Pattern Recognition 100 (2020) 107113

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Deformable face net for pose invariant face recognition

Mingjie He[a,b], Jie Zhang[a], Shiguang Shan[a,b,c,*], Meina Kan[a], Xilin Chen[a,b]

[a] Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing, 100190, China
[b] University of Chinese Academy of Sciences, Beijing 100049, China
[c] Peng Cheng Laboratory, Shenzhen, 518055, China

ARTICLE INFO

Article history:
Received 23 April 2019
Revised 9 September 2019
Accepted 15 November 2019
Available online 25 November 2019

Keywords:
Pose-invariant face recognition
Displacement consistency loss
Pose-triplet loss

ABSTRACT

Unconstrained face recognition still remains a challenging task due to various factors such as pose, expression, illumination, partial occlusion, etc. In particular, the most significant appearance variations are stemmed from poses which leads to severe performance degeneration. In this paper, we propose a novel Deformable Face Net (DFN) to handle the pose variations for face recognition. The deformable convolution module attempts to simultaneously learn face recognition oriented alignment and identity-preserving feature extraction. The displacement consistency loss (DCL) is proposed as a regularization term to enforce the learnt displacement fields for aligning faces to be locally consistent both in the orientation and amplitude since faces possess strong structure. Moreover, the identity consistency loss (ICL) and the pose-triplet loss (PTL) are designed to minimize the intra-class feature variation caused by different poses and maximize the inter-class feature distance under the same poses. The proposed DFN can effectively handle pose invariant face recognition (PIFR). Extensive experiments show that the proposed DFN outperforms the state-of-the-art methods, especially on the datasets with large poses.

1. Introduction

Face recognition, as a fundamental problem 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. However, under the uncontrolled environment, a number of factors (e.g., pose, illumination, resolution, occlusion, and expression) significantly affect the performance of face recognition system. Among these factors, self-occlusion from out-plane poses brings about large appearance variations. The misalignment problem heavily hurts the face recognition system. In this paper, we further push the frontier of this research area by simultaneously considering face recognition oriented alignment and identity-preserving feature extraction under deep neural networks, which aims at tackling the pose-invariant face recognition (PIFR) problem.

The conventional deep face recognition system usually firstly aligns faces with simple 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 variations, the intra-class appearance variations from out-plane poses still exists, resulting in face misalignment problem. As a consequence, the face recognition accuracy degenerates severely under large out-plane pose variations. To handle this problem, one can either align the face images with extra technology, e.g., 3D based face alignment [1] or improve the CNN's capacity of extracting pose-invariant features. Since human heads are nearly rigid 3D objects, following the former pipeline, many efforts are devoted to synthesizing well-aligned frontal face image from non-frontal faces by using 3D rigid motion models [2–8]. However, 3D model reconstruction with a single 2D image is an ill-conditioned problem and the synthesized image needs high fidelity refinement to improve the reality of faces. Since face recognition system extracts high-level feature to recognize identities, it is unnecessary to generated frontal faces. Thus aligning high level features is more convenient than aligning faces in pixel-level, leading to potentially more effective recognition results. The approaches following the latter pipeline focus on learning pose-invariant feature representations. Conventional approaches such as multiview subspace learning or pose-directed multi-task leaning significantly improve the large pose face recognition. Unfortunately, such subspace projections and multi-tasks are learnt corresponding to several discrete poses, it is difficult for those methods to handle face recognition under continuous pose variations. Moreover, it may be non-trivial for those methods to obtain pose-invariant feature robust to complex scenarios in no consideration of face alignment.

* Corresponding author.
E-mail address: sgshan@ict.ac.cn (S. Shan).

https://doi.org/10.1016/j.patcog.2019.107113
0031-3203/

In this paper, we propose a feature-level alignment method to handle pose variations in face recognition. In our approach, a convolution network, namely deformable face net (DFN) is designed to simultaneously learn feature-level alignment and feature extraction for face recognition. It is more favorable for CNNs to learn identity-relevant features after aligning faces, leading to better performance for face recognition under poses. Inspired by the deformable convolution [9], we propose to achieve feature-level alignment by a deformable convolution module which enables pose-aware spatial sampling based on displacement fields for the subsequent feature extraction. It should be noted that the conventional deformable convolution [9] is developed for detecting general objects which have diverse local and global non-rigid transformations, while human faces are approximately rigid and the most salient transformations are caused by the rigid pose change rather than other flexible variations. The difference in rigidness implies that the displacement field learnt for face recognition should be more consistent. With this in mind, we propose the displacement consistency loss (DCL) to enforce the local consistency of the learnt displacement field both in orientation and amplitude, leading to better alignment for face recognition. Moreover, the identity consistency loss (ICL) and the pose-triplet loss (PTL) are designed to minimize the intra-class feature variation caused by different poses and maximize the inter-class feature distance under the same poses. Specifically, the ICL minimizes the intra-class feature variation caused by different poses via taking two faces under different poses as input. The PTL emphasizes on improving the network discriminative ability of distinguishing faces with the same pose but from different identities. Besides, the DFN is quite efficient and can be end-to-end trained without additional supervision. Compared to the existing pose-invariant feature extraction methods, e.g., the PIM [10] and the p-CNN [11], the proposed DFN achieves better results for face recognition under poses, especially on the datasets with large poses.

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.
- The displacement consistency loss (DCL) is proposed to enforce the learnt displacement field to be locally consistent both in the orientation and amplitude, leading to better alignment for face recognition.
- The identity consistency loss (ICL) and the pose-triplet loss (PTL) are designed to minimize the intra-class feature variation caused by different poses and maximize the inter-class feature distance under the similar poses, leading to better performance for face recognition.
- DFN outperforms the state-of-the-art methods on MegaFace, MultiPIE and CFP, especially on the MultiPIE dataset with large poses.

The preliminary version of this work appears in [12]. We extend it in a number of ways. (i) We propose a new loss function named pose-triplets loss (PTL) for the Deformable Face Net (DFN). This new loss function improves the DFN's discriminative ability of distinguishing faces with the same pose but from different identities, leading to better results than the original DFN. (ii) Our pose-triplet loss (PTL) is evaluated together with our displacement consistency loss (DCL) on MultiPIE dataset and it significantly outperforms the state-of-the-art methods. (iii) Further experiments on the Celebrities in Frontal-Profile (CFP) dataset are conducted to demonstrate the superiority of our DFN in a wild setting.

The remainder of the paper is organized as follows: The related works are briefly reviewed in Section 2. In Section 3, the proposed DFN and loss functions are illustrated. Experimental results are detailed in Section 4. The conclusions are summarized in Section 5.

2. Related works

Recently, many efforts are devoted to exploring pose invariant face recognition (PIFR) methods, which can be roughly grouped into the following three categories: face frontalization methods, non-frontal face augmentation methods and pose-invariant feature learning methods. In this section, we give a brief review of the recent works which are most relevant to this paper.

2.1. Face frontalization methods

The face frontalization methods are essentially picture-level aligning method. The key point of nearly all these methods is how to construct a well-aligned 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 [2–8] or 2D images [13–16]. For the first category, [2] proposes an effective face frontalization approach by using a single and unchanged 3D shape to approximate the shape of all the input faces. In [3], 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 [15–17]. By modeling the face rotation process, DR-GAN [16] 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. However, 3D model reconstruction with a single 2D image is an ill-conditioned problem, so that the gap between the real 3D shape and the reconstructed 3D shape always exists. Furthermore, since the original non-frontal images have 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. Although current methods have improved the illumination trends and the texture details, the quality of the geometric frontalized images is still far from avoiding degeneration of face recognition performance. On the other side, the synthetic faces of GAN based methods usually have better visual effects. However, as the pixels are not directly collected from the input image, the major concern lies in how to guarantee that the frontalized faces can well preserve the identity information.

2.2. Face augmentation

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, data augmentation methods become more practical. The works in [18–20] enrich the diversity of poses by synthesizing massive images of sufficient pose variability from a frontal face. [18] employs 3DMM to augment the training data with faces of novel viewpoints. In [20], a multi-depth generic elastic model is developed to synthesize facial images with varying poses. To some extent, these methods relieve the poses influence, but the discrepancy between distributions of the synthetic and real face images still limits the recognition performance improvements. To improve the realism of synthetic training images, [19] proposes a dual-agent generative adversarial network (DA-GAN) to

refine the profile face images generated by the 3D face model. The compelling perceptual results improve the recognition significantly.

Although aforementioned augmentation methods enrich the training set with promising synthetic quality, but the misalignment issue inherited from the large poses still remains. The enriched training set relieves such issue by enforcing the feature extraction network to adapt to various poses. However, the final performance heavily relies on the fitting capacity of CNN, which may lead to increased computation cost.

2.3. 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 [21–27]. For instance, Sharma et al. [25] presents a discriminant coupled latent subspace framework for pose-invariant discriminative learning. In [26], GMA extracts unified multiview features by optimizing view-specific projections. In [27], MvDA is proposed to jointly solve the multiple linear transforms and meanwhile minimizes the within-class variations, resulting in very encouraging performance.

Recently, more works resort to the deep learning to extract more powerful pose-invariant features [10,11,28–34]. 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 [11,29,30,34] or design a single pose-invariant model [31–33], which uniformly tackles all poses. For the former category, [30] proposes a pose-directed multi-task CNN to learn pose-specific identity features. Similarly, in [34], 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 [33] to absorb moderate changes caused by poses. In [10], 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 face synthesis in PIM makes it essentially a pixel-level alignment method. In contrast, our method explicitly considers feature-level alignments. Furthermore, comparing to subspace methods and multi-tasks methods, our method can tackle arbitrary poses rather than several specific poses.

3. Method

The proposed Deformable Face Net (DFN) attempts to simultaneously learn feature-level alignment and feature extraction for face recognition via deformable convolutions with a spatial displacement field. This field is adaptively pose-aware, thus endowing the deformable convolution the ability to align features in case of pose variations. For this purpose, these displacement fields are learnt by introducing three loss functions, i.e., the displacement consistency loss (DCL), the identity consistency loss (ICL) and the pose-triplet loss (PTL). In this way, the DFN is able to well tackle the feature misalignment issue caused by poses, resulting in performance improvement in face recognition.

3.1. Overview of DFN

As shown in Fig. 1, a displacement field generator learns displacement fields at low-level features for face recognition oriented alignment. In consideration of the strong structure in faces, the displacement consistency loss (DCL) is proposed to improve the local consistency of the displacement fields and therefore assists the deformable convolution to well tackle the PIFR problem. Moreover, the identity consistency loss (ICL) are proposed to minimize the intra-class feature variation caused by different poses, so as to explicitly force the learnt displacement fields to well align features under different poses. When employing the ICL, the 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, thus providing compatibility with various normal training datasets. When extra pose information of training set is available, the proposed pose-triplet loss (PTL) can jointly minimize the intra-class feature variation and further maximize the inter-class feature distance under the same poses, so as the extracted features become more robust to poses. Both the ICL and the PTL losses are imposed on intermediate feature (i.e., the output feature of the deformable convolution) to supervise the learning of the displacement field generator, so that displacement fields are able to achieve the pose-aware feature alignment. The whole network is end-to-end trained jointly by using the softmax classification loss and the proposed loss functions recorded as DCL, ICL and PTL. The proposed method can be integrated with the existing powerful CNN architectures, e.g., the ResNet architecture [35,36]. We note that introducing the pose-aware deformation modules at different layers of the network have significant differences in performance. Details will be discussed in Section 4. Next, we present each component of the DFN in details.

3.2. Displacement consistency loss

Given an input feature map x, the kernels of the deformable convolution [9] samples irregular grids over the input x. For each gird i centered on location $\mathbf{p}_0^i$, such irregular sampling locations are obtained by an addition of offsets $\{\Delta\mathbf{p}_k^i = \{\Delta p_{kx}^i, \Delta p_{ky}^i\} | k = 1, \ldots, K\}$ (i.e., a displacement field) to a regular sampling grid $\mathcal{R}$. Δp_{kx}^i and Δp_{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 3×3 convolution kernels. Then, the output feature map f of the deformable convolution is computed as below:

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

where $\mathcal{R} = \{(-1,-1), (-1,0), \ldots, (0,1), (1,1)\}$ for a 3×3 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 salient 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

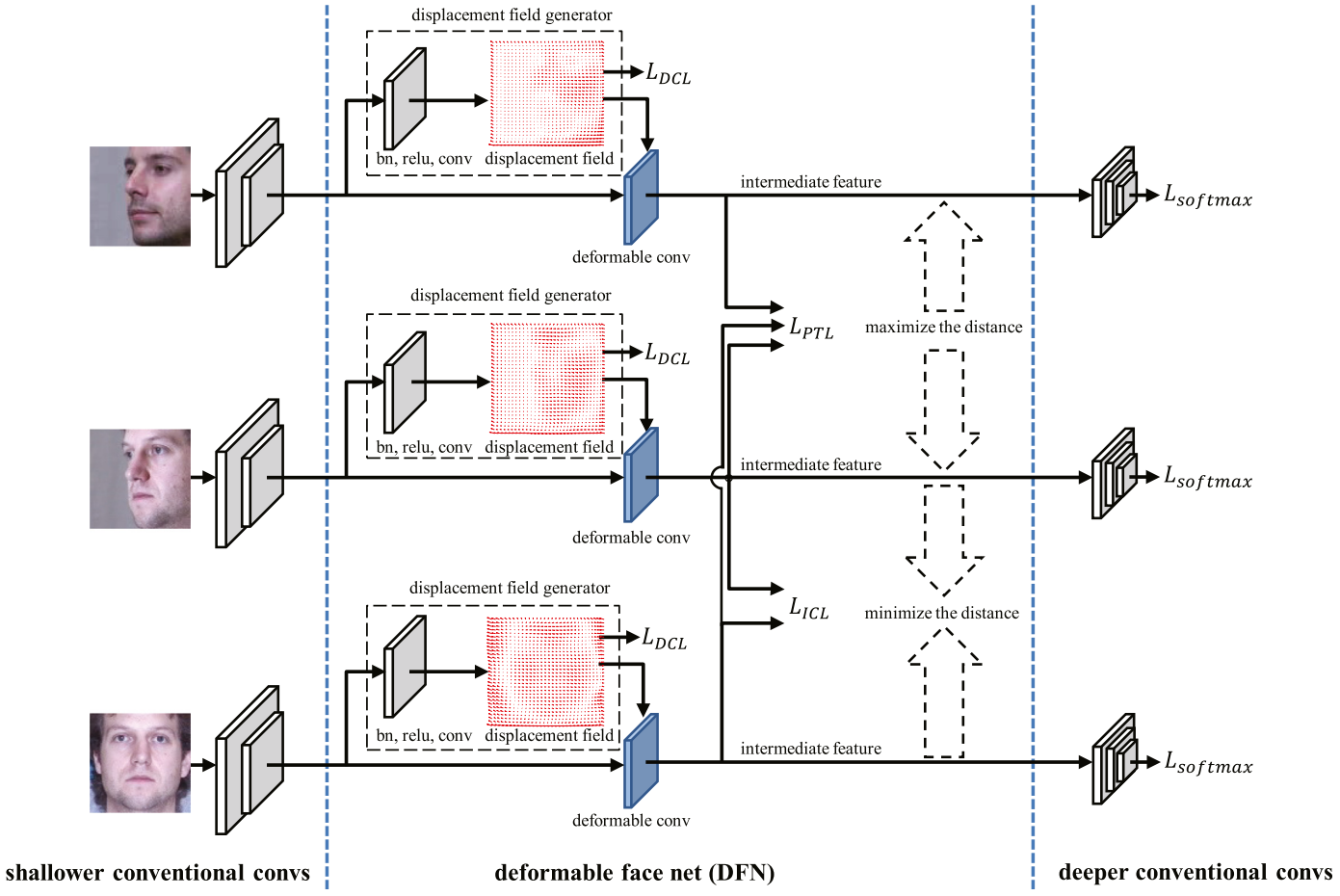


Fig. 1. Illustration of our proposed Deformable Face Net (DFN). DFN attempts to learn a pose-aware displacement field for the deformable convolution to extract pose-invariant features for face recognition. This field is adaptively pose-aware, thus endowing the deformable convolution the ability to align features in case of pose variations. For this purpose, these displacement fields are learnt by introducing three loss functions, i.e., the displacement consistency loss (DCL), the identity consistency loss (ICL) and the pose-triplet loss (PTL).

direction, as shown in Fig. 2. The DCL is formulated in Eq. (2) as:

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

where $\Delta \bar{\mathbf{p}}^i$ 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.

3.3. 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 [37,38], 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, benefited from the pose-aware deformation modules, DFN can naturally handle this problem more efficiently. 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.

3.4. Pose-triplets loss

The pose variation reduces the similarity of faces from the same identity. In addition, it even surpasses the intrinsic appearance differences between individuals, i.e., the features extracted from different identities under the same poses are more similar than those from the same identity across different poses. In this paper, we reformulate the triplet loss [39] as pose-triplets loss

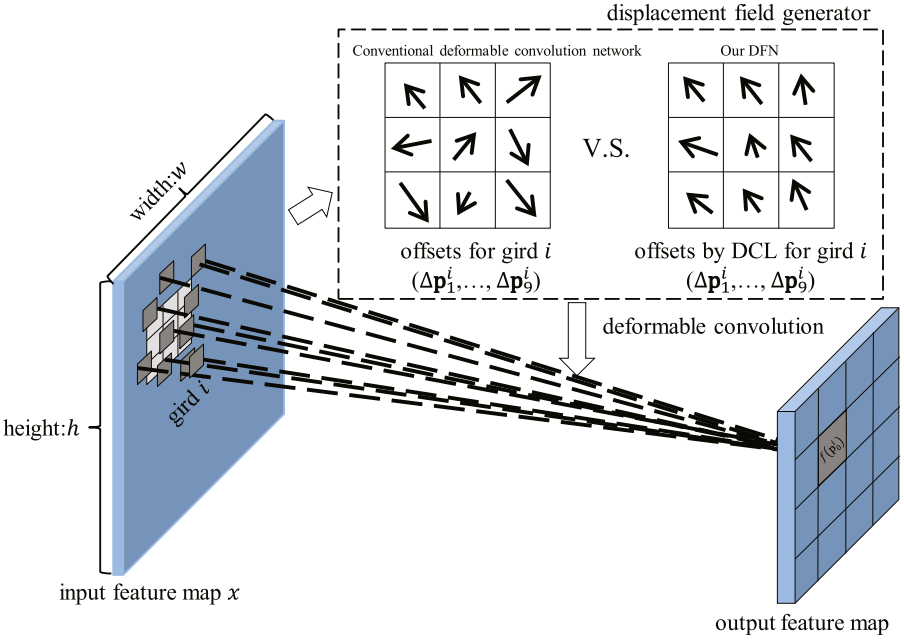


Fig. 2. Illustration of the offsets obtained with our displacement consistency loss (DCL).

(PTL) to improve the discriminative ability of separating images with same poses but from different identities.

Formally, $\mathbf{f}_i^a$ denotes the feature of the anchor face and $\mathbf{f}_i^p$ denotes the feature of positive sample from the same identity. The negative image is chosen from any other identity which has the same pose with the anchor face. Here, we want to ensure that the feature distance of the negative pair (recorded as $\mathbf{f}_i^a$ and $\mathbf{f}_i^n$) is larger than the distance of the positive pair (recorded as $\mathbf{f}_i^a$ and $\mathbf{f}_i^p$). The pose-triplets loss aims to separate the positive pair from the negative by a distance margin α. The PTL is formulated in Eq. (4) as:

$$L_{PTL} = \sum_{i=1}^{N}[\|\mathbf{f}_i^a - \mathbf{f}_i^p\|_2^2 - \|\mathbf{f}_i^a - \mathbf{f}_i^n\|_2^2 + \alpha]_+. \tag{4}$$

Additionally, similar to the aforementioned ICL, the features $\mathbf{f}_i^a$, $\mathbf{f}_i^p$ and $\mathbf{f}_i^n$ are normalized for better convergence. The Algorithm 1 summarizes the workflow of training our DFN with the proposed loss functions.

3.5. Discussion

3.5.1. Differences with the deformable convolution network

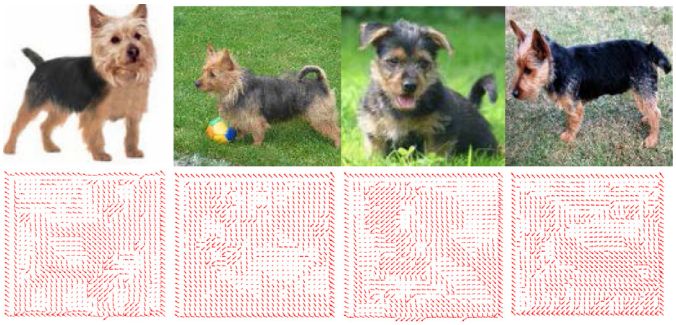
Both the deformable convolution network [9] and our DFN are feature-level alignment methods that attempt to handle the geometric transformations. The deformable convolution is firstly developed for detecting general objects which have diverse local and global non-rigid transformations, e.g., the dogs shown in Fig. 3 have significantly different postures. In contrast, human faces are approximately rigid objects and the most salient 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, three addition loss functions DCL, ICL and PTL are embedded in DFN 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. Moreover, when both the deformable convolution network and our DFN are applied to the human faces, the displacement fields of our DFN notably shows more structure consistency than those of the deformable convolution network, leading to better face alignment and further improved face recognition performance. The significant improvements in face recognition further demonstrate the effectiveness of our DFN, see details in Section 4.3.

Algorithm 1: Training deformable face net.

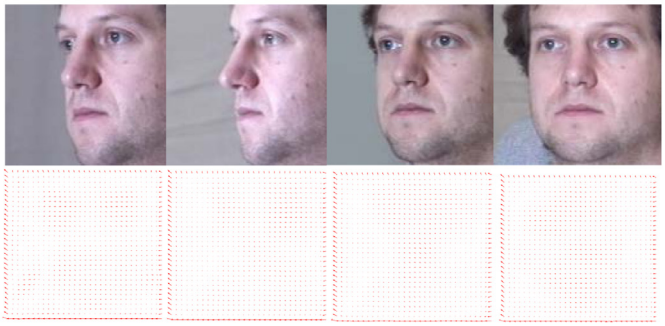
Input: A training batch containing N images and their labels.
while *not converged* **do**
  Compute the input feature map x for the deformable convolution;
  Compute the displacement field $\{\Delta\mathbf{p}_k^i | k = 1,\ldots,K\}$;
  Compute the displacement consistency loss L_{DCL};
  Compute the output feature map f of the deformable convolution;
  if *training set contains pose information* **then**
    Compute the pose-triplets loss L_{PTL};
    Compute the softmax loss $L_{softmax}$;
    Compute the total loss L_{total}:
      $L_{total} = L_{softmax} + \alpha L_{DCL} + \beta L_{PTL}$;
  else
    Compute the identity consistency loss L_{ICL};
    Compute the softmax loss $L_{softmax}$;
    Compute the total loss L_{total}:
      $L_{total} = L_{softmax} + \alpha L_{DCL} + \beta L_{ICL}$;
  end
  Backpropagation and update the weights of the DFN
end
Output: The trained DFN.

3.5.2. Differences with the face frontalization methods

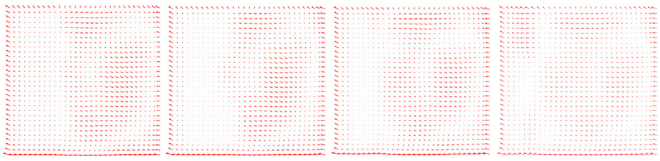
The face frontalization methods [2–8,13–17,19] which are image-level alignment attempt to generate frontal faces, while our DFN is feature-level alignment that attempts to align features under different poses. For face recognition, the generated frontal faces are further fed into CNNs for feature extraction, resulting in a two-stage process (i.e., the face frontalizaition and



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



(b) Displacement fields of the deformable convolution network for faces.



(c) 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. Since human heads are nearly rigid objects, the deformable transformations require more consistency. However, as show in (b), when directly applying conventional deformable convolution network for human faces, the generated displacement fields lack sufficient consistency, which are not good enough for aligning faces across poses. In contrast, as shown in (c), the displacement fields of our DFN are more consistent, which demonstrates the effectiveness of the proposed method.

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.

3.5.3. Differences with other pose-invariant feature learning methods

Different from most pose-invariant feature leaning methods [10,11,23–34] using multiple models in which each model correspond to a specific pose, our DFN presents a unified model to handle different poses. Besides, those subspace learning approaches [23–27] 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.

4. Experiments

4.1. Experimental setting

4.1.1. Dataset

To investigate the effectiveness of the proposed DFN, we evaluate our method on three main face recognition benchmarks, MegaFace [40], MultiPIE [41] and CFP [42]. The MegaFace [40] 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 $\pm$ 40 degrees. In this study, we evaluate the performance of our approach on the standard MegaFace challenge 1 (MF1) benchmark. This benchmark evaluates how face recognition method performs with a very large number of distractors in the gallery. For this purpose, the subjects in the MegaFace dataset [40] are used as the distractors, while the probes are from the Facescrub dataset [43]. The MegaFace dataset consists of more than 1 million face images from 690k different individuals and the Facescrub dataset contains 106,863 face images of 530 subjects. Specifically, in one test, each of the images per subject in the Facescrub dataset is added into the gallery, and each of the remaining images in the Facescrub of this subject is exploited as a probe. 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 [30,31], 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°. 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.

To evaluate how our DFN performs in a wild setting, we conduct experiments on the Celebrities in Frontal-Profile (CFP) database [42]. The CFP contains 7000 images of subjects and each subject has 10 frontal and 4 profile face images. The images in CFP are organized into 10 splits and each split contains 350 frontal-frontal pairs and 350 frontal-profile pairs. The evaluation follows the 10 fold cross-validation protocol defined in [42] and the mean and standard deviation of accuracy(ACC), Equal Error Rate (EER) and Area Under Curve (AUC) are used as the measurement.

4.1.2. Implementation details

In our experiments, we use [44] 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. The DFN (DCL&PTL) denotes the version trained with the DCL loss and

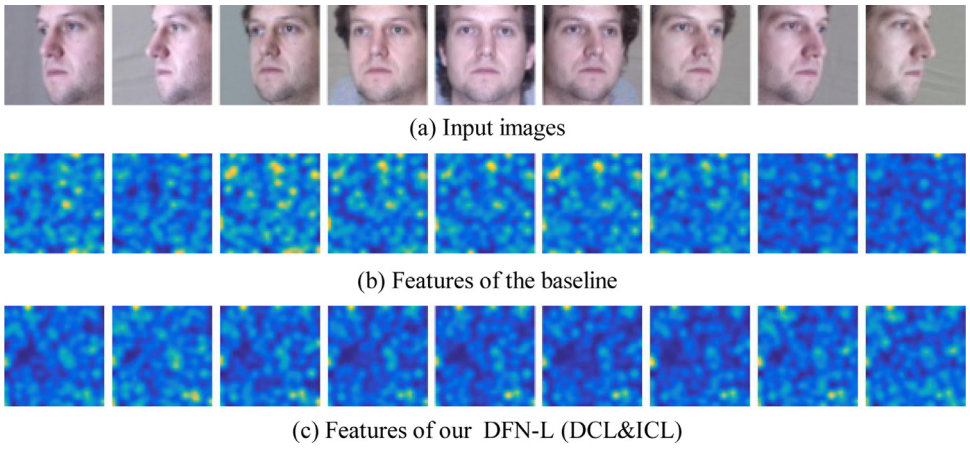


Fig. 4. An example of pose-invariant features of DFN-L (DCL&ICL) with various poses (-60° to $+60^{\circ}$). Even with the same identity, obvious differences are witnessed between features extracted from the baseline method. In contrast, the features obtained by the proposed DFN-L (DCL&ICL) show a similar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

Table 1
Architecture details of DFN-ResNet-50 and DFN-ResNet-152 with DCL, PTL or ICL.

Output size	DFN-ResNet-50	DFN-ResNet-152
62×62	conv, 7×7, 64, stride 4	conv, 7×7, 64, stride 2 *displacement field generator (conv, 3 × 3, 18)* *DCL loss* *deformable conv, 3 × 3, 64* *PTL loss or ICL loss* max pool, 3×3, stride 2
stage 1 31×31	max pool, 3×3, stride 2 $\begin{bmatrix} \text{conv}, 1 \times 1, 64 \\ \text{conv}, 3 \times 3, 64 \\ \text{conv}, 1 \times 1, 256 \end{bmatrix} \times 3$	$\begin{bmatrix} \text{conv}, 1 \times 1, 64 \\ \text{conv}, 3 \times 3, 64 \\ \text{conv}, 1 \times 1, 256 \end{bmatrix} \times 3$
stage 2 16×16	*displacement field generator (conv, 3 × 3, 18)* *DCL loss* *deformable conv, 3 × 3, 64* *PTL loss or ICL loss* $\begin{bmatrix} \text{conv}, 1 \times 1, 128 \\ \text{conv}, 3 \times 3, 128 \\ \text{conv}, 1 \times 1, 512 \end{bmatrix} \times 4$	$\begin{bmatrix} \text{conv}, 1 \times 1, 128 \\ \text{conv}, 3 \times 3, 128 \\ \text{conv}, 1 \times 1, 512 \end{bmatrix} \times 8$
stage 3 8×8	$\begin{bmatrix} \text{conv}, 1 \times 1, 256 \\ \text{conv}, 3 \times 3, 256 \\ \text{conv}, 1 \times 1, 1024 \end{bmatrix} \times 6$	$\begin{bmatrix} \text{conv}, 1 \times 1, 256 \\ \text{conv}, 3 \times 3, 256 \\ \text{conv}, 1 \times 1, 1024 \end{bmatrix} \times 36$
stage 4 1×1024	$\begin{bmatrix} \text{conv}, 1 \times 1, 512 \\ \text{conv}, 3 \times 3, 512 \\ \text{conv}, 1 \times 1, 2048 \end{bmatrix} \times 3$ avg pool, 4×4 fc, 1024	$\begin{bmatrix} \text{conv}, 1 \times 1, 512 \\ \text{conv}, 3 \times 3, 512 \\ \text{conv}, 1 \times 1, 2048 \end{bmatrix} \times 3$ avg pool, 7×7 fc, 1024
1×1	fc, softmax loss	

PTL loss jointly. For the MegaFace evaluation, the conventional ResNet-50 and ResNet-152 are used as our baselines.

We manually clean the MS-Celeb-1M [45] 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 three losses. The architecture details of the DFN-ResNet-50 and DFN-ResNet-152 are summarized in Table 1. For experiments on CFP dataset, we modify the conventional ResNet-18, forming a lightweight ResNet-10 as our baseline. The DFN-10 is then constructed by integrating the deformable module with ResNet-10. The architecture details of the DFN-Light and DFN-ResNet-10 are summarized in Table 2. Both the baseline and DFN-10 are pre-trained on the cleaned MS-Celeb-1M dataset and then fine-tuned on the CFP dataset following the 10 folds cross-validation protocol [42]. We implement our method on the MXNet [46] 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, PTL and ICL are 0.001, 0.1 and 0.01 respectively.

4.2. 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 the DFN-ResNet-50 (DFN-50 for short in the following sections). Table 1 exhibits an example of integrating the deformable module in the stage 2. One significant difference

Table 2
Architecture details of DFN-Light and DFN-ResNet-10 with DCL, PTL or ICL.

Output size	DFN-Light	DFN-ResNet-10
62 × 62	conv,7 × 7, 64, stride 4	conv, 7 × 7, 64, stride 4
stage 1 31 × 31	max pool, 3 × 3, stride 2 *displacement field generator (conv, 3 × 3, 18)* *DCL loss* *deformable conv, 3 × 3, 64* *PTL loss or ICL loss*	max pool, 3 × 3, stride 2 $\begin{bmatrix}\text{conv}, 3\times3, 256\\ \text{conv}, 3\times3, 256\end{bmatrix}\times 1$
stage 2 16 × 16	conv, 3 × 3, 64, stride 2 conv, 3 × 3, 64, stride 1	*displacement field generator (conv, 3 × 3, 18)* *DCL loss* *deformable conv, 3 × 3, 64* *PTL loss or ICL loss* $\begin{bmatrix}\text{conv}, 3\times3, 512\\ \text{conv}, 3\times3, 512\end{bmatrix}\times 1$
stage 3 8 × 8	conv, 3 × 3, 128, stride 2	$\begin{bmatrix}\text{conv}, 3\times3, 1024\\ \text{conv}, 3\times3, 1024\end{bmatrix}\times 1$
stage 4 1 × 1024	conv, 3 × 3, 128, stride 1 avg pool, 4 × 4 fc, 1024	$\begin{bmatrix}\text{conv}, 3\times3, 2048\\ \text{conv}, 3\times3, 2048\end{bmatrix}\times 1$ avg pool, 4 × 4 fc, 1024
1 × 1	fc, softmax loss	

Table 3
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 stage 1	75.25
DFN-50 with deformable conv embedded in stage 2	75.02
DFN-50 with deformable conv embedded in stage 3	72.48
DFN-50 with deformable conv embedded in stage 4	58.88

Table 4
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

between these four versions is the 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 3, 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 3 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 with 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 4 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 seen, 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. Nevertheless 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 is further improved to 78.21% which outperforms the plain ResNet-50 by 3.45%.

Table 5
Rank-1 identification accuracy on MegaFace challenge 1 compared to the state-of-the-art methods.

Method	MF1 Rank1
SphereFace-Small [47]	75.76
CosFace [48]	82.72
ArcFace [49]	81.03
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 5 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 result comparable to that of CosFace [48] trained with 90K identities and that of ArcFace [49] trained with 85K 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.

Table 6
Rank-1 recognition rates (%) on MultiPIE for different poses.

Method	±90°	±75°	±60°	±45°	±30°	±15°
CPF [50]	–	–	–	71.65	81.05	89.45
Hassner [2]	–	–	44.81	74.68	89.59	96.78
FV [51]	24.53	45.51	68.71	80.33	87.21	93.30
HPN [32]	29.82	47.57	61.24	72.77	78.26	84.23
FIP [31]	31.37	49.10	69.75	85.54	92.98	96.30
c-CNN [30]	47.26	60.66	74.38	89.02	94.05	96.97
TP-GAN [15]	64.03	84.10	92.93	98.58	99.85	99.78
PIM [10]	75.00	91.20	97.70	98.30	99.40	99.80
p-CNN [11]	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
DFN-L(DCL&PTL)	85.66	90.04	96.13	98.40	99.22	99.52

Table 7
Face recognition performance (%) comparison on CFP dataset. The results are average ± standard deviation over the 10 folds.

Method	Frontal-Profile		
	ACC	EER	AUC
Sengupta et al. [42]	84.91 ± 1.82	14.97 ± 1.98	93.00 ± 1.55
Sankarana et al. [52]	89.17 ± 2.35	8.85 ± 0.99	97.00 ± 0.53
Chen et al. [53]	91.97 ± 1.70	8.00 ± 1.68	97.70 ± 0.82
DR-GAN [16]	93.41 ± 1.17	6.45 ± 0.16	97.96 ± 0.06
PIM [10]	93.10 ± 1.01	7.69 ± 1.29	97.65 ± 0.62
Peng, et al. [18]	93.76	–	–
Human	94.57 ± 1.10	5.02 ± 1.07	98.92 ± 0.46
ResNet-10	92.89 ± 1.42	6.69 ± 1.43	97.90 ± 0.58
DFN-10	92.72 ± 1.57	7.03 ± 1.14	97.87 ± 0.50
DFN-10(ICL)	93.89 ± 2.25	5.71 ± 1.87	98.06 ± 0.89
DFN-10(DCL)	93.64 ± 2.39	5.69 ± 1.94	98.09 ± 0.87
DFN-10(ICL&DCL)	93.99 ± 2.75	5.51 ± 1.92	98.18 ± 0.98
DFN-10(PTL&DCL)	94.01 ± 2.79	5.40 ± 2.03	98.24 ± 1.02

4.3. Evaluations on the multipie benchmark

Table 6 summarizes the face recognition accuracy of our DFN-Light (DFN-L for short) on MultiPIE for different poses. The results of other state-of-the-arts are directly quoted from [2,10,11,15,30–32,50,51]. As seen from Table 6, the face frontalization method Hassner [2] performs better than CPF [50] since 3D facial shapes are utilized for the face synthesizing. Furthermore, benefitting from the patch based reconstruction and occlusion detection, HPN [32] achieves better results than [50] and [2]. Attributed to the powerful generation ability of GAN, the TP-GAN [15] outperforms all previous face frontalization methods. Differently, the methods of FV [51], FIP [31], c-CNN [30], p-CNN [11] and PIM [10] focus on pose-invariant feature learning. Among them, the deep methods FIP [31], c-CNN [30] and p-CNN [11] outperform the traditional feature representations method FV [51]. 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 [10] achieves almost the best results among the existing methods except the ± 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°.

As seen, our DFN-L generally 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 [11] with an improvement up to 7.11% for ± 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&PTL) achieves the comparable results with PIM and significantly outperforms PIM with an improvement up to 10.66% under faces of ± 90°. It is worth noting that the DFN-L has a very light network structure (as shown in Table 2), which is much more efficient than the GAN based PIM.

4.4. Evaluations on the CFP benchmark

Table 7 summarizes the Accuracy(ACC), Equal Error Rate (EER) and Area Under the Curve (AUC) on CFP dataset. The results of the other state-of-the-arts are directly quoted from [10,16,18,42,52,53]. As seen from Table 7, our DFN-10 (PTL&DCL) outperforms Peng, et al. [18], DR-GAN [16] and PIM [10], reaching a higher accuracy of 94.01%. Besides, attributed to the joint leaning with the proposed DCL and PTL loss functions, our DFN-L (PTL&DCL) achieves lower EER results than PIM [11] with an EER reduction up to 2%. It is worth noting that, without the proposed loss functions, DFN-10 performs worse than the baseline ResNet-10. The reason is that without the propose loss functions, it is non-trivial for the deformable module to learn appropriate pose-aware displacement fields for well face alignment. Moreover, it also increases the risk of over-fitting potentially. To be free from this, the experiments have illustrated that it is necessary to use the proposed loss functions with the deformable module jointly. For instance, with the DCL loss, the accuracy of DFN-10 (DCL) is improved to 93.64% which further demonstrates the effectiveness of enforcing the learnt displacement field to be locally consistent.

5. 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, three loss functions, namely displacement consistency loss (DCL), identity consistency loss (ICL) and pose-triplet loss (PTL) 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 and maximize the inter-class feature distance under the same poses. Extensive experiments show that the proposed DFN achieves quite promising performance with relatively light network structure, especially for those large poses.

Acknowledgments

This work is partially supported by National Key R&D Program of China (no. 2017YFA0700800), Natural Science Foundation of China (nos. 61806188 and 61772496).

References

[1] Y. Taigman, M. Yang, M. Ranzato, L. Wolf, Deepface: Closing the gap to human-level performance in face verification, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1701–1708.
[2] T. Hassner, S. Harel, E. Paz, R. Enbar, Effective face frontalization in unconstrained images, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4295–4304.
[3] X. Zhu, Z. Lei, J. Yan, D. Yi, 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, pp. 787–796.

[4] A. Asthana, T.K. Marks, M.J. Jones, K.H. Tieu, M. Rohith, Fully automatic pose-invariant face recognition via 3d pose normalization, in: IEEE International Conference on Computer Vision (ICCV), 2011, pp. 937–944.
[5] U. Prabhu, J. Heo, M. Savvides, Unconstrained pose-invariant face recognition using 3d generic elastic models, IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) 33 (10) (2011) 1952–1961.
[6] S. Li, X. Liu, X. Chai, H. Zhang, S. Lao, S. Shan, Morphable displacement field based image matching for face recognition across pose, in: European Conference on Computer Vision (ECCV), 2012, pp. 102–115.
[7] C. Ding, C. Xu, D. Tao, Multi-task pose-invariant face recognition, IEEE Trans. Image Process. (TIP) 24 (3) (2015) 980–993.
[8] J. Cao, Y. Hu, H. Zhang, R. He, Z. Sun, Learning a high fidelity pose invariant model for high-resolution face frontalization, in: Advances in Neural Information Processing Systems (NIPS), 2018, pp. 2867–2877.
[9] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, Y. Wei, Deformable convolutional networks, in: IEEE International Conference on Computer Vision (ICCV), 2017, pp. 764–773.
[10] J. Zhao, Y. Cheng, Y. Xu, L. Xiong, J. Li, F. Zhao, K. Jayashree, S. Pranata, S. Shen, J. Xing, S. Yan, J. Feng, Towards pose invariant face recognition in the wild, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2207–2216.
[11] X. Yin, X. Liu, Multi-task convolutional neural network for pose-invariant face recognition, IEEE Trans. Image Process. (TIP) 27 (2) (2018) 964–975.
[12] M. He, J. Zhang, S. Shan, M. Kan, X. Chen, Deformable face net: Learning pose invariant feature with pose aware feature alignment for face recognition, in: IEEE International Conference on Automatic Face Gesture Recognition (FG), 2019, pp. 1–8.
[13] L. Hu, M. Kan, S. Shan, X. Song, 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, pp. 9–16.
[14] M. Kan, S. Shan, H. Chang, X. Chen, Stacked progressive auto-encoders (spae) for face recognition across poses, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1883–1890.
[15] R. Huang, S. Zhang, T. Li, 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, pp. 2439–2448.
[16] L. Tran, X. Yin, X. Liu, Disentangled representation learning gan for pose-invariant face recognition, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1415–1424.
[17] X. Yin, X. Yu, K. Sohn, X. Liu, M. Chandraker, Towards large-pose face frontalization in the wild, in: IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3990–3999.
[18] X. Peng, X. Yu, K. Sohn, D.N. Metaxas, M. Chandraker, Reconstruction-based disentanglement for pose-invariant face recognition, in: IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1623–1632.
[19] J. Zhao, L. Xiong, P.K. Jayashree, J. Li, F. Zhao, Z. Wang, P.S. Pranata, P.S. Shen, S. Yan, J. Feng, Dual-agent gans for photorealistic and identity preserving profile face synthesis, in: Advances in Neural Information Processing Systems (NIPS), 2017, pp. 66–76.
[20] W. Deng, J. Hu, Z. Wu, J. Guo, From one to many: pose-aware metric learning for single-sample face recognition, Pattern Recognit. 77 (2018) 426–437.
[21] H. Hotelling, Relations between two sets of variates, Biometrika 28 (3/4) (1936). 321–277
[22] J. Rupnik, J. Shawe-Taylor, Multi-view canonical correlation analysis, in: Slovenian KDD Conference on Data Mining and Data Warehouses (SiKDD), 2010, pp. 1–4.
[23] A. Li, S. Shan, X. Chen, W. Gao, Maximizing intra-individual correlations for face recognition across pose differences, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009, pp. 605–611.
[24] G. Andrew, R. Arora, J. Bilmes, K. Livescu, Deep canonical correlation analysis, in: International Conference on Machine Learning (ICML), 2013, pp. 1247–1255.
[25] A. Sharma, M.A. Haj, J. Choi, L.S. Davis, D.W. Jacobs, Robust pose invariant face recognition using coupled latent space discriminant analysis, Comput. Vision Image Underst. (CVIU) 116 (11) (2012) 1095–1110.
[26] A. Sharma, A. Kumar, H. Daume, D.W. Jacobs, Generalized multiview analysis: a discriminative latent space, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012, pp. 2160–2167.
[27] M. Kan, S. Shan, H. Zhang, S. Lao, X. Chen, Multi-view discriminant analysis, IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) 38 (1) (2016) 188–194.
[28] Y. Zhang, M. Shao, E.K. Wong, Y. Fu, Random faces guided sparse many-to-one encoder for pose-invariant face recognition, in: IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2416–2423.
[29] I. Masi, S. Rawls, G. Medioni, P. Natarajan, Pose-aware face recognition in the wild, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4838–4846.
[30] C. Xiong, X. Zhao, D. Tang, K. Jayashree, S. Yan, T.-K. Kim, Conditional convolutional neural network for modality-aware face recognition, in: IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3667–3675.
[31] Z. Zhu, P. Luo, X. Wang, X. Tang, Deep learning identity-preserving face space, in: IEEE International Conference on Computer Vision (ICCV), 2013, pp. 113–120.
[32] C. Ding, D. Tao, Pose-invariant face recognition with homography-based normalization, Pattern Recognit. 66 (2017) 144–152.
[33] B.-S. Oh, K.-A. Toh, A.B.J. Teoh, Z. Lin, An analytic gabor feedforward network for single-sample and pose-invariant face recognition, IEEE Trans. Image Process. (TIP) 27 (6) (2018) 2791–2805.
[34] 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, R. Nevatia, Learning pose-aware models for pose-invariant face recognition in the wild, IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) 41 (2) (2019) 379–393.
[35] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.
[36] K. He, X. Zhang, S. Ren, J. Sun, Identity mappings in deep residual networks, in: European Conference on Computer Vision (ECCV), 2016, pp. 630–645.
[37] R. Hadsell, S. Chopra, Y. LeCun, Dimensionality reduction by learning an invariant mapping, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2006, pp. 1735–1742.
[38] S. Chopra, R. Hadsell, Y. LeCun, Learning a similarity metric discriminatively, with application to face verification, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2005, pp. 539–546.
[39] F. Schroff, D. Kalenichenko, J. Philbin, Facenet: a unified embedding for face recognition and clustering, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 815–823.
[40] I. Kemelmacher-Shlizerman, S.M. Seitz, D. Miller, E. Brossard, The megaface benchmark: 1 million faces for recognition at scale, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4873–4882.
[41] R. Gross, I. Matthews, J. Cohn, T. Kanade, S. Baker, Multi-pie, Image Vision Comput. (IVC) 28 (5) (2010) 807–813.
[42] S. Sengupta, J.-C. Chen, C. Castillo, V.M. Patel, R. Chellappa, D.W. Jacobs, Frontal to profile face verification in the wild, in: IEEE Winter Conference on Applications of Computer Vision (WACV), 2016, pp. 1–9.
[43] H.-W. Ng, S. Winkler, A data-driven approach to cleaning large face datasets, in: IEEE International Conference on Image Processing (ICIP), 2014, pp. 343–347.
[44] Z. He, M. Kan, J. Zhang, X. Chen, S. Shan, A fully end-to-end cascaded cnn for facial landmark detection, in: IEEE International Conference on Automatic Face Gesture Recognition (FG), 2017, pp. 200–207.
[45] Y. Guo, L. Zhang, Y. Hu, X. He, J. Gao, MS-Celeb-1M: a dataset and benchmark for large scale face recognition, in: European Conference on Computer Vision (ECCV), 2016, pp. 87–102.
[46] T. Chen, M. Li, Y. Li, M. Lin, N. Wang, M. Wang, T. Xiao, B. Xu, C. Zhang, Z. Zhang, MXNet: a flexible and efficient machine learning library for heterogeneous distributed systems, Neural Information Processing Systems, Workshop on Machine Learning Systems, 2015.
[47] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, L. Song, Sphereface: deep hypersphere embedding for face recognition, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 212–220.
[48] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, W. Liu, Cosface: large margin cosine loss for deep face recognition, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5265–5274.
[49] J. Deng, J. Guo, N. Xue, S. Zafeiriou, Arcface: Additive angular margin loss for deep face recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 4690–4699.
[50] J. Yim, H. Jung, B. Yoo, C. Choi, D. Park, J. Kim, Rotating your face using multi–task deep neural network, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 676–684.
[51] K. Simonyan, O.M. Parkhi, A. Vedaldi, A. Zisserman, Fisher vector faces in the wild, in: British Machine Vision Conference (BMVC), 2, 2013, p. 4.
[52] S. Sankaranarayanan, A. Alavi, C.D. Castillo, R. Chellappa, Triplet probabilistic embedding for face verification and clustering, in: IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS), 2016, pp. 1–8.
[53] J.-C. Chen, J. Zheng, V.M. Patel, R. Chellappa, Fisher vector encoded deep convolutional features for unconstrained face verification, in: IEEE International Conference on Image Processing (ICIP), 2016, pp. 2981–2985.

Mingjie He received the M.S. degree from the University of Science and Technology of China, Hefei, China, in 2014. Currently, he is a Ph.D. candidate at the University of Chinese Academy of Sciences and an engineer with the Institute of Computing Technology, Chinese Academy of Sciences (CAS). His research interests cover computer vision and machine learning.

Jie Zhang is an assistant professor with the Institute of Computing Technology, Chinese Academy of Sciences (CAS). He received the Ph.D. degree from the University of Chinese Academy of Sciences, Beijing, China. His research interests include deep learning and its application in face alignment, face recognition, object detection and localization.

Shiguang Shan received the M.S. degree in computer science from the Harbin Institute of Technology, Harbin, China, in 1999, and Ph.D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences (CAS), Beijing, China, in 2004. Currently, he is a professor with the Institute of Computing Technology, Chinese Academy of Sciences (CAS) and the University of Chinese Academy of Sciences. His research interests cover computer vision, pattern

recognition, and machine learning. He has published more than 200 papers in refereed journals and proceedings.

Meina Kan is an associate professor with the Institute of Computing Technology, Chinese Academy of Sciences (CAS). She received the Ph.D. degree from the University of Chinese Academy of Sciences, Beijing, China. Her research mainly focuses on Computer Vision especially face recognition, transfer learning, and deep learning.

Xilin Chen received the B.S., M.S., and Ph.D. degrees in computer science from the Harbin Institute of Technology, Harbin, China, in 1988, 1991, and 1994, respectively. He is a professor with the Institute of Computing Technology, Chinese Academy of Sciences (CAS) and the University of Chinese Academy of Sciences. He has authored one book and over 200 papers in refereed journals and proceedings in the areas of computer vision, pattern recognition, image processing, and multimodal interfaces. He is a fellow of the IEEE, IAPR, and CCF.