Direct Shape Regression Networks for End-to-End Face Alignment

Xin Miao[1,*] Xiantong Zhen[2,1,*] Xianglong Liu[2] Cheng Deng[3]
Vassilis Athitsos[1] Heng Huang[4,1*]
[1]University of Texas at Arlington, TX, USA, [2]Beihang University, Beijing, China
[3]Xidian University, Xi'an, China, [4] University of Pittsburgh, PA, USA
xin.miao@mavs.uta.edu, zhenxt@gmail.com, xlliu@nlsde.edu.cn, chdeng.xd@gmail.com
athitsos@uta.edu, heng.huang@pitt.edu

Abstract

Face alignment has been extensively studied in computer vision community due to its fundamental role in facial analysis, but it remains an unsolved problem. The major challenges lie in the highly nonlinear relationship between face images and associated facial shapes, which is coupled by underlying correlation of landmarks. Existing methods mainly rely on cascaded regression, suffering from intrinsic shortcomings, e.g., strong dependency on initialization and failure to exploit landmark correlations. In this paper, we propose the direct shape regression network (DSRN) for end-to-end face alignment by jointly handling the aforementioned challenges in a unified framework. Specifically, by deploying doubly convolutional layer and by using the Fourier feature pooling layer proposed in this paper, DSRN efficiently constructs strong representations to disentangle highly nonlinear relationships between images and shapes; by incorporating a linear layer of low-rank learning, DSRN effectively encodes correlations of landmarks to improve performance. DSRN leverages the strengths of kernels for nonlinear feature extraction and neural networks for structured prediction, and provides the first end-to-end learning architecture for direct face alignment. Its effectiveness and generality are validated by extensive experiments on five benchmark datasets, including AFLW, 300W, CelebA, MAFL, and 300VW. All empirical results demonstrate that DSRN consistently produces high performance and in most cases surpasses state-of-the-art.

1. Introduction

Face alignment or facial landmark detection has recently drawn significant attention in computer vision due to its fundamental role in various applications, including facial image analysis e.g. face recognition [36, 35], face verification [45], and facial attribute analysis [2]. Face alignment is the task of estimating a set of predefined key points, known as landmarks, providing the semantic description of facial shapes. Face alignment has been studied extensively in recent years, but several aspects of it remain unresolved. Its great challenges stem from the nonlinear relationship between input images and output shapes, since images are usually represented by low-level features while facial shapes contain high-level semantic meanings. Meanwhile, landmarks are spatially correlated, which can also be exploited for more robust and accurate alignment.

Cascaded regression has been a popular method for face alignment and made significant progress in the past decades. Nevertheless, the cascaded regression model suffers from intrinsic shortcomings. It is an indirect method and progressively estimates shape increments in an iterative way, with results highly dependent on initialization. Therefore, the final solution of cascade models is prone to getting trapped in local optima when the initialized shape is far from the true shape. Cascade models rely on local feature descriptors, and only the regions around landmarks are passed through the feature extractor. As a result, the semantic information of faces and correlations between landmarks are largely overlooked. Moreover, cascaded models extract handcrafted features, *e.g.*, SIFT [19], which fail to leverage the strength of convolutional neural networks. In addition, those local descriptors need to be calculated in each iteration based on updated shapes, which can be time-consuming and makes it hard to integrate feature learning into one single architecture for end-to-end learning.

In this paper, we propose direct shape regression networks (DRSN) to directly predict facial landmarks from images without relying on cascaded regression. DSRN tackles the aforementioned challenges by jointly modeling input-output relationships and landmark correlations in a compact end-to-end learning architecture which is composed of one doubly convolutional layer, one Fourier feature pooling layer, and one low-rank learning layer as illustrated in Fig 1.

*Corresponding Authors.

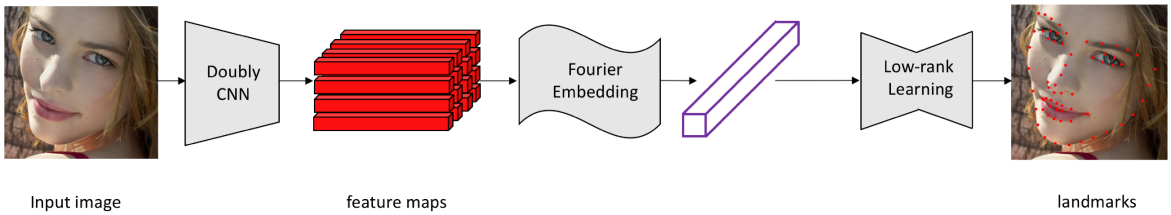


Figure 1: The learning architecture of our proposed direct shape regression network (DSRN) for end-to-end face alignment.

For feature extraction, DRSN incorporates the doubly convolutional module [46], which is computationally more efficient due to fewer parameters while improving performance compared to regular convolution. It fits well face alignment tasks where training samples are rather limited compared to other vision tasks, such as image classification. In conjunction with the doubly convolutional module, we introduce a Fourier feature pooling into the last convolutional layer, so as to build strong holistic representations. The Fourier feature pooling is derived from kernel approximation to leverage the strong ability of kernel methods for nonlinear feature extraction [30], and it greatly enhances the capability of handling the nonlinear relationship between images and shapes. More importantly, the Fourier pooling provides a nonlinear layer with a cosine activation function, which is readily learned in an end-to-end way by back-propagation.

Meanwhile, previous methods for cascaded face alignment have often overlooked landmark correlation, which has not yet been modeled explicitly. Properly modeling the correlation can not only help recover occluded landmarks, but also improve the overall estimation performance. We propose encoding the correlation in a principled way. Specifically, we design a simple but effective layer with linear low-rank learning to replace the fully connected layer as the output layer. The low-rank learning is able to explicitly encode the intrinsic correlations by forcing correlated outputs to share subsets of features. More importantly, the low-rank layer can be efficiently learned due to its nature of linearity without suffering from bad local minima.

In this work, our major contributions can be summarized in the following three aspects.

- We propose the first direct shape regression model for end-to-end face alignment without relying on cascaded regression. Our method provides a novel compact convolutional learning architecture, which leverages the strengths of kernel methods for nonlinear feature extraction and convolutional neural networks for multivariate structured prediction.

- We propose a new feature extraction layer which is composed of a doubly convolutional layer and a Fourier feature pooling layer to efficiently build strong representations, which greatly enhances its capability of disentangling the highly nonlinear relationship between images and the associated shape of facial landmarks.

- We propose a new linear layer of low-rank learning to explicitly encode the intrinsic correlation of facial landmarks. The low-rank learning layer not only improves the estimation performance but also offers a principled way to model inter-correlation of multiple outputs in structured prediction.

The effectiveness of the proposed DSRN has been verified by extensive experiments on five benchmark datasets including AFLW, 300W, CelebA, MAFL, and 300VW. Results have shown that DSRN consistently achieves high estimation accuracy on all datasets and produces the new state-of-the-art performance which, in most cases, largely surpasses previous methods. In contrast to cascaded regression models, once learned in the training stage, DSRN can efficiently predict landmarks on new input face images by simple matrix multiplications without further iterative optimization. More importantly, our DSRN offers a general compact convolutional learning architecture for multivariate estimation, which can be readily used for diverse visual tasks of structured prediction (though in this paper we focus on face alignment).

2. Related Work

Face alignment has been extensively studied and remarkable progress has occurred over the past decades [47, 8, 56, 7, 39, 40, 1, 7, 28, 49]. Previous work mainly focused on cascaded regression, which relies on iterative optimization. Cascaded regression starts with an initial shape which can be a random guess or the mean shape of training samples, and iteratively refines the shape by a cascade of regressors. Building upon cascaded regression, many improved variants have been developed which distinguish themselves by the shape initialization strategies [7], shape-indexed features [55] or regressors [42].

Xiong *et al.* proposed a supervised descent method (SDM) [42] to address the cascaded regression problem by optimizing non-linear least squares based on SIFT [19] features. Zhu *et al.* use a coarse-to-fine shape searching method to locate the landmarks. That method is robust to large pose variation [54]. To achieve high performance, they employ multiple hybrid handcrafted features, *e.g.*, SIFT, HOG and BRIEF etc, as local descriptors. Support vector regression and random forests are used by [41] for face alignment from the local image patch. By using Markov random field to model the spatial relations of landmarks, they try to resolve the predictions uncertainties. Although direct face alignment without using cascaded regression has been previously explored in [52], it is based on handcrafted features and not in an end-to-end manner.

With the great success of deep learning in feature representation, some methods use convolutional neural networks (CNNs) to learn the features or deep models to represent the regressors. Sun *et al.* [34] constructed a deep convolutional network cascaded structure to detect facial points, with multi-level regression networks. Liu *et al.* [17] not only consider the spatial domain, but also use recurrent neural networks (RNN) to get the temporal information in the video-based face alignment datasets.

However, most deep learning-based models are still based on cascaded regression, which is sensitive to improper shape initialization. Some recent methods [7, 6] attempt to solve this problem by running algorithms more than one times, but the dependence on shape initialization is still not totally avoided. Lv *et al.* [20] use a two-stage regression method. It uses spatial transformer networks [12] to transform the full face and face parts to canonical shape respectively in two stages. They call this step re-initialization. However, this method does not optimize the network parameters in the two stages jointly. The first end-to-end recurrent convolutional system for face alignment was proposed in [37]. They use CNNs to extract features and a connected RNN to approximate the cascaded process. The main difference from our end-to-end learning is that our method is direct shape regression which starts with a raw image and directly predicts coordinates of landmarks on facial shapes rather than estimating shape increments iteratively. Bulat *et al.* [5] propose a method that can also map 2D facial landmarks to 3D. We should also mention the method in [4], which is a facial alignment method explicitly designed to be lightweight and suitable for devices with limited computational resources. Obviously, our method has a different scope as it is designed for usage with modern desktop computers.

Recently, Zhang *et al.* [50] develop a multi-task deep learning framework to do the landmark detection and simultaneously learning the auxiliary attributes, such as beard, gender, wearing glasses. Unlike other related methods, they do not use cascaded steps, formulating instead face alignment as a multi-task learning problem. However, they need to use auxiliary information, e.g, facial attributes, during the training stage to ensure the performance for face alignment in the test stage. In contrast, our method directly associates images with the facial shapes by learning the mapping between them with no need for other training information.

In contrast to the existing methods for face alignment, our DSRN is, to the best of our knowledge, the first method that achieves direct shape regression in an end-to-end learning framework, without relying on cascaded regression. DSRN addresses the central issue of face alignment by effectively disentangling the highly nonlinear relationship between images and facial shapes while simultaneously encoding correlations of landmarks on the shape. It leverages the strengths of neural networks for structured prediction and kernels for nonlinear feature extraction.

3. Direct Shape Regression Network

In this section, we introduce our direct shape regression network (DSRN). We start with the problem formulation in §3.1 and describe in detail the key components of DSRN, that is, the doubly convolutional layer in §3.2, the Fourier feature pooling layer in §3.3, and the linear low-rank learning layer in §3.4. We conclude by summarizing the end-to-end learning architecture for direct face alignment in §3.5.

3.1. Preliminaries

Face alignment is the task of finding a mapping from an input image I to the facial shape S represented by the coordinates of landmarks in the form of a vector, $[x_1, y_1, \cdots, x_N, y_N]^\top \in \mathbb{R}^{2N}$, where N is the number of landmarks. DSRN directly predicts shapes from images in an end-to-end learning architecture, which handles major challenges of face alignment in one single framework. Specifically, the doubly convolutional layer in conjunction with the Fourier pooling layer are used for effective nonlinear feature extraction, to model the nonlinear relationship between images and shapes; the linear low-rank learning layer explicitly encodes intrinsic correlations of landmarks in a data-driven way for robust and improved estimation.

3.2. Doubly Convolutional Layer

Image representation plays a fundamental role in face alignment. Hand-crafted features, *e.g.*, SIFT [19] and HoGs [9], were extensively used in previous methods [54, 55, 10, 42]. The convolutional neural network (CNN) has recently emerged as a powerful tool for feature extraction and shown great success in diverse visual tasks [57].

However, the size of training data is relatively small in face alignment, while images exhibit great appearance variation and face shapes show huge variability. This poses great challenges to conventional CNNs. Instead of using

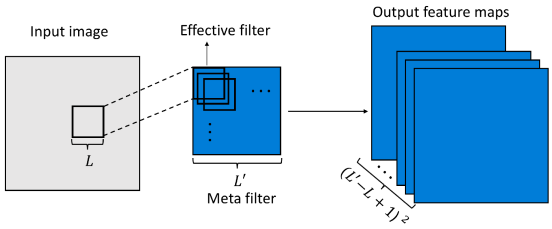


Figure 2: The structure of the doubly convolutional module. The effective filters within every meta filter are considered to be overlapped and translated versions of each other, so as to enforce parameter sharing.

regular convolutions, we use a doubly convolutional module [46], which has shown improved performance in term of both efficiency and effectiveness. The double convolution is inspired by the fact that many of filters in regular convolutions are very similar or almost translated version of each other, which induces huge redundancy. It can largely reduce the number of parameters while improving the performance, which is well suited for face alignment.

In double convolutions, there are a set of meta filters with size $L^{'} \times L^{'}$. The size of effective filters is $L \times L$ where $L < L^{'}$. Thus, we can consider that there are $(L^{'} - L + 1)^2$ effective filters within each meta filter, and the group of effective filters are forced to be translated versions of each other. When the input image is convolved with one meta filter, it convolves with each effective filter in this meta filter, to produce $(L^{'} - L + 1)^2$ feature maps for this meta filter. As a consequence, we use only one meta filter with $L^{'} \times L^{'}$ parameters, while obtaining the same number of feature maps as using $(L^{'} - L + 1)^2$ individual filters with $(L^{'} - L + 1)^2 \times L \times L$ parameters. The structure of doubly convolutional layer is shown in Fig 2.

3.3. Fourier Pooling Layer

To handle the complicated relationship between images and facial shapes, nonlinear feature extraction is usually required to achieve high-level representations. The doubly convolutional module produces a set of feature maps contained in $X \in \mathbb{R}^{w \times h \times c}$ with width w, height h and the number of maps c. For a c dimensional vector of a spatial location across the feature maps in X, we use notation $\mathbf{x} = [x_1, x_2, \cdots, x_c]^\top \in \mathbb{R}^c$. We need to pool those $w \times h$ c-dimensional feature vectors into in a compact holistic representation for shape regression.

We propose a Fourier pooling layer to aggregate feature maps by leveraging the great strength of kernels for nonlinear feature extraction [51], which enables filling the semantic gap between images and shapes. The Fourier pooling layer is derived from the approximation of shift invariant kernels [23] which is underpinned by the well-known Bochner's theorem [3].

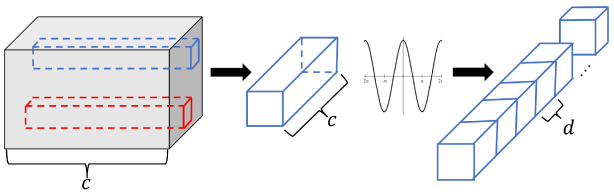


Figure 3: The structure of Fourier feature pooling.

Theorem 1 (Bochner [3]) *A continuous shift-invariant kernel function $k(\mathbf{x}, \mathbf{x}') = k(\mathbf{x} - \mathbf{x}')$ on $\mathbb{R}^d$ is positive definite if and only if it is the Fourier transform of a unique finite non-negative measure on $\mathbb{R}^d$. Defining $\zeta_\omega(\mathbf{x}) = e^{j\boldsymbol{\omega}^\top \mathbf{x}}$, for any $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$:*

$$k(\mathbf{x} - \mathbf{x}') = \int_{\mathbb{R}^d} p(\boldsymbol{\omega}) e^{j\boldsymbol{\omega}^\top (\mathbf{x} - \mathbf{x}')} d\boldsymbol{\omega} = \mathbb{E}_{\boldsymbol{\omega}}[\zeta_{\boldsymbol{\omega}}(\mathbf{x}) \zeta_{\boldsymbol{\omega}}(\mathbf{x}')^*] , \tag{1}$$

*where * is the conjugate and $p(\boldsymbol{\omega})$ is the Fourier transform of the kernel.*

The kernel $k(\mathbf{x}, \mathbf{x}')$ can be approximated by drawing d random samples as:

$$k(\mathbf{x}, \mathbf{x}') \approx \sum_{i=1}^{d} \left\langle \sqrt{\frac{2}{d}} \cos(\boldsymbol{\omega}_i^\top \mathbf{x} + b_i), \sqrt{\frac{2}{d}} \cos(\boldsymbol{\omega}_i^\top \mathbf{x}' + b_i) \right\rangle , \tag{2}$$

where $\boldsymbol{\omega}$ is sampled from the probablity distribution $p(\boldsymbol{\omega})$, and b is uniformly sampled over $[0, 2\pi]$.

Therefore, the corresponding approximated feature map $\boldsymbol{\phi}(\mathbf{x})$ is:

$$\boldsymbol{\phi}(\mathbf{x}_i) = \sqrt{\frac{2}{d}} [\cos(\boldsymbol{\omega}_i^\top \mathbf{x}_i + b_i)]_{1:d} , \tag{3}$$

where $\boldsymbol{\phi}(\mathbf{x})$ is called the random Fourier feature [23], and has been successfully used in various kernel methods.

However, the great power of kernel approximation based on random Fourier features remains largely underdeveloped, and this topic has recently attracted attention [22, 33]. In most of the existing kernel approximation methods, the sampling is independent of input data distributions, and this usually requires high-dimensional feature maps to achieve kernel approximation with satisfactory performance. Moreover, since no learning is involved, the approximate feature maps would be of high redundancy and of low discriminant ability, which compromises performance while inducing unnecessary computational cost. In addition, approximating the kernel with a fixed configuration does not necessarily lead to high performance since it remains an open question how to choose the best kernel configuration.

Instead of approximating kernels by random sampling from data-independent distributions, we learn the parameters $\{\boldsymbol{\omega}, b\}$ from data in a supervised way, which enables more compact but highly discriminative feature representations. Define $W = [\boldsymbol{\omega}_1, \cdots, \boldsymbol{\omega}_d] \in \mathbb{R}^{d \times c}$ and $\boldsymbol{b} = [b_1, \cdots, b_d]$. We define a nonlinear layer of neural networks with cosine activations:

$$\boldsymbol{\phi}_i = \cos(W\mathbf{x}_i + \mathbf{b}), \tag{4}$$

where $\cos$ is an element-wise function, i indicates the i-th location in the feature map X, and W is the weight matrix of the nonlinear layer. The induced Fourier feature pooling layer can be seamlessly integrated with the doubly convolutional layer to achieve a fully end-to-end learning architecture that can be trained via back-propagation.

To achieve a holistic representation, we concatenate the embedded feature vectors into a single vector $\mathbf{z} = [\boldsymbol{\phi}_1, \cdots, \boldsymbol{\phi}_i, \cdots, \boldsymbol{\phi}_p] \in \mathbb{R}^D$, where $p = w \times h$, *i.e.*, the number of locations. In contrast to feature pooling techniques by directly summing up the feature vectors, the concatenation can well preserve the spatial information of images, which is of great importance for predicting the spatial locations of facial landmarks.

3.4. Low-rank Learning Layer

We propose a simple but effective layer to encode correlations of landmarks by linear low-rank learning. Having the holistic representation $\mathbf{z}$, a straightforward way for prediction is to use a fully connected layer with the regression matrix represented by $M \in \mathbb{R}^{Q \times D}$, where Q is the number of outputs, *i.e.*, $Q = 2N$, which gives $\mathbf{y} = M\mathbf{z}$. An identity activation function is used by default. Although sharing the holistic representations, landmark correlations are not explicitly encoded. Low-rank constraints, such as the nuclear norm [51], could be simply imposed to force the regression matrix M to be low rank, but this does not always guarantee low-rankness of M, and can fail to fully capture the correlations. Instead of using one fully connected layer, we propose linear low-rank learning layer to explicitly encoding correlations of landmarks.

Specifically, we propose the low-rank learning layer by replacing the single matrix M with multiplication of two low-rank matrices, which gives rise to

$$\mathbf{y} = M\mathbf{z} = U^\top V\mathbf{z}, \tag{5}$$

where $U \in \mathbb{R}^{P \times Q}$, $V \in \mathbb{R}^{P \times D}$ and $P \leq Q$. The linear function provides a low-rank layer to explicitly encode inter-output correlations. U and V are learned in a data-driven way without relying on any specific assumptions, and can adaptively capture specific correlations in different applications.

Low-rank learning brings two attractive advantages compared to nuclear norm based minimization. First, it establishes an overall mapping of M with guaranteed low rankness to explicitly encode correlations; related outputs are forced to share similar regression parameter patterns [53], and thus knowledge is transferred across correlated outputs. This can significantly improve the overall prediction performance. Second, low-rank learning avoids solving complicated rank-constrained problems and leverages the great effectiveness of linear learning, which enjoys great computational efficiency; by setting $P \ll Q$, the low-rank learning can greatly reduce the number of parameters, which is especially advantageous when using iterative optimization with stochastic gradient descent [29].

3.5. End-to-End Direct Face Alignment

The doubly convolutional layer, the Fourier pooling layer and the low-rank learning layer are used to define our direct shape regression network (DSRN), which is a novel compact end-to-end learning architecture for direct face alignment. In contrast to the cascaded regression models, DSRN is trained in one single framework by back-propagation by directly associating images with the coordinates of landmarks on facial shapes; in the test stage, DSRN predicts facial shapes of input images by simple matrix multiplications rather than iterative optimization, which leads to improved efficiency. More importantly, the proposed DSRN is highly generalizable and can be readily adapted to other structured prediction tasks with multiple continuous outputs.

4. Experiments Results

We have conducted extensive experiments on five benchmark datasets, and we provide a comprehensive comparison with state-of-the-art methods. The proposed direct shape regression network (DSRN) consistently yields high accuracy for face alignment, and in most cases outperforms previous methods by large margins. Moreover, the consistently high performance on the five diverse face alignment tasks demonstrates the generality of our method.

4.1. Datasets

The five datasets used in our experiments are commonly used benchmarks for face alignment. Faces in the datasets are collected in uncontrolled scenarios, demonstrating great variations, which pose significant challenges for face alignment. We provide the detailed description of those datasets to facilitate direct comparison with previous work under the same experimental settings.

AFLW [15] contains a total of 24386 face images gathered from Flickr. In contrast to other databases limited to frontal views or acquired under controlled conditions. AFLW faces are collected in the wild, have large-scale

Table 1: Comparison on AFLW.

Method	Error	Year
CDM [43]	5.43	2013
PCPR [6]	3.73	2013
ERT [13]	4.35	2014
SDM [42]	4.05	2013
LBF [24]	4.25	2014
PO-CR [38]	5.32	2015
CFSS [54]	3.92	2015
CLL [55]	2.72	2016
DAC-CSR [10]	2.27	2017
DRA-TSR [20]	2.17	2017
DSRN	**1.86**	

pose variations up to $\pm 90^{\circ}$ and also have large variety in face appearance (*e.g.*, pose, expression, ethnicity, gender). Each image is annotated with 21 landmarks. Following the experimental settings of cascaded compositional learning (CLL) [55], we ignore the two ear points and use the same 20000 and 4386 images for training and test, respectively.

300W [26, 27] consists of several datasets including AFW [56], HELEN [16], LFPW [2], XM2VTS [21]. In addition, it contains a challenging 135-image IBUG [31] set. Following the same dataset configuration in [54], our training set of 3148 images consists of the full set of AFW and the training sets of HELEN and LFPW. The full test set (689 images) is divided into a "common subset" (554 images), which contains the test sets from LFPW and HELEN, and a "challenging subset" (135 images) which is from IBUG. 300W has a 68-points annotation for each face image.

CelebA [18] is a large-scale face dataset with 202599 images. CelebA provides 5 landmarks of the facial shape for each image. The images show large pose variations and background clutter. Because of large diversities and large quantities, CelebA is suitable for training and testing a deep learning model. Following the original work [18], 182631 and 19926 images are used respectively for the training and test sets.

MAFL is a subset of CelebA. To benchmark with previous methods, we follow the experimental settings in [50]. Specifically, we sample the same 20000 faces from CelebA and select the same 1000 faces for testing as in [50].

300VW [31] is a video-based face alignment dataset which contains 114 videos from different conditions. We extract face images from the same 50 videos as [31] to train the model, and the remaining 64 videos are divided into three test sets.

4.2. Implementation Details

We use four doubly convolutional layers and four pooling layers for the feature extraction task. Multiple feature maps are produced in each convolutional layer. Following each convolution operation, we use rectified linear unit as activation function and the $5 \times 5, 5 \times 5, 3 \times 3, 3 \times 3$ max pooling. After that, the Fourier pooling layer is added to the feature maps $X \in \mathbb{R}^{8\times 8\times 256}$. In Fourier pooling, we obtain $X^{'} \in \mathbb{R}^{8\times 8\times d}$ first, where the value of d may be changed depending on the size of training samples and the number of landmarks in the task. Then we do simple concatenation for $X^{'}$ to achieve the holistic representation.

In the low-rank learning layer, we do not use any nonlinear activtion functions but just the linear function with identity activations. The commonly used weight decay and batch normalization [11] techniques are also used. The parameter for weight dacay is 0.001. We employ the stochastic optimization algorithm Adam [14] to learn the parameters of the neural network. The mini-batch size is set to 64. The codes are available at *https://github.com/xinxinmiao/DSRN*.

For all experiments, the original bounding box given by the dataset is used, without any data augmentation. For the 300W dataset, due to the size of the training set being relative small, we pre-train our model on the large-scale 300VW dataset which has the same number, 68, of landmarks, and fine tune it on the training set of 300W to obtain the final model.

We use the normalized mean error (NME) as the evaluation metric, which is defined as follows:

$$\mathrm{NME} = \frac{\frac{1}{N}\sum_{i=1}^{N}\sqrt{(\hat{x}_i - x_i)^2 + (\hat{y}_i - y_i)^2}}{d}, \quad (6)$$

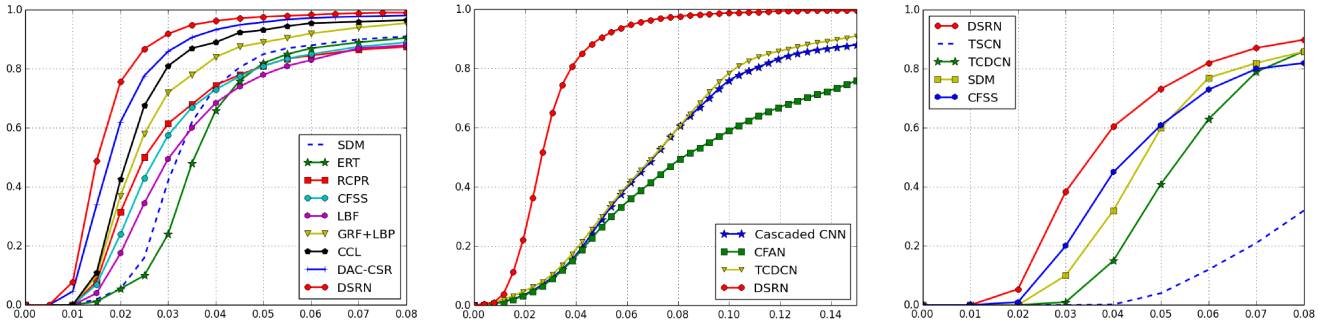
where (x, y) and $(\hat{x}, \hat{y})$ denotes the ground truth and predicted coordinates, respectively, N denotes the number of landmarks on facial shapes, and d is the distance for normalization.

Following previous work, for 300W, CelebA, MAFL and 300VW, we use the inter-ocular distance to normalize the mean error; for AFLW, we use face size to normalize mean error since the inter-ocular distance of many faces is close to zero. For brevity, $\%$ is omitted in all tables. We also show the evaluation results in the form of cumulative error distribution (CED) curve for comprehensive comparison.

4.3. Performance and Comparison

Our DSRN consistently achieves high performance on all five datasets and outperforms previous methods in most cases by large margins.

On AFLW, as shown in Table 1, DSRN achieves the best error rate, 1.86%, compared to the previous best error rate of 2.17% [20]. In Fig 4 (a), the curve of our DSRN is clearly above those of other methods, which also indicates the performance advantages. Compared with those methods based on cascaded regression, our DSRN can detect the landmarks for side faces accurately as shown by the intuitive illustration in the fourth and seventh images of Fig 5 (a).



AFLW MAFL 300VW

Figure 4: Comparions on AFLW, MAFL and 300VW in terms of CED.

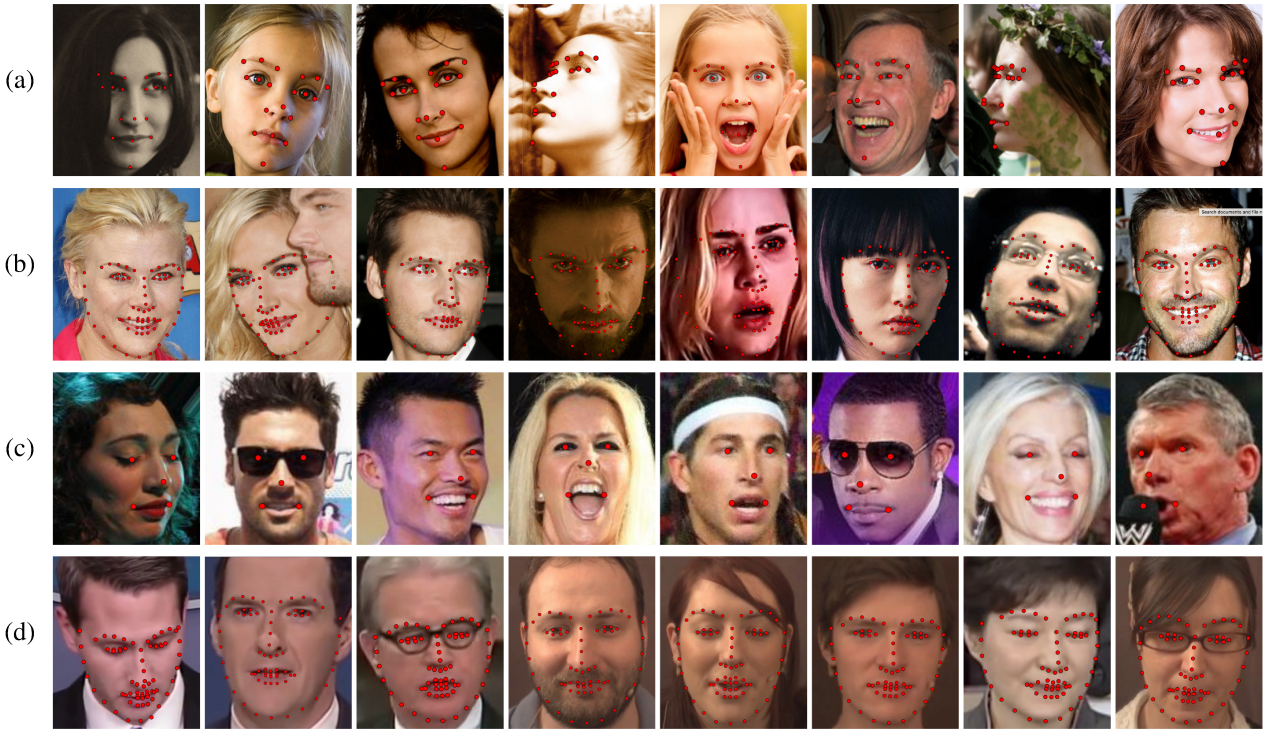


Figure 5: Illustrative results on (a) AFLW (b) 300W (c) CelebA (d)300VW .

On 300W, our DSRN achieves competitive performance, which is better than all previous methods except for [20], which gives better results on the challenging set and the full test set. The challenges of 300W stem from the great variations of images while with limited training data. As shown in Fig 5, our DSRN can accurately predict the landmarks on faces with large orientations and diverse expressions.

On CelebA and MAFL, as can be seen in Table 3, our DSRN achieves the best performance on both datasets, with error rates of 3.08% and 3.15% respectively, which are significant improvements over the previous best error rates of 3.95% and 7.95% respectively. In Fig 4 (b), we can see that there is a big gap between DSRN and TCDCN, which uses the similar convolutional network with DSRN and takes advantage of face attributes, but without Fourier pooling and low-rank learning layer. In the second and sixth images of Fig 5 (c), when the eyes in face images are occluded by sunglasses, DSRN can still predict the landmarks correctly. This is because our low-rank learning can encode the intrinsic correlation of landmarks.

On 300VW, as shown in Table 4, DSRN produces the highest accuracy on Tests 1 and 3, where Test 3 is regarded

Figure 6: Illustration of Fourier features.

Table 2: Comparison on 300W.

Method	Common Subset	Challenging Subset	Full Test set
RCPR [6]	6.18	17.26	8.35
SDM [42]	5.57	15.40	7.52
ESR [7]	5.28	17.00	7.58
GN-DPM [40]	5.78	-	-
ERT [13]	-	-	6.40
CFAN [48]	5.50	16.78	7.69
LBF [25]	4.95	11.98	6.32
DDN [44]	-	-	5.59
CFSS [54]	4.73	9.98	5.76
MDM [37]	4.83	10.14	5.88
DRA-TSR [20]	4.36	**7.56**	**4.99**
DSRN	**4.12**	9.68	5.21

Table 3: Comparison on CelebA and MAFL.

Method	CelebA	MAFL
TCDCN [50]	-	7.95
Cascaded CNN [34]	-	9.73
CFAN [48]	-	15.84
RCPR [6]	4.12	-
SDM [42]	4.35	-
CFSS [54]	3.95	-
DSRN	**3.08**	**3.15**

as the most challenging subset. We have also compared with TSTN [17] designed specifically for video-based face alignment by modeling the temporal relationship across frames. Our method achieves overall better performance than TSTN. Moreover, DSRN can run very fast with about 500 frames per second excluding face detection in the platform of NVIDIA GTX 1080Ti GPU, which is promising for the prospect of practical application. The intuitive results of 300VW are shown in Fig 5, our DSRN can accurately predict the shapes of images with great appearance variations.

In addition, since the Fourier pooling layer serves as important role in feature learning, it would be interesting to look inside into features after Fourier pooling. Assume that feature maps from the last convolutional layer are denoted by $X^{'} \in \mathbb{R}^{8\times8\times d}$, where. We illustrate the features output from Fourier pooling and their corresponding input images in Fig 6. It is easy to see that the Fourier pooling layer tries to learn the shape of a face, largely preserving spatial structure information, which will be of great benefit for accurate facial landmark prediction.

Table 4: Comparison on 300VW.

Method	Test 1	Test 2	Test 3	Year
SDM [42]	7.41	6.18	13.04	2013
TSCN [32]	12.54	7.25	13.13	2014
CFSS [54]	7.68	6.42	13.67	2015
TCDCN [50]	7.66	6.77	14.98	2016
TSTN [17]	5.36	**4.51**	12.84	2017
DSRN	**5.33**	4.92	**8.85**	-

5. Conclusions

In this paper, we propose the direct shape regression network (DSRN) for end-to-end face alignment. DSRN consists of the doubly convolutional layer, the novel Fourier pooling layer, and the low-rank learning layer. These layers enable jointly handling nonlinear image-shape relationships and the intrinsic correlations between landmarks. Our DSRN offers a new learning architecture that combines the strengths of kernels for nonlinear feature extraction and neural networks for structured prediction. Experimental results on five benchmark datasets have shown that our DSRN achieves superior performance on all datasets.

Acknowledgments XM was supported by U.S. NSF IIS 1633753, IIS 1565328 and IIP 1719031. XZ was supported by NSF of China 61571147. XZ worked at UTA for this project first and joined Beihang University later. VA was supported by U.S. NSF IIS 1565328 and IIP 1719031. HH was supported by U.S. NSF IIS 1302675, IIS 1344152, DBI 1356628, IIS 1619308, IIS 1633753, NIH R01 AG049371.

References

[1] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic. Robust discriminative response map fitting with constrained local models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3444–3451, 2013.

[2] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar. Localizing parts of faces using a consensus of exemplars. *IEEE transactions on pattern analysis and machine intelligence*, 35(12):2930–2940, 2013.

[3] S. Bochner. *Lectures on Fourier Integrals.(AM-42)*, volume 42. Princeton University Press, 2016.

[4] A. Bulat and G. Tzimiropoulos. Binarized convolutional landmark localizers for human pose estimation and face alignment with limited resources. *arXiv preprint arXiv:1703.00862*, 2017.

[5] A. Bulat and G. Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). *arXiv preprint arXiv:1703.07332*, 2017.

[6] X. P. Burgos-Artizzu, P. Perona, and P. Dollár. Robust face landmark estimation under occlusion. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1513–1520, 2013.

[7] X. Cao, Y. Wei, F. Wen, and J. Sun. Face alignment by explicit shape regression. *International Journal of Computer Vision*, 107(2):177–190, 2014.

[8] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *IEEE Transactions on pattern analysis and machine intelligence*, 23(6):681–685, 2001.

[9] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005.

[10] Z.-H. Feng, J. Kittler, W. Christmas, P. Huber, and X.-J. Wu. Dynamic attention-controlled cascaded shape regression exploiting training data augmentation and fuzzy-set sample weighting. *arXiv preprint arXiv:1611.05396*, 2016.

[11] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456, 2015.

[12] M. Jaderberg, K. Simonyan, A. Zisserman, et al. Spatial transformer networks. In *Advances in Neural Information Processing Systems*, pages 2017–2025, 2015.

[13] V. Kazemi and J. Sullivan. One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1867–1874, 2014.

[14] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[15] M. Koestinger, P. Wohlhart, P. M. Roth, and H. Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 2144–2151. IEEE, 2011.

[16] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang. Interactive facial feature localization. In *European Conference on Computer Vision*, pages 679–692. Springer, 2012.

[17] H. Liu, J. Lu, J. Feng, and J. Zhou. Two-stream transformer networks for video-based face alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[18] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3730–3738, 2015.

[19] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[20] J. Lv, X. Shao, J. Xing, C. Cheng, and X. Zhou. A deep regression architecture with two-stage reinitialization for high performance facial landmark detection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[21] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre. Xm2vtsdb: The extended m2vts database. In *Second international conference on audio and video-based biometric person authentication*, volume 964, pages 965–966, 1999.

[22] J. Moeller, V. Srikumar, S. Swaminathan, S. Venkatasubramanian, and D. Webb. Continuous kernel learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 657–673. Springer, 2016.

[23] A. Rahimi and B. Recht. Random features for large-scale kernel machines. In *Advances in neural information processing systems*, pages 1177–1184, 2008.

[24] S. Ren, X. Cao, Y. Wei, and J. Sun. Face alignment at 3000 fps via regressing local binary features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1685–1692, 2014.

[25] S. Ren, X. Cao, Y. Wei, and J. Sun. Face alignment via regressing local binary features. *IEEE Transactions on Image Processing*, 25(3):1233–1245, 2016.

[26] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: Database and results. *Image and Vision Computing*, 47:3–18, 2016.

[27] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 397–403, 2013.

[28] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. A semi-automatic methodology for facial landmark annotation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 896–903, 2013.

[29] T. N. Sainath, B. Kingsbury, V. Sindhwani, E. Arisoy, and B. Ramabhadran. Low-rank matrix factorization for deep neural network training with high-dimensional output targets. In *ICASSP*, 2013.

[30] B. Scholkopf and A. J. Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2001.

[31] J. Shen, S. Zafeiriou, G. G. Chrysos, J. Kossaifi, G. Tzimiropoulos, and M. Pantic. The first facial landmark track-

ing in-the-wild challenge: Benchmark and results. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 50–58, 2015.

[32] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576, 2014.

[33] A. Sinha and J. C. Duchi. Learning kernels with random features. In *Advances in Neural Information Processing Systems*, pages 1298–1306, 2016.

[34] Y. Sun, X. Wang, and X. Tang. Deep convolutional network cascade for facial point detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3476–3483, 2013.

[35] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1701–1708, 2014.

[36] G. Toderici, G. Passalis, S. Zafeiriou, G. Tzimiropoulos, M. Petrou, T. Theoharis, and I. A. Kakadiaris. Bidirectional relighting for 3d-aided 2d face recognition. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2721–2728. IEEE, 2010.

[37] G. Trigeorgis, P. Snape, M. A. Nicolaou, E. Antonakos, and S. Zafeiriou. Mnemonic descent method: A recurrent process applied for end-to-end face alignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4177–4187, 2016.

[38] G. Tzimiropoulos. Project-out cascaded regression with an application to face alignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3659–3667, 2015.

[39] G. Tzimiropoulos and M. Pantic. Optimization problems for fast aam fitting in-the-wild. In *Proceedings of the IEEE international conference on computer vision*, pages 593–600, 2013.

[40] G. Tzimiropoulos and M. Pantic. Gauss-newton deformable part models for face alignment in-the-wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1851–1858, 2014.

[41] M. Valstar, B. Martinez, X. Binefa, and M. Pantic. Facial point detection using boosted regression and graph models. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2729–2736. IEEE, 2010.

[42] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 532–539, 2013.

[43] X. Yu, J. Huang, S. Zhang, W. Yan, and D. N. Metaxas. Pose-free facial landmark fitting via optimized part mixtures and cascaded deformable shape model. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1944–1951, 2013.

[44] X. Yu, F. Zhou, and M. Chandraker. Deep deformation network for object landmark localization. In *European Conference on Computer Vision*, pages 52–70. Springer, 2016.

[45] S. Zafeiriou, G. Tzimiropoulos, M. Petrou, and T. Stathaki. Regularized kernel discriminant analysis with a robust kernel for face recognition and verification. *IEEE transactions on neural networks and learning systems*, 23(3):526–534, 2012.

[46] S. Zhai, Y. Cheng, Z. M. Zhang, and W. Lu. Doubly convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1082–1090, 2016.

[47] J. Zhang, M. Kan, S. Shan, and X. Chen. Occlusion-free face alignment: deep regression networks coupled with de-corrupt autoencoders. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3428–3437, 2016.

[48] J. Zhang, S. Shan, M. Kan, and X. Chen. Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment. In *European Conference on Computer Vision*, pages 1–16. Springer, 2014.

[49] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.

[50] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. Learning deep representation for face alignment with auxiliary attributes. *IEEE transactions on pattern analysis and machine intelligence*, 38(5):918–930, 2016.

[51] X. Zhen, M. Yu, X. He, and S. Li. Multi-target regression via robust low-rank learning. *IEEE transactions on pattern analysis and machine intelligence*, 40(2):497–504, 2018.

[52] X. Zhen, M. Yu, A. Islam, M. Bhaduri, I. Chan, and S. Li. Descriptor learning via supervised manifold regularization for multioutput regression. *IEEE transactions on neural networks and learning systems*, 28(9):2035–2047, 2017.

[53] X. Zhen, M. Yu, F. Zheng, I. B. Nachum, M. Bhaduri, D. Laidley, and S. Li. Multitarget sparse latent regression. *IEEE transactions on neural networks and learning systems*, 2017.

[54] S. Zhu, C. Li, C. Change Loy, and X. Tang. Face alignment by coarse-to-fine shape searching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4998–5006, 2015.

[55] S. Zhu, C. Li, C.-C. Loy, and X. Tang. Unconstrained face alignment via cascaded compositional learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3409–3417, 2016.

[56] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2879–2886. IEEE, 2012.

[57] Y. Zhu, Z. Lan, S. Newsam, and A. G. Hauptmann. Hidden two-stream convolutional networks for action recognition. *arXiv preprint arXiv:1704.00389*, 2017.