Leveraging Intra and Inter-Dataset Variations for Robust Face Alignment

Wenyan Wu
Department of Computer Science and Technology
Tsinghua University
wwy15@mails.tsinghua.edu.cn

Shuo Yang
Department of Information Engineering
The Chinese University of Hong Kong
ys014@ie.cuhk.edu.hk

Abstract

Face alignment is a critical topic in the computer vision community. Numerous efforts have been made and various benchmark datasets have been released in recent decades. However, two significant issues remain in recent datasets, e.g.*, Intra-Dataset Variation and Inter-Dataset Variation. Inter-Dataset Variation refers to bias on expression, head pose,* etc. *inside one certain dataset, while Intra-Dataset Variation refers to different bias across different datasets. In this study, we show that model robustness can be significantly improved by leveraging rich variations within and between different datasets. This is non-trivial because of inconsistent landmark definitions between different datasets and the serious data bias within one certain dataset.*

To address the mentioned problems, we proposed a novel Deep Variation Leveraging Network (DVLN), which consists of two strong coupling sub-networks, e.g.*, Dataset-Across Network (DA-Net) and Candidate-Decision Network (CD-Net). In particular, DA-Net takes advantage of different characteristics and distributions across different datasets, while CD-Net makes a final decision on candidate hypotheses given by DA-Net to leverage variations within one certain dataset. Extensive evaluations show that our approach demonstrates real-time performance and dramatically outperforms state-of-the-art methods on the challenging 300-W dataset.*

1. Introduction

Face alignment aims at locating a sparse set of fiducial facial landmarks. It is a critical component in many face analysis tasks, such as face recognition [9, 50], face verification [35, 36], and robust face frontalisation [14]. Although many efforts have been devoted in solving this task and rapid progress has been made during the past decades [25, 8, 12, 46, 45, 38, 5]. Face alignment remains a very challenging problem. The challenge mainly comes from the large variations of facial appearance, *e.g.*, most changes come from different poses, lightening conditions and expressions.

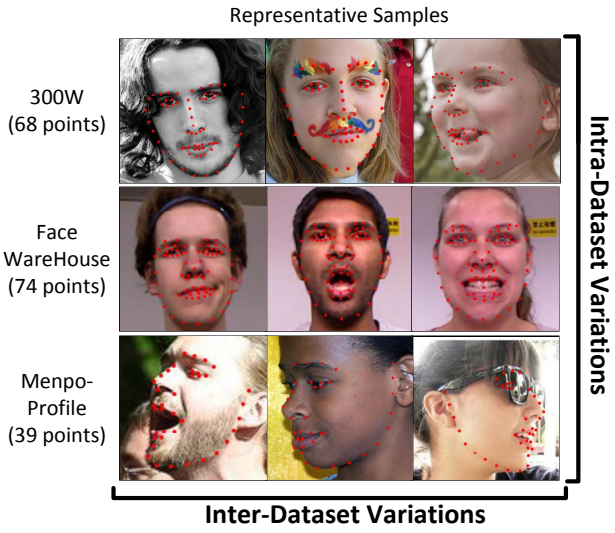


Figure 1: Intra and Inter-Dataset Variations appear in recent popular datasets, *e.g.*, 300-W [27], FaceWareHouse [7], and Menpo-Profile [43]. Between the rows, great bias and the inherent annotation gaps issue are shown across datasets, while between columns, severe bias are also shown inside one certain dataset.

Numerous approaches have been proposed over these years. Classic methods including Active Shape Models (ASMs) [24, 16], Active Appearance Models (AAMs) [11, 29, 22, 18], and Constrained Local Models (CLMs) [20, 30] estimate a parametric model for the spatial configuration of landmarks, often referred as shape models. Regression based methods [8, 5, 40, 10], which map the discriminative features around landmarks to the desired landmark positions have been proposed and shown high effectiveness. Besides traditional models, very recent works

adapt convolutional neural network (CNN) on face alignment [34, 46, 4, 31] with the advent of Deep Learning. These methods show superior accuracy compared to previous methods and existing commercial systems.

At the same time, various benchmarks have been released, from datasets under laboratory condition like Multi-PIE [13] and XM2VTSDB [23], to more recent in-the-wild datasets like LFPW [3], AFLW [19], AFW [25], HELEN [21], IBUG [28], 300-W [27] and Menpo [43]. In this study, we highlight two significant issues among recent datasets, *e.g.*, *Intra-Dataset Variation* and *Inter-Dataset Variation*. Such issues are briefly illustrated in Fig.1.

Intra-Dataset Variation refers to recent datasets show very different bias while being compared with each other by considering different aspects, *e.g.*, expressions, occlusions and head poses. For instance, AFLW [19] has 16.4% profile faces, while AFW [25] consists of only 3.1%. Such bias is also common in expression, *e.g.*, 300-W [27] has 14.16% scream face, while HELEN [21] has even no scream face. One obvious problem is that a model trained in one dataset will lead to severe over-fitting easily and perform poor generalization in recent in-the-wild datasets with large variations of facial appearance. Naturally, unifying these datasets from different distributions to train one model can significantly help alleviate this problem. This though, however, is hindered by the annotation gaps, *e.g.*, different datasets have different landmark definitions (for example, Menpo [43] has 68-landmark markup, while AFLW [19] has 21-landmark markup). Ideally, different datasets can be re-announced using unified annotation. Nevertheless, it is quite labor-intensive and time-consuming. One objective of this study is to formulate an approach to train a single model by leveraging datasets with varying annotations.

To address the *Intra-Dataset Variation* issue, we propose a *Dataset-Across Network (DA-Net)*, which can integrate different datasets and take advantage of different distributions across datasets. As shown in Fig.2, for each dataset with different annotation, images are fed into a convolutional networks for feature extraction. In this stage, weights are shared across different datasets to get high-level representations. We argue that this mechanism can implicitly guide the network to learn the general feature of faces, *e.g.*, face shape, component relationship and so on, from different datasets. While the last fully connected layers are sliced into n splits, where n is the number of datasets. We regard these split layers as a mechanism to learn different landmark definitions from different datasets. The experiments show the proposed method is effective in robust facial points detection and can significantly prevent the network from over-fitting. The light-weight CNN network also makes our framework an efficient solution (66 FPS on a single core i5-4300u CPU).

Inter-Dataset Variation refers to recent datasets have great bias even inside the dataset itself, especially in form of the *yaw* angle of head pose. For instance, Menpo-Profile [43] dataset has about 27.1% extreme left view faces (with $yaw < -80°$) while 72.9% extreme right view faces (with $yaw > 80°$). It will lead to poor generalization due to the lack of training samples of certain facial characteristic, for example, left view faces.

To address this issue, we further propose a *Candidate-Decision Network (CD-Net)*, which is closely coupled with *DA-Net*. Specially, all of the images are split into two parts, *e.g.*, left and right view. Then, all of the right view faces are flipped to get one united dataset with only left view. A *DA-Net* is trained on this dataset to enjoy all of the variations of both views. Moreover, to handle the test faces with right view, a *CD-Net* is further trained. Thus, in the test process, each image and its flipped version are taken together into the trained *DA-Net* to generate two candidate predictions. The trained *CD-Net* will act like a arbiter, whose job is to decide which candidate hypothesis given by *DA-Net* will be finally taken.

In summary, the contributions of this work are:

1. We proposed a *Dataset-Across Network (DA-Net)* to combine different datasets together to train an end-to-end single model. This framework makes it possible to take advantage of different variations in different datasets and greatly improves the generalization performance.

2. We further introduced a *Candidate-Decision Network (CD-Net)*, which is closely coupled with *DA-Net* to choose one better prediction. This network can take advantage of various characteristics of faces in one certain dataset and meanwhile maintain the performance while encountering diverse views of the faces.

3. Detailed experimental evaluations show our method demonstrates real-time performance and outperforms existing state-of-the-art methods on the challenging 300-W [27] dataset.

The reminder of this paper is organised as follows. In Sec.2, we provide an overview of the related work. Subsequently, in Sec.3, we describe in detail the architecture design of our proposed *Deep Variation Leveraging Network (DVLN)*, which is composed of *Dataset-Across Network (DA-Net)* and *Candidate-Decision Network (CD-Net)*. The results of experiments are shown in Sec.4 to evaluate the effectiveness of our method. Finally, the paper is concluded in Sec.5.

2. Related Work

2.1. Generic Face Alignment

In the past decades, a number of achievements have been made including the classic Active Shape Models (ASMs)

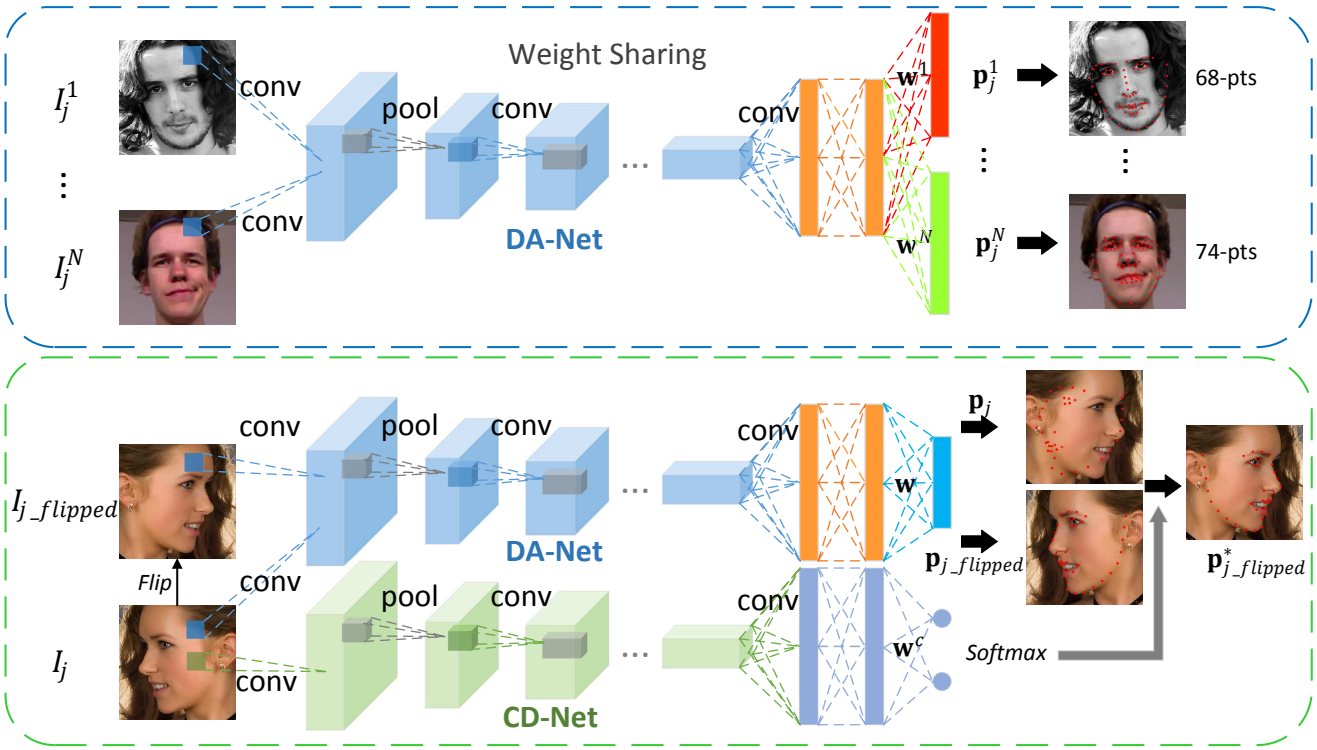


Figure 2: Structure of our proposed Deep Variation Leveraging Network (DVLN). In the top half of Figure.2, Dataset-Across Alignment process using a Dataset-Across Network (DA-Net) is shown. Images from different datasets firstly pass a set of shared parameters and then do a dataset-specific regression respectively to generate predictions with different landmark definitions. In the bottom half of Figure.2, Candidate Decision mechanism using a Candidate-Decision Network (CD-Net) is shown. A image from one certain dataset and its flipped version are taken together into the trained DA-Net to generate two candidate predictions. Then, the trained CD-Net makes a final decision. (Best viewed in color)

[24, 16], Active Appearance Models (AAMs) [11, 29, 22, 18], and Constrained Local Models (CLMs) [20, 30]. Subsequently, regression based face alignment approaches are proposed, which directly regress shape update based on local feature extracted from all current estimated landmarks. Xiong *et al*. [41] predict shape increment by applying linear regression on SIFT features. Both Cao *et al*. [8] and Burgos-Artizzu *et al*. [5] use boosted ferns to regress the shape increment.

Besides traditional models, convolutional neutral network (CNN) has also been employed in and achieved superior accuracy. Such methods are close to our approach. Sun *et al*. [34] firstly formulate the face alignment as a regression problem and use CNN to locate the landmarks in a coarse-to-fine manner. Zhang *et al*. [46] frame the problem as a multi-task learning problem. Our proposed method is somewhat analogous to the multi-task learning mechanism. However, the significant difference is that we do not need any re-annotation while multi-task learning network proposed in [46] needs extra facial attributes annotation which is sometimes impractical. MDM [37] is the first end-to-end recurrent convolutional system for deformable object alignment and improved on the state-of-the-art in face alignment on the challenging test-set of 300-W [27] competition by a large margin. RAR [39] follows the pipeline of cascaded regressions. By refining the landmark locations sequentially and introducing LSTM models, it demonstrates superior performance on several benchmarks.

2.2. Face Alignment by Leveraging Data Variations

There are a few works have attempted to solve the datasets bias issue intra-datasets. Smith *et al*. [33] proposed a method that takes multiple source datasets as input and labels a partially labeled target dataset using a union of landmarks defined in the source datasets. It is the first effort to combine multiple face landmark datasets with different landmark definitions for prediction. However, it can only jointly align all testing images together but can not handle the single test image scenario and suffers from high computational cost (more than 30 seconds per image). Zhu *et*

al. [48] take SDM as basis and formulate a novel Transductive Cascaded Regression (TCR) method, whose core is a transductive alignment approach. This method is capable of transferring annotation style from one dataset to another. It performs well in cross-dataset evaluation and even unseen domain. However, it suffers from some limitation. For example, it cannot handle samples unsuited for SDM [41] (*e.g*. with one eye totally unseen in AFLW [19]). Zhang *et al*. [44] present a unified deep regression network coupled with the sparse shape regression (DRN-SSR) to predict the union of all types of landmarks. However, it is integrated with cascaded regression, in which each linear regressors are learnt independently. The correlations between semantically related images are not taken into account. In addition, shape-indexed feature which used to drive the cascade may be sub-optimal and leading performance degradation. In addition, to solve the datasets bias issue inter-datasets, Yang *et al*. [42] proposed a supervised initialisation scheme for cascaded face alignment based on explicit head pose estimation. It can decrease failures casused by improper initialisation with large head pose variations. However, it brings much more complexity as an additional head pose estimation model is needed to be trained separately.

3. Our Approach

In this section, we describe in detail the architecture design of our proposed *Deep Variation Leveraging Network (DVLN)*, which consists of *Dataset-Across Network (DA-Net)* and *Candidate-Decision Network (CD-Net)*. For both of these two networks, we firstly illustrate a brief overview and then describe the formulation and detail the learning process.

3.1. Dataset-Across Network

Overview of DA-Net. Given N training datasets $\mathbf{T} = \{T^i\}_{i=1}^N$ with different landmarks definition. The landmarks defined in T^i can be denoted as $\mathbf{s}^i \in \mathbb{R}^{l^i}$, where l^i represent the number of x-,y-coordinate of the landmarks in T^i. For instance, l^i equals $68 \times 2 = 136$ for Menpo [43] with 68-landmark markup. The corresponding ground truth landmarks set is denoted as $\mathbf{S} = \{\mathbf{s}^i\}_{i=1}^N$, where $\mathbf{s}^i \in \mathbb{R}^{l^i \times M^i}$ and M^i is the number of images in dataset T^i. The goal of face alignment is to learn a model Θ, to estimate the location of landmarks. Specifically, for a image $I^i \in T^i$, the model learns to recover the landmarks locations $\mathbf{p}^i = \Theta(I^i)$ to predict $\mathbf{s}^i$.

As shown in Fig.2, the *Feature Extraction* part of *DA-Net* takes in a batch of images. One batch consists of equal numbers of images from N training datasets $\mathbf{T}$. Convolutional filters are shared among these training images, thus take advantage of rich variations of facial characteristics of different datasets. It leads to a much more robust model for feature extraction. Then, in the *Regression* part, a set of linear regressors defined by the last fully connected layers are used to generate different landmarks defined in $\mathbf{T}$. In this part, each image have been transformed to its high-level dataset invariable representation, *e.g*., $\mathbf{x}^i \in \mathbb{R}^F$ is a feature vector of one image belong to dataset T^i. For each landmark definition, a specific linear regressor is used to transform dataset invariable representations into dataset specific landmarks. These individual regressors naturally handle the specification of annotations definition in different datasets. Owing to the robust feature extraction before, these individual linear regressors avoid the lack of variations of a single dataset.

Learning for DA-Net. Suppose one individual regressor $\mathbf{w}^i = [\mathbf{w}_1^i, \mathbf{w}_2^i, \cdots, \mathbf{w}_{l_i}^i]$ corresponding to T_i be a weight matrix, $\mathbf{w}^i \in \mathbb{R}^{F \times l_i}$, where each column vector corresponds to the parameters of a single coordinate value. For example, $\mathbf{w}_1^i \in \mathbb{R}^F$ indicates the parameter vector for the x-coordinary of the first landmark for one image in T_i. Given a total of N datasets, and for each dataset T^i, there are M^i corresponding images. Training data are denoted as $(I_j^i, \mathbf{s}_j^i)$, with $I_j^i \in T^i$ and $\mathbf{s}_j^i \in \mathbb{R}^{l^i}$. The goal of the *DA-Net* is to minimize

$$\arg\min_{\theta, \mathbf{w}^i} \sum_{i=1}^{N} \sum_{j=1}^{M^i} \lambda_i \|\mathbf{s}_j^i - {\mathbf{w}^i}^\top \Phi(I_j^i, \theta)\| \qquad (1)$$

where λ_i denotes the importance coefficient of i-th datsets error, and $\Phi(I_j^i, \theta)$ represents the feature vector of one sample $I_j^i \in T^i$. θ represents the parameters of the Feature Extraction part of *DA-Net*, while $\mathbf{w}^i$ denotes the linear function maps features to dataset specific landmarks. These two set of parameters are optimized together in Eq.1. Note that under this scheme, θ can learn the shared feature space and $\mathbf{w}^i$ can learn the landmarks mappings of different landmark definitions, and loss function gradients of images from different datasets are propagated back together to refine θ and the shared feature space.

The objective function in Eq.1 is optimized using stochastic gradient descent with the standard backpropagation. The batch size is set 128 and is composed of equal number of images from N training datasets $\mathbf{T}$.

3.2. Candidate-Decision Network

Overview of CD-Net. Sometimes one dataset T will have severe imbalance distribution of facial characteristics. The most common scenario is the bias of the yaw angle of head pose. In general, taking such dataset with extremely different yaw angles directly to train one model will lead to poor performance. It is mainly caused by the lack of data variations in a specific angle of view, *e.g*., left view or right view. One straightforward solution is to augment raw training set with mirrored images to fully utilize all of

the data variations. However, even though a balance dataset is constructed with the original of additional mirrored images. Training a model directly with these augmented data still suffers from the relationship inconsistency inside the dataset. Take Menpo-Profile [43] dataset as an example, for the faces in left view, the semantic point of tip of nose is always in the left of corner of the mouth, while it is totally opposite for the faces in right view. This inconsistency will seriously degrade the learning of relationships between semantic points and further lead to an unsatisfactory performance.

We proposed Dataset Standardization, together with a Candidate Decision mechanism to solve this problem. Data Standardization simply splits all faces into left and right view, then flips all images from right view to get one united dataset with only one view, *e.g.*, left view. This dataset will be used to train model without suffering from the relationship inconsistency problem. Specifically, we firstly split T into T_{left} and T_{right} according to the yaw angle. Then, the T_{left} and $T_{right\_flipped}$, *e.g.*, the flipped image set of T_{right} are merged to construct a new dataset with a single view of faces. Correspondingly, the ground truth landmarks $\mathbf{s}^{'} \in \mathbb{R}^{l \times M}$ is constructed with $\mathbf{s}_{left}$ and $\mathbf{s}_{right\_flipped}$. Thus, training data can be denoted as $(I_j^{'}, \mathbf{s}_j^{'})$, with $I_j^{'} \in T^{'}$ and $\mathbf{s}_j^{'} \in \mathbb{R}^l$. Moreover, $T^{'}$ can also be trained across dataset using *DA-Net*. One evident problem is that right view faces in testset necessarily perform poor with no training data in right view angles. We further introduce *Candidate-Decision Network (CD-Net)* to solve this problem. Specifically, as shown in the bottom of Fig.2, for each image in testset, take it and its flipped version to a *DA-Net* trained on $T^{'}$. Two candidate predictions will be generated. Then, *CD-Net* should give a higher score to the candidate that is closer to its corresponding ground truth.

Learning for CD-Net. To construct training data for *CD-Net*. We firstly do a flip on the j-th image $I_j \in T$ to generate $I_{j\_flipped}$. Also, the corresponding flipped version of the ground truth landmarks $\mathbf{s}_j$ are denoted as $\mathbf{s}_{j\_flipped}$. The ground truth label $c_j \in \{0, 1\}$ of I_j can be denoted as

$$c_j = \mathbf{1}\{\|\mathbf{s}_j - \mathbf{p}_j\| < \|\mathbf{s}_{j\_flipped} - \mathbf{p}_{j\_flipped}\|\} \quad (2)$$

where $\mathbf{1}$ in Eq.2 is a indicator function. Given model parameters, *e.g.*, θ and $\mathbf{w}$, trained on $T^{'}$. $\mathbf{p}_j$ and $\mathbf{p}_{j\_flipped}$, which represent the prediction landmarks corresponding to I_j and $I_{j\_flipped}$ respectively, can be doneted as $\mathbf{p}_j = \mathbf{w}^\top \Phi(I_j; \theta)$ and $\mathbf{p}_{j\_flipped} = \mathbf{w}^\top \Phi(I_{j\_flipped}; \theta)$. Note that, as shown in Fig.2, if $c_j = 1$, the corresponding prediction landmarks $\mathbf{p}_{j\_flipped}$ will be further flipped to get the final prediction.

Now that training data for *CD-Net* can be denoted as (I_j, c_j), with $I_j \in T$ and $c_j \in \{0, 1\}$, where $j = \{1, 2, \cdots, M\}$. It is reasonable to employ the cross-entropy as the loss functions for the optimization of *CD-Net*. Therefore, the objective function can be written as

$$\arg\min_{\theta^c, \mathbf{w}^c} \sum_{j=1}^{M} -log(p(c_j|\mathbf{x}_j; \mathbf{w}^c)) \quad (3)$$

where $\mathbf{x}_j \in \mathbb{R}^F$ represent feature vector of image I^i. It can be denoted as $\mathbf{x}_j = \Phi(I_j; \theta^c)$. θ^c and $\mathbf{w}^c$ represent parameters in *Feature Extraction* and *Regression* parts of *CD-Net* respectively. We use a softmax function $p(c_j|\mathbf{x}_j; \mathbf{w}^c)$ to model the class posterior probability, which can be denoted as

$$p(c_j = m|\mathbf{x}_j; \mathbf{w}^c) = \frac{exp\{{\mathbf{w}_m^c}^\top \mathbf{x}_j\}}{\sum_{t \in \{0,1\}} exp\{{\mathbf{w}_t^c}^\top \mathbf{x}_j\}} \quad (4)$$

where $\mathbf{w}_t^c$ denotes the t-th column of the weight matrix $\mathbf{w}^c$. Following the optimization scheme of *DA-Net*, objective function in Eq.3 is optimized using SGD with the standard backpropagation. A batch-mode learning method with a batch size of 128 is also used.

4. Experiments

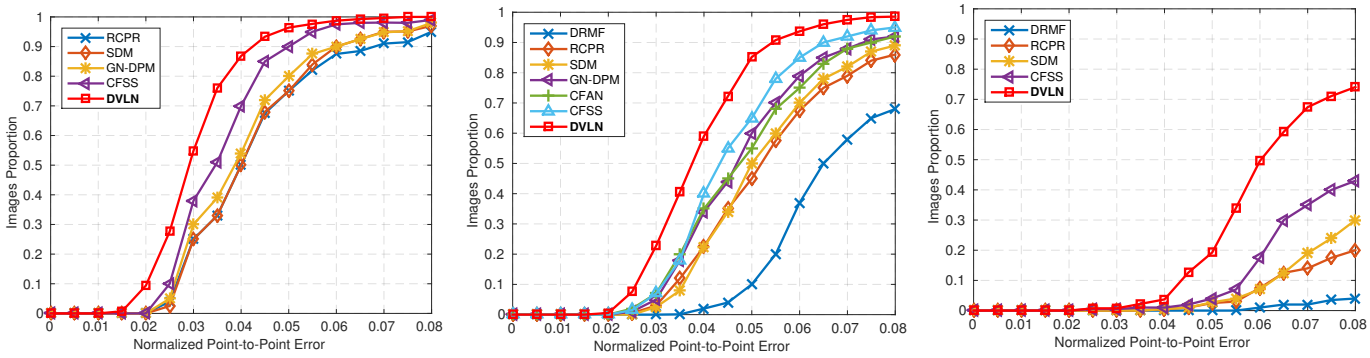
Datesets To facilitate the training of *DA-Net* and *CD-Net*, we construct two new datasets annotating facial landmarks on faces randomly collected in-the-wild. One set is *Semifrontal Facial Landmarks (SFL)* using a 106 landmarks mark-up and another is *Profile Facial Landmarks (PFL)* annotated using a 39 landmarks mark-up. Even though PFL has the same numbers of points as Menpo-Profile [43] has, the definitions of some points are actually different. Evaluations are performed on the two well-known benchmark datasets. These datasets are challenging due to images with large head pose, occlusions, and illumination variations.

300-W [27] dataset: 300-W [27] is short for 300 Faces in-the-Wild and is currently the most widely-used in-the-wild dataset. It is created from existing datasets, including LFPW [3], AFW [25], Helen [21], and a new dataset called IBUG [28] and where each image was re-annotated in a consistent manner using the 68-point landmark configuration of CMU Multi-PIE [13]. Commonly, these annotations are split into the following subsets: *(i)* the training set (3148 images) consisting of AFW [25] (337), LFPW [3] training set (811) and HELEN [21] training set (2000) *(ii)* the common subset (554 images) of HELEN [21] testing set (330) and LFPW [3] testing set (224) *(iii)* the challenging subset (135 images) named IBUG [28] and *(iv)* the full set (689 images) consisting of both the common (554) and challenging subsets (135).

Menpo [43] dataset: The Menpo dataset is a very recently introduced dataset containing landmark annotations for 8978 faces from FDDB [17] and AFLW [19]. This

Model Name	Network	Training Set	# of Training Samples	Testing Set	# of Testing Samples	Point	Normalising Factor
Eva-DA-Base	DA-Net-A	300W train-set	3,148	300W full-set	689	68	inter-pupil distance
Eva-DA	DA-Net-A	300W train-set SFL	>10K	300W full-set	689	68	inter-pupil distance
Eva-CD-Base	DA-Net-A	Menpo-Pr-Train	1,840	Menpo-Pr-Val	460	39	face size
Eva-CD	DA-Net-A CD-Net-A	Menpo-Pr-Train* Menpo-Pr-Train$^{\Delta}$	1,840	Menpo-Pr-Val	460	39	face size
Test-Semifrontal	DA-Net-B	Menpo-Fr-Train SFL	>10K	Menpo-Fr-Test	5335	68	face size
Test-Profile	DA-Net-B CD-Net-B	Menpo-Pr-Train PFL	10K	Menpo-Pr-Test	1946	39	face size

Table 1: Detailed evaluation/test settings for our experiments. Specifically, model Eva-DA-Base in Sec.4.1 gives a baseline of a *DA-Net* and make comparison with state-of-the-art methods. Models named Eva-DA-Base and Eva-DA evaluate the effectiveness of our proposed *DA-Net* (more detail is described in Sec.4.2). Meanwhile, models named Eva-CD-Base and Eva-CD evaluate the effectiveness of our proposed *CD-Net* (more detail is described in Sec.4.3). Finally, Test-Semifrontal and Test-Profile give the test results of the 2nd Facial Landmark Localisation Competition. Note that Menpo-Pr-Train* is different from Menpo-Pr-Train. Menpo-Pr-Train is the raw training set, while Menpo-Pr-Train* is the standardized version with only left view faces. Menpo-Pr-Train$^{\Delta}$ is generated by a trained *DA-Net-A* and used to train a *CD-Net-A* in model Eva-CD.



(a) CED for 49-pts Common Subset. (b) CED for 68-pts Common Subset. (c) CED for 68-pts Challenging Subset.

Figure 3: Comparisons of cumulative errors distribution (CED) curves on 300-W [27] Dataset. The proposed method outperforms various state-of-the-art methods.

dataset consists of two categories of faces: *(i)* Menpo-Fr (6679 semifrontal faces) with regards to the same mark-up used in the 300-W [27] competition (a total of 68 landmarks) and *(ii)* Menpo-Pr (2300 profile faces) using a 39 landmarks mark-up. For Menpo-Fr, we randomly sample 1,200 of 6,679 images to construct a validation set, *e.g.*, Menpo-Fr-Val. The corresponding training set is Menpo-Fr-Train with 5,479 images. Menpo-Pr-Val (460) and Menpo-Pr-Train (1,840) are constructed in the same way.

Evaluation metric Following most previous works, we evaluate the alignment accuracy using the standard normalised landmarks mean error. Note that the difference of normalising factor used between different experiments. All evaluation settings for different experiments are noted in Table.1. Cumulative Error Distribution (CED) is adopted for different evaluation schemes in the literature. We should point out that inherent difficulty exits in fitting the boundary points of the face contour for semifrontal faces with 68-points landmarks. Thus, an advisable evaluation is also taken in 49-points landmarks, *e.g.*, the facial feature landmarks.

Implementation In the experiments, two hyper-

parameters settings are used for the network topology. For *DA-Net-A*, we modified a VGG-16 [32] with the purpose of enabling it to adapt to the scenario of face alignment. Note that the network structure design is not the main task of this work and other structures, *e.g*., ResNet-18 [15], also show similar phenomenon in our experiments. Fully connected layers are used to produce a feature vector for the final linear projection. For *DA-Net-B*, we gain further performance advance by adding more convolutional layers and enlarging input image size. *DA-Net-B* is only used for 2nd Facial Landmark Localisation Competition in Sec.4.4. Note that in our experiments, the setting of *CD-Net* always keep consistent with its coupled *DA-Net*. We train the network using Nesterov SGD with a mini-batch size of 128. We choose the best model according to our validation set (300W full-set).

4.1. Comparison with State-of-the-art Methods

We compare our approach against state-of-the-art methods. In order to build a baseline of *DA-Net*, we train a DA-Net-A on 300-W [27] training set which consists of 3,148 face images without cross dataset training to get the model Eva-DA-Base. We also augment the training set by affine transformation.

The evaluation is performed on the 300-W [27] dataset, which incudes Common Subset (554), Challenging Subset (135) and Fullset (689). The mean error results normalised by the inter-pupil distance are listed in Table.2. From the table, we can observe that our model outperforms the best ever reported regression-based method, *e.g*., CFSS [49] and the best ever reported CNN-based method, *e.g*., RAR [39] by a large margin. As shown in Fig.3, we also plot the CED curves for various methods in order to compare with literatures reporting CED performance. Again, our proposed method achieves the best performance.

4.2. Evaluation of the Effectiveness of Cross Dataset Alignment

In this experiment, we wish to verify the effectiveness of cross dataset alignment method using *DA-Net* architecture, by evaluating the accuracy of annotations on the target testset, *e.g*., 300W full-set. To compare with Eva-DA-Base which is trained only on 3,148 images in 300-W [27], we perform a cross dataset training together on 300-W [27] training set (68pt) and SFL (106pt). Note that we do not need to re-annotate each of the datasets, *e.g*., re-annotate SFL using a 68 landmarks mark-up. All we need to do is feeding batches consist of images from different datasets in *DA-Net* and training together.

It is evident from Table.2 that optimizing landmark detection under dataset-across mechanism are better. In particular, Eva-DA outperforms Eva-DA-Base much more obvious on the Chanllenging Subset with large pose variations and severe partial occlusions. This result illustrates the proposed *DA-Net* architecture can leverage different variations in different datasets to greatly improve the generalization and leading a much more robust model. Further experiments show that the accuracy of the validation set of SFL is also improved, which indicate different datasets can benefit from each other by implicit feature sharing and dataset-specific landmarks regression.

Method	Commont Subset	Challenging Subset	Fullset
DRMF [2]	6.65	19.79	9.22
RCPR [6]	6.18	17.26	8.35
CFAN [45]	5.50	16.78	7.69
ESR [8]	5.28	17.00	7.58
SDM [41]	5.57	15.40	7.50
LBF [26]	4.95	11.98	6.32
CFSS [49]	4.73	9.98	5.76
TCDCN [47]	4.80	8.60	5.54
RAR [39]	4.12	8.35	4.94
DVLN (Eva-DA-Base)	3.94	7.62	4.66
DVLN (Eva-DA)	**3.79**	**7.15**	**4.45**

Table 2: Mean Errors (Percent) on 300-W [27] Dataset (68 Landmarks)

4.3. Evaluation of the Effectiveness of Dataset Standardization and Candidate Decision

To examine the influence of Dateset Standardization and Candidate Decision mechanism in our proposed approach, we do experiments on Menpo-Pr [43] on account of its extreme variations in yaw angle of head pose. A baseline model Eva-CD-Base is trained directly on Menpo-Pr-Train with *DA-Net-A* network. Note that we also augment Menpo-Pr-Train with their mirrored images. For dateset standardization, Menpo-Pr-Train is split into two subsets based on head pose, *e.g*., Menpo-Pr-Train-Left and Menpo-Pr-Train-Right. Menpo-Pr-Train-Left and the flipped version of Menpo-Pr-Train-Right are merged to construct single Menpo-Pr-Train*. Note that we always flipped the annotations together with images. A *DA-Net-A* network is then trained on Menpo-Pr-Train*. Next, for realization of Candidate Decision mechanism, a *CD-Net-A* is trained. As mentioned in Sec.3.2, to generate the training set, each image I_j on Menpo-Pr-Train will pass the trained *DA-Net-A* and generate the groundtruth class c_j according to the errors of its two candidate predictions. Then, Menpo-Pr-Train$^\Delta$, which consists of several pairs of image and its generated class c_j is used to train a *CD-Net-A*. For a test image, the original image and its flipped version will be given to *DA-Net-A* to generate two plausible hypotheses for landmarks and *CD-*

Net-A will choose one of them as final output.

The evaluation is performed on the Menpo-Pr-Val (the validation set of Menpo-Pr). The comparison result is illustrated in Table.3 and Fig.4. It is obvious that Dataset Standardization and Candidate Decision mechanism which concentrate different facial variations to a smaller distribution space can significantly alleviate over-fitting problem caused by lack of data. Meanwhile, the performance degradation caused by relationship inconsistency is automatically alleviated.

Model Name	Mean Error
Eva-CD-Base	0.0542
Eva-CD	**0.0327**

Table 3: Mean Errors on Menpo-Pr-Val

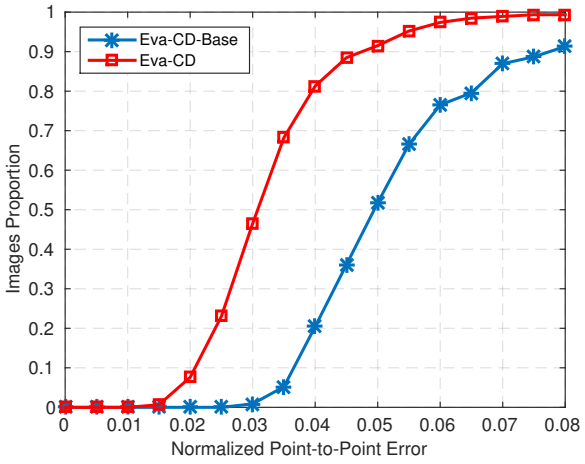
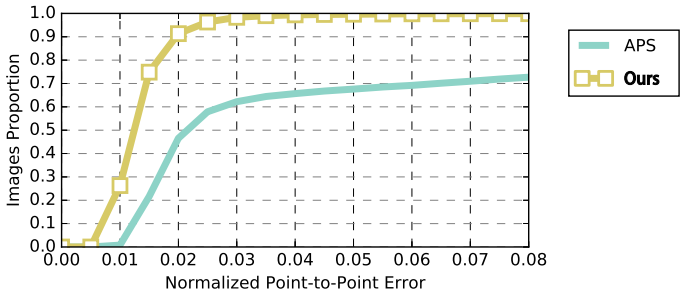


Figure 4: CED for Menpo-Pr-Val

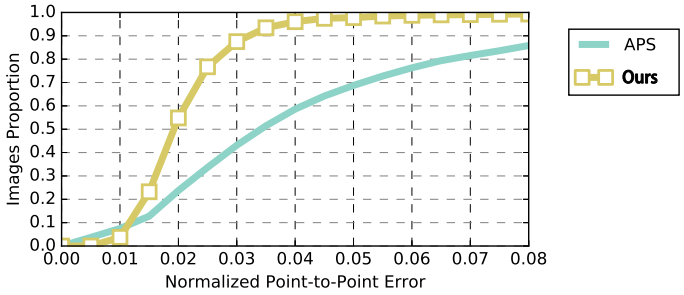
4.4. Results on the 2nd Facial Landmark Localisation Competition

In this section, we report the result of our submissions on the 2nd Facial Landmark Localisation Competition [43]. All of the test faces are firstly detected by a face detector. Then, these face images are converted to gray-scale and cropped with the input of our models. For the Semifrontal Challenge, model Test-Semifrontal is trained on Menpo-Fr-Train and SFL under the dataset-across mechanism with a *DA-Net-B*. For the Profile Challenge, model Test-Profile is trained on Menpo-Pr-Train and PFL under the Dataset Standardization and Candidate Decision mechanism. Note that model Test-Profile also benefits from dataset-across training together with Menpo-Pr-Train and PFL.

The returned results are shown in Fig.5. The baseline is Active Pictorial Structures (APS) [1] which combines the main ideas behind Pictorial Structures and Active Appearance Models. Based on the individual errors provided by the organizers, we also calculate the mean errors. For Semifrontal and Profile Challenge, the mean errors, which normalised by the bounding box size are 0.013 and 0.022 respectively.



(a) CED for Semifrontal Challenge



(b) CED for Profile Challenge

Figure 5: Results on the 2nd Facial Landmark Localisation Competition.

5. Conclusion

In this paper, we have presented a novel *Deep Variation Leveraging Network (DVLN)* to take advantage of rich data variations among and within different datasets. The proposed network uniquely combined two sub-networks, *e.g.*, *Dataset-Across Network (DA-Net)* and *Candidate-Decision Network (CD-Net)* to work in a strong coupling schema. In *DA-Net*, data variations can be shared among different datasets without any re-annotation by one implicit feature sharing mechanism, while data variations within dataset can be further utilized with a coupled *CD-Net*. Thanks to learning with diverse variations, the proposed model is much more robust compared to existing methods. Future work will concentrate on exploring the relationship representations of datasets to further improve the accuracy and robustness of the proposed model.

6. Acknowledgement

We thank the organisers of the 2nd Facial Landmark Localisation Competition for providing data and evaluating our submission.

References

[1] E. Antonakos, J. Alabort-i-Medina, and S. Zafeiriou. Active pictorial structures. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 5435–5444, 2015.

[2] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic. Robust discriminative response map fitting with constrained local models. In *2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, June 23-28, 2013*, pages 3444–3451, 2013.

[3] P. N. Belhumeur, D. W. Jacobs, D. Kriegman, and N. Kumar. Localizing parts of faces using a consensus of exemplars. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 545–552, 2011.

[4] 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). *CoRR*, abs/1703.07332, 2017.

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

[6] X. P. Burgos-Artizzu, P. Perona, and P. Dollár. Robust face landmark estimation under occlusion. In *IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, December 1-8, 2013*, pages 1513–1520, 2013.

[7] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou. Facewarehouse: A 3d facial expression database for visual computing. *IEEE Trans. Vis. Comput. Graph.*, 20(3):413–425, 2014.

[8] 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.

[9] C. Chen, A. Dantcheva, and A. Ross. Automatic facial makeup detection with application in face recognition. In *International Conference on Biometrics, ICB 2013, 4-7 June, 2013, Madrid, Spain*, pages 1–8, 2013.

[10] D. Chen, S. Ren, Y. Wei, X. Cao, and J. Sun. Joint cascade face detection and alignment. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI*, pages 109–122, 2014.

[11] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(6):681–685, 2001.

[12] T. F. Cootes, M. C. Ionita, C. Lindner, and P. Sauer. Robust and accurate shape model fitting using random forest regression voting. In *European Conference on Computer Vision*, pages 278–291, 2012.

[13] R. Gross, I. A. Matthews, J. F. Cohn, T. Kanade, and S. Baker. Multi-pie. *Image Vision Comput.*, 28(5):807–813, 2010.

[14] T. Hassner, S. Harel, E. Paz, and R. Enbar. Effective face frontalization in unconstrained images. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 4295–4304, 2015.

[15] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778, 2016.

[16] D. C. Hogg and R. Boyle, editors. *Proceedings of the British Machine Vision Conference, BMVC 1992, Leeds, UK, September, 1992*. BMVA Press, 1992.

[17] V. Jain and E. G. Learned-Miller. Fddb: A benchmark for face detection in unconstrained settings. In *UMass Amherst Technical Report*, 2010.

[18] F. Kahraman, M. Gökmen, S. Darkner, and R. Larsen. An active illumination and appearance (AIA) model for face alignment. In *2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2007), 18-23 June 2007, Minneapolis, Minnesota, USA*, 2007.

[19] M. Köstinger, 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 *IEEE International Conference on Computer Vision Workshops, ICCV 2011 Workshops, Barcelona, Spain, November 6-13, 2011*, pages 2144–2151, 2011.

[20] N. Kumar, P. Belhumeur, and S. Nayar. Facetracer: A search engine for large collections of images with faces. In *European Conference on Computer Vision*, pages 340–353, 2008.

[21] 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, 2012.

[22] I. A. Matthews and S. Baker. Active appearance models revisited. *International Journal of Computer Vision*, 60(2):135–164, 2004.

[23] K. Messer, J. Matas, J. Kittler, and K. Jonsson. Xm2vtsdb: The extended m2vts database. In *In Second International Conference on Audio and Video-based Biometric Person Authentication*, pages 72–77, 1999.

[24] S. Milborrow and F. Nicolls. Locating facial features with an extended active shape model. In *European Conference on Computer Vision*, pages 504–513, 2008.

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

[26] S. Ren, X. Cao, Y. Wei, and J. Sun. Face alignment at 3000 FPS via regressing local binary features. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pages 1685–1692, 2014.

[27] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *2013 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 *IEEE Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2013, Portland, OR, USA, June 23-28, 2013*, pages 896–903, 2013.

[29] J. M. Saragih and R. Göcke. A nonlinear discriminative approach to AAM fitting. In *IEEE 11th International Conference on Computer Vision, ICCV 2007, Rio de Janeiro, Brazil, October 14-20, 2007*, pages 1–8, 2007.

[30] J. M. Saragih, S. Lucey, and J. F. Cohn. Deformable model fitting by regularized landmark mean-shift. *International Journal of Computer Vision*, 91(2):200–215, 2011.

[31] B. Shi, X. Bai, W. Liu, and J. Wang. Deep regression for face alignment. *CoRR*, abs/1409.5230, 2014.

[32] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[33] B. M. Smith and L. Zhang. Collaborative facial landmark localization for transferring annotations across datasets. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI*, pages 78–93, 2014.

[34] Y. Sun, X. Wang, and X. Tang. Deep convolutional network cascade for facial point detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3476–3483, 2013.

[35] Y. Sun, X. Wang, and X. Tang. Deep learning face representation from predicting 10, 000 classes. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pages 1891–1898, 2014.

[36] Y. Sun, X. Wang, and X. Tang. Hybrid deep learning for face verification. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(10):1997–2009, 2016.

[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 *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 4177–4187, 2016.

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

[39] S. Xiao, J. Feng, J. Xing, H. Lai, S. Yan, and A. A. Kassim. Robust facial landmark detection via recurrent attentive-refinement networks. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I*, pages 57–72, 2016.

[40] X. Xiong and F. D. la Torre. Global supervised descent method. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 2664–2673, 2015.

[41] X. Xiong and F. D. L. Torre. Supervised descent method and its applications to face alignment. In *Computer Vision and Pattern Recognition*, pages 532–539, 2013.

[42] H. Yang, W. Mou, Y. Zhang, I. Patras, H. Gunes, and P. Robinson. Face alignment assisted by head pose estimation. In *Proceedings of the British Machine Vision Conference 2015, BMVC 2015, Swansea, UK, September 7-10, 2015*, pages 130.1–130.13, 2015.

[43] S. Zafeiriou, G. Trigeorgis, G. Chrysos, J. Deng, and J. Shen. The menpo facial landmark localisation challenge: A step closer to the solution. In *CVPRW*, 2017.

[44] J. Zhang, M. Kan, S. Shan, and X. Chen. Leveraging datasets with varying annotations for face alignment via deep regression network. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 3801–3809, 2015.

[45] 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*, 2014.

[46] Z. Zhang, P. Luo, C. L. Chen, and X. Tang. Facial landmark detection by deep multi-task learning. In *European Conference on Computer Vision*, pages 94–108, 2014.

[47] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. Learning deep representation for face alignment with auxiliary attributes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(5):918–930, 2016.

[48] S. Zhu, C. Li, C. C. Loy, and X. Tang. Transferring landmark annotations for cross-dataset face alignment. *CoRR*, abs/1409.0602, 2014.

[49] S. Zhu, C. Li, C. C. Loy, and X. Tang. Face alignment by coarse-to-fine shape searching. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 4998–5006, 2015.

[50] Z. Zhu, P. Luo, X. Wang, and X. Tang. Deep learning identity-preserving face space. In *IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, December 1-8, 2013*, pages 113–120, 2013.