer vision, has received more and more attentions in recent years. Equipped with powerful convolutional neural networks (CNNs), the accuracy has a rapid boost that face recognition under controlled settings (i.e., near-frontal poses, neutral expressions, normal illuminations, etc.) seems to be solved. The conventional deep face recognition system firstly aligns faces with affine transformations and then feeds the aligned faces into convolutional neural networks to extract identity-preserving features. Since the affine transformations can only remove in-plane pose influences, the face recognition accuracy degenerates severely under large out-plane pose variations. Face recognition under large poses still remains a challenging problem.

To handle large pose variations, enlarging the training datasets with faces under diversified poses may be an effective way to obtain features robust to different poses. However, such training sets of a mass of identities are extremely rare. Alternatively, the works in [7], [29] enrich the diversity of poses by synthesizing massive images of sufficient pose variability from a frontal face. To some extent, the above methods relieve the pose influence, but it still leaves a long way to go. Recently, many efforts are devoted to exploring pose invariant face recognition (PIFR) methods, which can be roughly grouped into the following two categories: face frontalization methods and pose-invariant feature learning methods.

978-1-7281-0089-0/19/$31.00 ©2019 IEEE

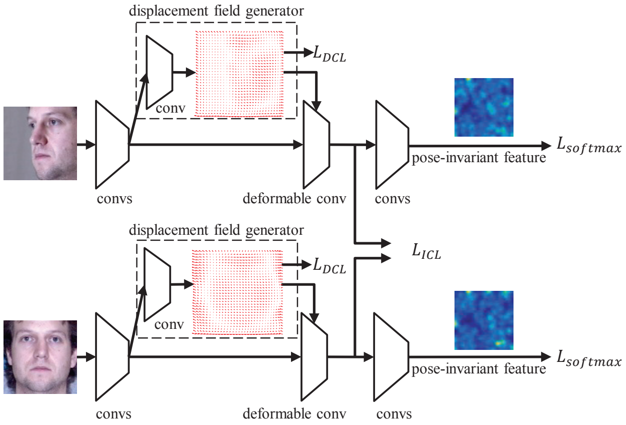


Fig. 1. Illustration of our proposed Deformable Face Net (DFN). DFN attempts to learn a pose-aware displacement field for the deformable convolution to extract pose-invariant features for face recognition. The pose-aware displacement field is learnt by introducing the displacement consistency loss (DCL) and the identity consistency loss (ICL) for minimizing the intra-class feature variation under different poses.

### A. Face Frontalization Methods

The key point of nearly all face frontalization methods is how to construct a frontal face from faces under diverse poses. In terms of the generation ways, these methods are generally categorized into synthesizing frontal faces with 3D information [12], [44], [2], [30], [23], [9], [3] and 2D images [17], [19], [18], [35]. For the first category, [12] proposes an effective face frontalization approach by using a single and unchanged 3D shape to approximate the shape of all the input faces. In [44], a high-fidelity pose and expression normalization method with 3D Morphable Model (3DMM) is proposed to generate a frontal face under neutral expression. Without using the 3D structure model, the promising image synthesis approach Generative Adversarial Network (GAN) has also been used to frontalize faces [18], [43], [40], [35]. By modeling the face rotation process, DR-GAN [35] learns a disentangled representation which can frontalize extreme poses in the wild.

The face frontalization methods above have shown promising results of transforming non-frontal faces to frontal ones. Since the original non-frontal images has invisible face pixels due to self-occlusion, the details of the transformed faces highly rely on the invisible region filling approaches. Even though the facial structure is symmetrical, the symmetry of illumination cannot always hold. Both the blurry details and the weird illumination may make the transformed images unreal under large poses. For GAN based methods, the identity information may not be well preserved for the synthetic faces. Although current methods have improved the illumination trends and the texture details, the quality of the transformed images is still far from avoiding degeneration of face recognition performance.

### B. Pose-Invariant Feature Learning Methods

These methods focus on learning pose-invariant feature representations for face recognition in the wild. Conventional multiview subspace approaches learn complex nonlinear transformations that respectively project images captured under different poses to the common space, where the intra-class variation is minimized [16], [31], [22], [1], [32], [33], [20]. For instance, [32] presents a discriminant coupled latent subspace framework for pose-invariant discriminative learning. In [33], GMA extracts unified multiview features by optimizing view-specific projections. In [20], MvDA is proposed to jointly solve the multiple linear transforms and meanwhile minimize the within-class variations, resulting in very encouraging performance.

Recently, more works resort to the deep learning to extract more powerful pose-invariant features [41], [26], [37], [45], [8], [42], [39], [28], [25]. To address the above mentioned problem, one may either group multiple pose-specific models or pose-specific activations, i.e., each one corresponding to a specific pose [26], [37], [39], [25] or design a single pose-invariant model [28], [45], [8], which uniformly tackles all poses. For the former category, [37] proposes a pose-directed multi-task CNN to learn pose-specific identity features. Similarly, in [25], a face image is processed by utilizing several pose-specific deep convolution neural networks. Although a significant improvement in accuracy has been witnessed, the efficiency concern of such a multi-model framework needs to be further tackled. For the other category, a unified model is exploited to extract pose-invariant features. For instance, an analytic Gabor feedforward network is proposed in [28] to absorb moderate changes caused by poses. In [42], a face frontalization sub-net (FFN) and a discriminative learning sub-net (DLN) is aggregated at a pose invariant model (PIM) which generates both high fidelity frontalized face images and pose invariant facial representations. The PIM can also be categorized as a hybrid approach which combines the face synthesis and the pose-robust feature extraction. In consideration of complementary advantages, the hybrid framework trends to be the most promising approach for PIFR problems.

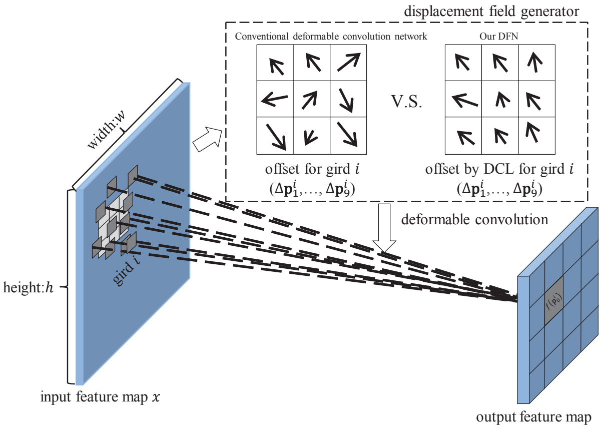

displacement field generator
Conventional deformable convolution network
Our DFN
V.S.
offset for gird $i$
$(\Delta\mathbf{p}_1^i,\ldots,\Delta\mathbf{p}_9^i)$
offset by DCL for gird $i$
$(\Delta\mathbf{p}_1^i,\ldots,\Delta\mathbf{p}_9^i)$
deformable convolution
width:$w$
height:$h$
gird $i$
$f(\mathbf{p}_0^i)$
input feature map $x$
output feature map

Fig. 2. Illustration of the offset obtained with our displacement consistency loss (DCL).

### C. The Proposed Approach

As illustrated above, the pixel-level frontalization methods can generate a frontal face under different poses but it may be difficult for these approaches to strongly preserve the identity information, leading to performance degeneration. On the contrary, the pose-invariant feature learning approaches emphasize on projecting faces into new feature space to eliminate the influence of poses rather than frontalizing faces. However, it is non-trivial to obtain pose-invariant feature robust to complex scenarios, due to the lack of explicit alignment process.

In this paper, we further push the frontier of the area by proposing a pose-robust feature extraction neural network with explicitly considering the feature-level alignment. Inspired by the deformable convolution [6], we carefully design the deformable convolution modules to implement feature-level geometric transformations for face recognition oriented alignment, named as Deformable Face Net (DFN), as illustrated in Fig. 1. Moreover, two loss functions, namely displacement consistency loss (DCL) and identity consistency loss (ICL) are designed to learn more accurate pose-aware displacement fields in feature-level. Specifically, the DCL enforces the learnt displacement fields to be locally consistent both in the orientation and amplitude since faces possess strong structures. The ICL is designed to minimize the intra-class feature variation caused by different poses via taking two faces under different poses as input. Besides, the DFN is quite efficient and can be end-to-end trained without additional supervision. Extensive experiments demonstrate that the feature-level alignment is more effective than the pixel-level frontalization and more robust pose-invariant features can be obtained by coupling feature-level alignment.

Briefly, the main contributions of this paper are summarized as follows:

- A novel Deformable Face Net (DFN) is proposed to handle pose variations in face recognition with explicitly considering the feature-level alignment.

- Two loss functions, namely displacement consistency loss (DCL) and identity consistency loss (ICL) are designed to minimize the intra-class feature variation caused by different poses, leading to better performance for face recognition.
- DFN outperforms the state-of-the-art methods on MegaFace and MultiPIE, especially on the MultiPIE dataset with large poses.

## II. Method

The proposed Deformable Face Net (DFN) attempts to learn a pose-aware displacement field in deformable convolutions to extract pose-invariant features for face recognition. The objective is that the spatial displacement field of deformable convolutions should be adaptively adjusted according to the head pose. These displacement fields are jointly learnt by introducing two novel loss functions, i.e., displacement consistency loss (DCL) and the identity consistency loss (ICL). In this way, the DFN is able to well tackle the feature misalignment issue caused by poses, resulting in performance improvement in face recognition.

### A. Overview of DFN

As shown in Fig. 1, DFN takes paired images as input, of which each pair contains two faces randomly sampled from the same person. It should be noted that the two faces are not limited to one frontal image and one non-frontal image. A displacement field generator consisting of a convolutional layer learns displacement fields from low-level features for face recognition oriented alignment. In consideration of the strong structure lying in faces, the displacement consistency loss (DCL) is proposed to improve the locality consistency of the learnt displacement fields. Moreover, the identity consistency loss (ICL) is proposed to minimize the intra-class feature variation, so as to explicitly force the learnt displacement fields to well align features under different poses. The whole network is end-to-end trained jointly by using the softmax classification loss and the proposed two loss functions recorded as DCL and ICL. The proposed method can be integrated with the existing powerful CNN architectures, e.g., the ResNet architecture [13], [14]. We note that introducing the pose-aware deformation modules at different layers of the network have significant difference in performance. Details will be discussed in Sec. III. Next, we present each component of the DFN in details.

### B. Displacement Consistency Loss

Given an input feature map $x$, the kernels of the deformable convolution [6] sample irregular grids over the input $x$. For gird $i$ centered on location $\mathbf{p}_0^i$, such irregular sampling locations are obtained by an addition of offsets $\{\Delta\mathbf{p}_k^i = \{\Delta p_{kx}^i, \Delta p_{ky}^i\} | k = 1, ..., K\}$ (i.e., a displacement field) to a regular sampling grid $\mathcal{R}$. $\Delta p_{kx}^i$ and $\Delta p_{ky}^i$ denote the x-axis and the y-axis component of $\Delta\mathbf{p}_k^i$ respectively. The size of $\mathcal{R}$ is $K$, e.g., $K = 9$ for $3 \times 3$ convolution kernels. Then, the output feature map $f$ of the deformable convolution is computed as bellows:

$$f(\mathbf{p}_0^i) = \sum_{k=1}^{K} \mathbf{w}(\mathbf{p}_k^i) \cdot x(\mathbf{p}_0^i + \mathbf{p}_k^i + \Delta\mathbf{p}_k^i), \tag{1}$$

where $\mathcal{R} = \{(-1,-1), (-1,0), ..., (0,1), (1,1)\}$ for a $3\times3$ kernel, $\mathbf{p}_k^i$ enumerates the locations in $\mathcal{R}$ and $\mathbf{w}$ denotes the convolution kernel. The offsets are represented as a $h\times w\times 2K$ tensor for a $h\times w$ input feature map with stride 1. The spatial dimension $h\times w$ corresponds to the sliding sampling grids of the convolution operations and the channel dimension $2K$ corresponds to $K$ offsets for each sampling grid $\mathcal{R}$.

To solve the PIFR problem, we expect that all the $h\times w\times 2K$ offsets to compensate both rigid and non-rigid global geometric transformations, such as poses and expressions. Since the general objects have diverse local and global transformations in the wild, it is reasonable to learn those offsets without additional constraints for conventional object detections. However, different faces share the same structure and the most visible transformation is caused by the poses, which means the deformation module should focus more on the distribution of the global displacement field along the spatial dimension of the input feature maps. Moreover, redundant capacity of modeling the local transformations increases the risk of over-fitting potentially, especially for the face images. To be free from this, the displacement consistency loss (DCL) is proposed to learn the displacement field within each grid towards a consistent direction, as shown in Fig. 2. The DCL is formulated in Eq. (2) as:

$$L_{DCL} = \frac{1}{h \times w \times K} \sum_{i=1}^{h\times w} \sum_{k=1}^{K} \|\Delta\mathbf{p}_k^i - \Delta\overline{\mathbf{p}}^i\|_2^2, \tag{2}$$

where $\Delta\overline{\mathbf{p}}^i$ is the mean offset along $k$ for $i$-th grid. By limiting the solution searching space of the displacement field, the DCL makes the training process more feasible, meanwhile the obtained displacement field drives the deformable convolutions to well compensate the intra-class feature variation caused by poses.

### C. Identity Consistency Loss

The final objective of PIFR is to learn robust features that the difference across poses is minimized as much as possible. It is natural to introduce the Euclidean distance loss such as the contrastive loss [11], [5], whose minimization can pull the features of the same identity under different conditions (e.g., poses) together. Moreover, the formulation of pair-wise Euclidean distance loss is frequently applied to face recognition. However, due to the limited geometric transformation capacity of conventional CNN structures, the pair-wise loss function is not always helpful. On the contrary, DFN can naturally handle this problem more efficiently, benefited from the pose-aware deformation modules. In this paper, we reformulate the Euclidean distance loss as the identity consistency loss (ICL) by constraining the distance between features of the same person from the deformable convolutions rather than final features from the penultimate layer. In this way, the identity consistency loss has more profound supervision effects on learning the deformable offsets such that the PIFR can be further improved.

Formally, to train the DFN, a training batch containing $N$ images is randomly chosen from $N/2$ identities, where two images for the identity $j$, namely $\mathbf{I}_1^j$ and $\mathbf{I}_2^j$. The identity consistency loss minimizes the difference between the output deformable features $\mathbf{f}_1^j$ and $\mathbf{f}_2^j$ corresponding to the input images $\mathbf{I}_1^j$ and $\mathbf{I}_2^j$ respectively, i.e.,

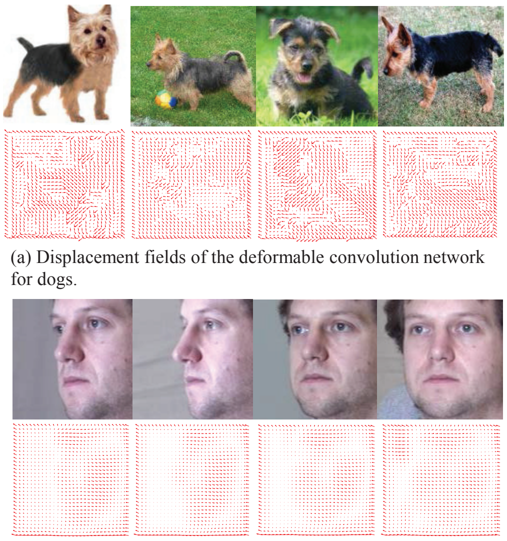
$$L_{ICL} = \sum_{j=1}^{N/2} \|\mathbf{f}_1^j - \mathbf{f}_2^j\|_2^2. \tag{3}$$

It should be noted that the normalization of $\mathbf{f}_1^j$ and $\mathbf{f}_2^j$ is necessary, otherwise the norm of features will implicitly affect the scale of the loss function, leading to un-convergence. By employing the ICL, the deformable module is optimized to enforce features under varied poses to be well aligned.

### D. Discussion

*1) Differences with the deformable convolution network:* Both the deformable convolution network [6] and our DFN are feature-level alignment methods that attempt to handle the geometric transformations. The deformable convolution is firstly developed for detecting general objects which have diverse local and global non-rigid transformations, e.g., the dogs shown in Fig. 3 have significantly different postures. In contrast, human faces are approximately rigid objects and the most viable transformations are caused by the rigid pose variations rather than the non-rigid expressions, which means the displacement field learnt for face recognition should be more consistent in directions. To this end, two addition loss functions DCL and ICL are embedded in deformable convolution networks for better face alignment. As illustrated in Fig. 3, the displacement fields of faces from our DFN are more consistent than those of dogs from deformable convolution networks, which are more favorable for face recognition oriented face alignment. The significant improvements in face recognition further demonstrate the effectiveness of our DFN, see details in Sec. III-C.

*2) Differences with the face frontalization methods:* The face frontalization methods [12], [44], [2], [30], [23], [9], [3], [17], [19], [18], [43], [40], [35] resort to do pixel-level alignment that attempts to generate frontal faces, while our DFN performs feature-level alignment that attempts to align features under different poses. For face recognition, the generated frontal faces are further fed into CNNs for feature extraction, resulting in a two-stage process (i.e., the face frontalizaition and the feature extraction). Differently, our method learns the pose invariant features in a unified framework by designing an effective feature-level deformable convolutional module, leading to better recognition results.



(a) Displacement fields of the deformable convolution network for dogs.

(b) Displacement fields of our DFN for faces.

Fig. 3. Illustration of the displacement fields. As seen, adjacent offsets share similar direction, meaning that local consistency inheres in the distribution of displacement field. However, for rigid human head, such local consistency is greater than the non-rigid dogs, meaning that the offsets for larger face area share the similar direction.

*3) Differences with other pose-invariant feature learning methods:* Different from most pose-invariant feature leaning methods [22], [1], [32], [33], [20], [41], [26], [37], [45], [8], [42], [39], [28], [25] using multiple models in which each model correspond to a specific pose, our DFN present a unified model to handle different poses. Besides, those subspace learning approaches [22], [1], [32], [33], [20] directly learn projections to achieve pose-invariant features. Since such projections are learnt corresponding to several specific poses, those methods are limited to handle these discrete poses. Besides, it may be non-trivial for those methods to obtain features robust to more complex pose variations without explicitly considering alignments. Differently, our method can tackle arbitrary poses rather than several specific poses. Furthermore, our method learns pose-invariant features in consideration of explicit feature-level alignments, resulting in significant improvement for face recognition across poses.

## III. Experiments

### A. Experimental Setting

*1) Dataset:* To investigate the effectiveness of the proposed DFN, the MegaFace [21] benchmark is employed for the evaluations as this challenging benchmark contains more than 1 million face images among which more than 197K faces have yaw angles larger than ±40 degrees. In this study, we evaluate the performance of our approach on Megaface challenge 1 (MF1). The gallery set in MegaFace datasets consists more than 1 million face images from 690k different individuals. The Facescrub dataset [27] containing 106,863 face images of 530 people is used as

TABLE I
ARCHITECTURE DETAILS OF DFN (DCL&ICL).

| Output size | DFN-Light (DCL&ICL) | DFN-ResNet-50 (DCL&ICL) | DFN-ResNet-152 (DCL&ICL) |
|---|---|---|---|
| 62×62 | conv,7 × 7, 64, stride 4 | conv, 7 × 7, 64, stride 4 | conv, 7 × 7, 64, stride 2<br>*displacement field generator*<br>*with DCL and ICL*<br>*deformable conv,* 3 × 3, *64, stride 1*<br>max pool, 3 × 3, stride 2 |
| stage 1<br>31×31 | max pool, 3 × 3, stride 2<br>*displacement field generator*<br>*with DCL and ICL*<br>*deformable conv,* 3 × 3, *64, stride 1* | max pool, 3 × 3, stride 2<br>$\begin{bmatrix}\text{conv}, 1\times 1, 64\\ \text{conv}, 3\times 3, 64\\ \text{conv}, 1\times 1, 256\end{bmatrix}\times 3$ | $\begin{bmatrix}\text{conv}, 1\times 1, 64\\ \text{conv}, 3\times 3, 64\\ \text{conv}, 1\times 1, 256\end{bmatrix}\times 3$ |
| stage 2<br>16×16 | conv, 3 × 3, 64, stride 2 | *displacement field generator*<br>*with DCL and ICL*<br>*deformable conv,* 3 × 3, *64, stride 1*<br>$\begin{bmatrix}\text{conv}, 1\times 1, 128\\ \text{conv}, 3\times 3, 128\\ \text{conv}, 1\times 1, 512\end{bmatrix}\times 4$ | $\begin{bmatrix}\text{conv}, 1\times 1, 128\\ \text{conv}, 3\times 3, 128\\ \text{conv}, 1\times 1, 512\end{bmatrix}\times 8$ |
| stage 3<br>8×8 | conv, 3 × 3, 128, stride 2 | $\begin{bmatrix}\text{conv}, 1\times 1, 128\\ \text{conv}, 3\times 3, 128\\ \text{conv}, 1\times 1, 512\end{bmatrix}\times 6$ | $\begin{bmatrix}\text{conv}, 1\times 1, 128\\ \text{conv}, 3\times 3, 128\\ \text{conv}, 1\times 1, 512\end{bmatrix}\times 36$ |
| stage 4<br>1×1024 | conv, 3 × 3, 128, stride 2<br>avg pool, 4 × 4, stride 1<br>fc, 1024 | $\begin{bmatrix}\text{conv}, 1\times 1, 512\\ \text{conv}, 3\times 3, 512\\ \text{conv}, 1\times 1, 2048\end{bmatrix}\times 3$<br>avg pool, 4 × 4, stride 1<br>fc, 1024 | $\begin{bmatrix}\text{conv}, 1\times 1, 512\\ \text{conv}, 3\times 3, 512\\ \text{conv}, 1\times 1, 2048\end{bmatrix}\times 3$<br>avg pool, 7 × 7, stride 1<br>fc, 1024 |
| 1×1 | fc, softmax loss | | |

the probe set. It should be noted that the uncleaned MegaFace datasets are used in evaluation for fair comparison.

To systematically evaluate how our DFN handles various pose angles, we conduct experiments on the MultiPIE dataset as it contains images captured with varying poses. The MultiPIE dataset is recorded during four sessions and contains images of 337 identities under 15 view points and 20 illumination levels. To compare with state-of-the-arts, we employ the following setting since it is an extremely challenging setting with more pose variations. The setting follows the protocol introduced in [45], [37], images of 250 identities in session one are used. For training, we utilize the images of the first 150 identities with 20 illumination levels and poses ranging from $+90^{\circ}$ to $-90^{\circ}$. For testing, one frontal image with neutral expression and illumination is used as the gallery image for each of the remaining 100 identities and the other images are used as probes. The rank-1 recognition rate is used as the measurement of the face recognition performance.

*2) Implementation Details:* In our experiments, we use [15] for landmark detection and crop the face images into size of $256 \times 256$ by affine transformations. Some examples of the cropped images are shown in Fig. 4. The DFNs are constructed by integrating the deformable module between two adjacent original CNN blocks and trained with the softmax loss function. It is flexible to be directly applied to the standard CNNs so that we develop DFN-ResNets by stacking it into two adjacent residual blocks of the ResNets. Extensive experiments are conducted to explore the impact of the deformable module integrated at different stages of the ResNet architectures. The DFN (DCL) and DFN (ICL) denote the DFN versions trained with the proposed DCL and ICL respectively. The DFN (DCL&ICL) denotes the version trained with the two loss function jointly. For the MegaFace evaluation, the conventional ResNet-50 and ResNet-152 are used as our baselines. We manually clean the MS-Celeb-1M [10] dataset and finally collect 3.7 Million images from 50K identities. The revised dataset is used as our training set for the evaluations on MegaFace challenge 1. For experiments on MultiPIE dataset, the limited amount of training images may incur over-fitting issue for deep networks like ResNet-50/152. To this end, we design a light CNN, namely, DFN-Light which is pre-trained on the cleaned MS-Celeb-1M dataset and then fine-tuned on the MultiPIE training set. The baseline denotes the plain network without the deformable modules and the proposed two losses. The architecture details are summarized in Table I. We implement our method on the MXNet [4] platform and train all the models using SGD with four NVIDIA TITAN XP GPUs. The loss weight of the softmax loss is set to 1 and the loss weights of DCL and ICL are 0.001 and 0.01 respectively.

### *B. Evaluations on the MegaFace Benchmark*

Since the stage where the deformable convolution is integrated plays an important role in the resulting network architectures, we firstly conduct experiments to investigate the best construction with only softmax loss. By integrating the deformable convolution at four different stages of the plain ResNet-50 respectively, we construct four versions of

TABLE II

Rank-1 identification accuracy on MegaFace challenge 1 with deformable convolution embedded at different stages.

| Method | MF1 Rank1 |
|---|---|
| Baseline ResNet-50 | 74.76 |
| DFN-50 with deformable conv embedded in stage1 | 75.25 |
| DFN-50 with deformable conv embedded in stage2 | 75.02 |
| DFN-50 with deformable conv embedded in stage3 | 72.48 |
| DFN-50 with deformable conv embedded in stage4 | 58.88 |

TABLE III

Rank-1 identification accuracy on MegaFace challenge 1 with different loss functions.

| Loss | MF1 Rank1 |
|---|---|
| DFN-50: softmax | 75.02 |
| DFN-50: softmax + contrastive | 76.82 |
| DFN-50: softmax + ICL | 78.14 |
| DFN-50: softmax + DCL | 77.51 |
| DFN-50: softmax + DCL + ICL | 78.21 |

the DFN-ResNet-50 (DFN-50 for short in the following sections). Table I exhibits an example of integrating the deformable module in the stage 2. One significant difference between these four versions is size of the input feature map which varies from 62 × 62 to 8 × 8. We train the four versions on the 3.7 Million images and test them on the MegaFace challenge 1 benchmark. As illustrated in Table II, the performance is gradually improved from stage 4 to stage 1, which means the deformable convolution works better on larger input feature maps from the shallow stage. Since the size of the receptive field in the shallow stage is much smaller than that in the deep stage, the learnt displacement field of the shallow stage is more elaborative, leading to better alignment for face recognition.

Furthermore, integrating the deformable module in shallow stage significantly outperforms the baseline, indicating that the DFN is superior to its plain version, i.e., the ResNet-50 baseline. Since models in Table II are trained only with the softmax loss, the capability of DFN has not been fully excavated. Here, we further explore the effectiveness of applying the DCL and ICL loss functions to DFN. Firstly, we train the DFN-50 integrated the deformable module in stage 2 with the DCL and ICL respectively. Then, we train the same network structure with both the DCL and ICL.

Table III. summarizes the rank-1 identification accuracy on MegaFace challenge 1 of our models trained with the proposed DCL and ICL loss functions. When the two loss functions are used separately, both of them can significantly improve the performance, which demonstrates the effectiveness of the two proposed loss functions. Specifically, when only using the DCL loss, the rank-1 accuracy is improved by 2.49%. We also compare the proposed ICL with the contrastive loss function. As shown in III, both the ICL and the contrastive loss improve the rank-1 accuracy and our ICL outperforms the conventional contrastive loss by 1.32%. It is reasonable that the conventional contrastive loss function is usually applied at the penultimate layer, which may weaken the effect of the loss function to well align faces under poses. On the contrary, our ICL is applied directly after the deformable module, enforcing the transformed features to be well aligned for better face recognition. Furthermore, by employing the ICL and DCL jointly, the performance of DFN-50 is further improved to 78.21% which outperforms the plain ResNet-50 by 3.45%.

TABLE IV

Rank-1 identification accuracy on MegaFace challenge 1 compared to the state-of-the-art methods.

| Method | MF1 Rank1 |
|---|---|
| SphereFace-Small [24] | 75.76 |
| CosFace [36] | 82.72 |
| ResNet-152 | 80.60 |
| DFN-152 | 80.99 |
| DFN-152 (ICL) | 81.85 |
| DFN-152 (DCL) | 81.53 |
| DFN-152 (DCL&ICL) | 82.11 |

We then evaluate the DFN with deeper architectures. The DFN-ResNet-152 (DFN-152 for short in the following sections) and its corresponding plain ResNet-152 are trained under the same optimization scheme. Table IV shows the results of different networks on MegaFace challenge 1. Similar to the observation under the DFN-50, the performance of DFN-152 is consistently improved with the proposed loss functions. Trained with only 50K identities, our DFN-152 (DCL&ICL) achieves the comparable result with CosFace [36] whose training set contains more faces of 90K identities. Moreover, compared to the ResNet-152, our DFN-152 (DCL&ICL) improves the rank-1 accuracy by 1.51% with only 0.2M extra parameters.

### C. Evaluations on the MultiPIE Benchmark

Table V summarizes the face recognition accuracy of our DFN-Light (DFN-L for short) on multi-PIE for different poses. The results of other state-of-the-arts are directly quoted from [38], [12], [8], [18], [34], [45], [37], [39], [42]. As seen from Table V, the face frontalization method Hassner [12] performs better than CPF [38] since 3D facial shapes are utilized for the face synthesizing. Furthermore,

TABLE V

Rank-1 recognition rates (%) on MultiPIE for different poses

| Method | ±90° | ±75° | ±60° | ±45° | ±30° | ±15° |
|---|---|---|---|---|---|---|
| CPF [38] | - | - | - | 71.65 | 81.05 | 89.45 |
| Hassner [12] | - | - | 44.81 | 74.68 | 89.59 | 96.78 |
| FV [34] | 24.53 | 45.51 | 68.71 | 80.33 | 87.21 | 93.30 |
| HPN [8] | 29.82 | 47.57 | 61.24 | 72.77 | 78.26 | 84.23 |
| FIP [45] | 31.37 | 49.10 | 69.75 | 85.54 | 92.98 | 96.30 |
| c-CNN [37] | 47.26 | 60.66 | 74.38 | 89.02 | 94.05 | 96.97 |
| TP-GAN [18] | 64.03 | 84.10 | 92.93 | 98.58 | 99.85 | 99.78 |
| PIM [42] | 75.00 | 91.20 | 97.70 | 98.30 | 99.40 | 99.80 |
| p-CNN [39] | 76.96 | 87.83 | 92.07 | 90.34 | 98.01 | 99.19 |
| Baseline | 74.22 | 80.40 | 89.30 | 95.59 | 97.83 | 98.39 |
| DFN-L | 82.42 | 87.64 | 94.44 | 97.76 | 98.88 | 99.22 |
| DFN-L (ICL) | 83.65 | 88.62 | 94.97 | 98.00 | 99.12 | 99.51 |
| DFN-L (DCL) | 83.71 | 88.59 | 94.68 | 97.87 | 99.15 | 99.47 |
| DFN-L (DCL&ICL) | 84.07 | 88.97 | 95.16 | 98.05 | 99.23 | 99.58 |

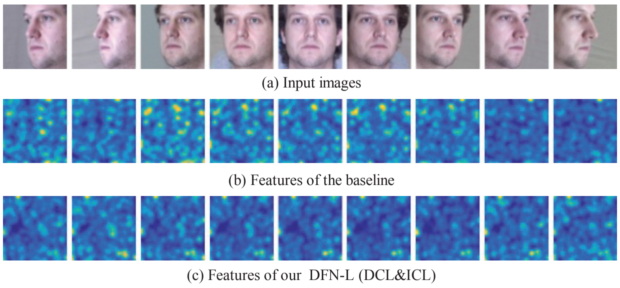

(a) Input images
(b) Features of the baseline
(c) Features of our DFN-L (DCL&ICL)

Fig. 4. An example of pose-invariant features of DFN-L (DCL&ICL) with various poses ($-60°$ to $+60°$). Even with the same identity, obvious differences are witnessed between features extracted from the baseline method. In contrast, the features obtained by the proposed DFN-L (DCL&ICL) show a similar pattern across all poses.

benefitting from the patch based reconstruction and occlusion detection, HPN [8] achieves better results than [38] and [12]. Attributed to the powerful generation ability of GAN, the TP-GAN [18] outperforms all previous face frontalization methods. Differently, the methods of FV [34], FIP [45], c-CNN [37], p-CNN [39] and PIM [42] focus on pose-invariant feature learning. Among them, the deep methods FIP [45], c-CNN [37] and p-CNN [39] outperform the traditional feature representation method FV [34]. Furthermore, owing to learning pose-specific models or pose-specific adaptive routes, the c-CNN and p-CNN perform much better than the unified model FIP. By integrating face frontalization and discriminative feature learning, the PIM [42] achieves almost the best results among the existing methods except the $\pm 90°$. The reason is that as PIM is a face frontalization method, it may be hard for it to well maintain the realness of synthesis, especially on the pose of +/-90 degrees. As seen, our DFN-L significantly outperforms the p-CNN for all poses, demonstrating the effectiveness of introducing deformable convolutions for face recognition oriented alignment. Besides, attributed to the joint leaning with the proposed DCL and ICL loss functions, our DFN-L (DCL&ICL) achieves better results than p-CNN [39] with an improvement up to 7.11% for $\pm 90°$. As shown in Fig. 4, the features extracted by our DFN have a similar pattern across all poses, while obvious differences are witnessed between features extracted from the baseline, which demonstrates the superiority of our DFN again. Moreover, the DFN-L (DCL&ICL) achieves the comparable results with PIM and significantly outperforms PIM with an improvement up to 9% under faces of $\pm 90°$. It is worth noting that the DFN-L has a very light network structure (as shown in Table I), which is much more efficient than the GAN based PIM.

## IV. CONCLUSIONS

To deal with the pose invariant face recognition problem, we proposed a novel Deformable Face Net (DFN) to align features across different poses. To achieve the feature-level alignments, the proposed method, DFN introduces deformable convolution modules to simultaneously learn face recognition oriented alignment and feature extraction. Besides, two loss functions, namely displacement consistency loss (DCL) and identity consistency loss (ICL) are designed to learn pose-aware displacement fields for deformable convolutions in DFN and consequently minimize the intra-class feature variation caused by different poses. Extensive experiments show that the proposed DFN achieves quite promising performance with relatively light network structure, especially for those large poses.

## V. ACKNOWLEDGMENTS

This work was partially supported by Natural Science Foundation of China under contracts Nos. 61806188 and Beijing Municipal Science and Technology Program: Z181100003918012.

## REFERENCES

[1] G. Andrew, R. Arora, J. Bilmes, and K. Livescu. Deep canonical correlation analysis. In *International Conference on Machine Learning (ICML)*, 2013.
[2] A. Asthana, T. K. Marks, M. J. Jones, K. H. Tieu, and M. Rohith. Fully automatic pose-invariant face recognition via 3d pose normalization. In *IEEE International Conference on Computer Vision (ICCV)*, 2011.
[3] J. Cao, Y. Hu, H. Zhang, R. He, and Z. Sun. Learning a high fidelity pose invariant model for high-resolution face frontalization. In *Advances in Neural Information Processing Systems (NIPS)*, 2018.
[4] T. Chen, M. Li, Y. Li, M. Lin, N. Wang, M. Wang, T. Xiao, B. Xu, C. Zhang, and Z. Zhang. MXNet: A flexible and efficient machine learning library for heterogeneous distributed systems. *arXiv preprint arXiv:1512.01274*, 2015.
[5] S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.
[6] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei. Deformable convolutional networks. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.
[7] W. Deng, J. Hu, Z. Wu, and J. Guo. From one to many: Pose-aware metric learning for single-sample face recognition. *Pattern Recognition*, 2018.
[8] C. Ding and D. Tao. Pose-invariant face recognition with homography-based normalization. *Pattern Recognition*, 2017.
[9] C. Ding, C. Xu, and D. Tao. Multi-task pose-invariant face recognition. *IEEE Transactions on Image Processing (TIP)*, 2015.
[10] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. MS-Celeb-1M: A dataset and benchmark for large scale face recognition. In *European Conference on Computer Vision (ECCV)*, 2016.
[11] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006.
[12] T. Hassner, S. Harel, E. Paz, and R. Enbar. Effective face frontalization in unconstrained images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.
[13] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
[14] K. He, X. Zhang, S. Ren, and J. Sun. Identity mappings in deep residual networks. In *European Conference on Computer Vision (ECCV)*, 2016.
[15] Z. He, M. Kan, J. Zhang, X. Chen, and S. Shan. A fully end-to-end cascaded cnn for facial landmark detection. In *IEEE International Conference on Automatic Face Gesture Recognition (FG)*, 2017.
[16] H. Hotelling. Relations between two sets of variates. *Biometrika*, 1936.
[17] L. Hu, M. Kan, S. Shan, X. Song, and X. Chen. LDF-Net: Learning a displacement field network for face recognition across pose. In *IEEE International Conference on Automatic Face Gesture Recognition (FG)*, 2017.
[18] R. Huang, S. Zhang, T. Li, and R. He. Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[19] M. Kan, S. Shan, H. Chang, and X. Chen. Stacked progressive auto-encoders (spae) for face recognition across poses. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[20] M. Kan, S. Shan, H. Zhang, S. Lao, and X. Chen. Multi-view discriminant analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2016.

[21] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, and E. Brossard. The megaface benchmark: 1 million faces for recognition at scale. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[22] A. Li, S. Shan, X. Chen, and W. Gao. Maximizing intra-individual correlations for face recognition across pose differences. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[23] S. Li, X. Liu, X. Chai, H. Zhang, S. Lao, and S. Shan. Morphable displacement field based image matching for face recognition across pose. In *European Conference on Computer Vision (ECCV)*, 2012.

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

[25] I. Masi, F. Chang, J. Choi, S. Harel, J. Kim, K. Kim, J. Leksut, S. Rawls, Y. Wu, T. Hassner, W. AbdAlmageed, G. Medioni, L. Morency, P. Natarajan, and R. Nevatia. Learning pose-aware models for pose-invariant face recognition in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2019.

[26] I. Masi, S. Rawls, G. Medioni, and P. Natarajan. Pose-aware face recognition in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[27] H.-W. Ng and S. Winkler. A data-driven approach to cleaning large face datasets. In *IEEE International Conference on Image Processing (ICIP)*, 2014.

[28] B.-S. Oh, K.-A. Toh, A. B. J. Teoh, and Z. Lin. An analytic gabor feedforward network for single-sample and pose-invariant face recognition. *IEEE Transactions on Image Processing (TIP)*, 2018.

[29] X. Peng, X. Yu, K. Sohn, D. N. Metaxas, and M. Chandraker. Reconstruction-based disentanglement for pose-invariant face recognition. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[30] U. Prabhu, J. Heo, and M. Savvides. Unconstrained pose-invariant face recognition using 3d generic elastic models. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2011.

[31] J. Rupnik and J. Shawe-Taylor. Multi-view canonical correlation analysis. In *Slovenian KDD Conference on Data Mining and Data Warehouses (SiKDD)*, 2010.

[32] A. Sharma, M. A. Haj, J. Choi, L. S. Davis, and D. W. Jacobs. Robust pose invariant face recognition using coupled latent space discriminant analysis. *Computer Vision and Image Understanding (CVIU)*, 2012.

[33] A. Sharma, A. Kumar, H. Daume, and D. W. Jacobs. Generalized multiview analysis: A discriminative latent space. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[34] K. Simonyan, O. M. Parkhi, A. Vedaldi, and A. Zisserman. Fisher vector faces in the wild. In *British Machine Vision Conference (BMVC)*, 2013.

[35] L. Q. Tran, X. Yin, and X. Liu. Representation learning by rotating your faces. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2018.

[36] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu. Cosface: Large margin cosine loss for deep face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[37] C. Xiong, X. Zhao, D. Tang, K. Jayashree, S. Yan, and T.-K. Kim. Conditional convolutional neural network for modality-aware face recognition. In *IEEE International Conference on Computer Vision (ICCV)*, 2015.

[38] J. Yim, H. Jung, B. Yoo, C. Choi, D. Park, and J. Kim. Rotating your face using multi-task deep neural network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[39] X. Yin and X. Liu. Multi-task convolutional neural network for pose-invariant face recognition. *IEEE Transactions on Image Processing (TIP)*, 2018.

[40] X. Yin, X. Yu, K. Sohn, X. Liu, and M. Chandraker. Towards large-pose face frontalization in the wild. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[41] Y. Zhang, M. Shao, E. K. Wong, and Y. Fu. Random faces guided sparse many-to-one encoder for pose-invariant face recognition. In *IEEE International Conference on Computer Vision (ICCV)*, 2013.

[42] J. Zhao, Y. Cheng, Y. Xu, L. Xiong, J. Li, F. Zhao, K. Jayashree, S. Pranata, S. Shen, J. Xing, S. Yan, and J. Feng. Towards pose invariant face recognition in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[43] J. Zhao, L. Xiong, P. K. Jayashree, J. Li, F. Zhao, Z. Wang, P. S. Pranata, P. S. Shen, S. Yan, and J. Feng. Dual-agent gans for photorealistic and identity preserving profile face synthesis. In *Advances in Neural Information Processing Systems (NIPS)*. 2017.

[44] X. Zhu, Z. Lei, J. Yan, D. Yi, and S. Z. Li. High-fidelity pose and expression normalization for face recognition in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[45] Z. Zhu, P. Luo, X. Wang, and X. Tang. Deep learning identity-preserving face space. In *IEEE International Conference on Computer Vision (ICCV)*, 2013.