Facial Landmarks Localization using Cascaded Neural Networks

Shahar Mahpod[a], Rig Das[b,**], Emanuele Maiorana[b], Yosi Keller[a], Patrizio Campisi[b]

[a]*S. Mahpod & Y. Keller are with the Faculty of Engineering, Bar-Ilan University, Ramat Gan 52900, Israel (e-mail: mahpods@biu.ac.il, yosi.keller@gmail.com)*
[b]*R. Das, E. Maiorana & P. Campisi are with the Section of Applied Electronics, Department of Engineering, Roma Tre University, Via Vito Volterra 62, 00146, Rome, Italy (e-mail: {rig.das, emanuele.maiorana, patrizio.campisi}@uniroma3.it)*

ABSTRACT

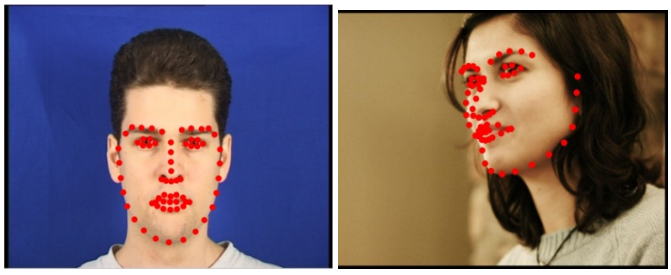
The accurate localization of facial landmarks is at the core of face analysis tasks, such as face recognition and facial expression analysis, to name a few. In this work, we propose a novel localization approach based on a deep learning architecture that utilizes cascaded subnetworks with convolutional neural network units. The cascaded units of the first subnetwork estimate heatmap-based encodings of the landmarks' locations, while the cascaded units of the second subnetwork receive as input the output of the corresponding heatmap estimation units, and refine them through regression. The proposed scheme is experimentally shown to compare favorably with contemporary state-of-the-art schemes, especially when applied to images depicting challenging localization conditions.

1. Introduction

The localization of facial landmark points, such as eyebrows, eyes, nose, mouth, and jawline, is one of the core computational components in visual face analysis, with applications in face recognition (Huang et al., 2013), kinship verification (Mahpod and Keller, 2018), and facial attribute inference (Kumar et al., 2008), to cite a few. Robust and accurate localization entails several difficulties, due to variations in face pose, illumination, and resolution, as well as to occlusions, as depicted in Figure 1.

Traditional approaches for facial landmark localization have relied on appearance models, providing parametric or non-parametric descriptors of a face shape. In this context, fitting strategies have been defined to minimize the residual error between the training face images and their synthesized model (Cootes and Taylor, 1992).

Regression-based approaches have been then successfully proposed, showing improved accuracy compared to their predecessors, especially when applied to in-the-wild face images (Cao et al., 2012). Starting from an initial estimate of the landmarks' positions, typically obtained computing local image features from an average face template, a high-dimensional regression model is iteratively estimated. Besides achieving high localization accuracy, these schemes are also computationally efficient, commonly requiring less than 1ms processing time per frame (Ren et al., 2014). Yet, given that they rely on an initial estimate of the landmarks' positions, they are in general limited to yaw, pitch, and head roll angles of less than 30°, and are thus susceptible to initialization and convergence issues. Contemporary approaches might therefore underperform when employed in challenging conditions (Shao et al., 2017).



(a) (b)

Fig. 1. Facial landmark localization. Each image feature, marked by a point, is considered a particular landmark, and is localized individually. (a) A frontal face image from the XM2VTS datasets (Messer et al., 2003). (b) An image from the Helen dataset (Le et al., 2012) with non-frontal pose and expression variation, making the localization challenging.

Further improvements in facial landmarks localization have been achieved with the exploitation of deep-learning-based approaches. In particular, a notable innovation allowed by con-

**Please cite this work as: S. Mahpod, R. Das, E. Maiorana, Y. Keller, P. Campisi, "Facial Landmarks Localization using Cascaded Neural Networks," Computer Vision and Image Understanding, Vol. 205, April 2021. Digital Object Identifier 10.1016/j.cviu.2021.103171

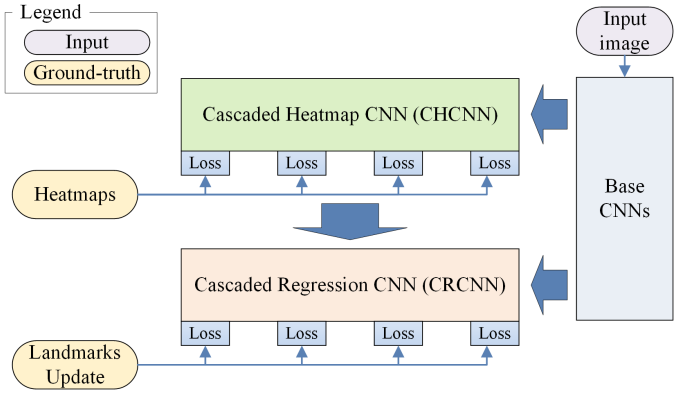


Fig. 2. Outline of the proposed CCNN framework. The CCNN consists of base CNNs, preceding the cascaded heatmap subnetwork (CHCNN) estimating the heatmaps, and the cascaded regression CNN (CRCNN) refining the heatmaps localization via point-wise regression.

volutional neural networks (CNNs) has consisted in the amelioration of heatmap-based techniques, following the seminal work of Pfister et al. (2015). Heatmaps are general-purpose descriptors which can be employed to represent sets of points. Typically, they can be obtained by applying smoothing filters, such as diffusion kernels, to characterize a point depending on the geometry of its surroundings (Coifman and Lafon, 2006). Although generating heatmaps through CNNs has guaranteed state-of-the art results in terms of robustness, the achievable localization accuracy is inherently limited, due to the coarse spatial resolution of the created heatmaps, typically much lower than the one of the original image.

In this work we exploit the upsides of the most commonly employed regression- and heatmap-based approaches, by proposing a novel deep-learning-based framework for facial landmark localization, formulated as a cascaded CNN (CCNN) comprising two paired cascaded heatmap and regression subnetworks. An outline of the architecture of our proposed CCNN is depicted in Figure 2. In more detail, the cascaded heatmap subnetwork (CHCNN) consists of multiple successive heatmap-based localization units, which perform a *robust* and *coarse* localization. The following cascaded regression CNN (CR-CNN) subnetwork *refines* the heatmap-based localization performing a coarse-to-fine estimate. Cascaded architectures have been employed in the proposed subnetwork due to their proven ability in improving the localization accuracy of regression- and heatmap-based schemes. The cascaded layers in both the CHCNN and CRCNN are non-weight-sharing, allowing each to separately learn a different localization range. The proposed CCNN is experimentally shown to compare favourably with contemporary state-of-the-art face localization schemes. Although this work exemplifies the use of the proposed approach in the localization of facial landmarks, it is of general applicability, and can be used for any class of objects.

Thus, the contributions of this work are as follows:

- we derive a CNN-based face localization scheme using a coarse and robust heatmap estimate, followed by a subsequent regression-based refinement;
- the heatmap estimation and regression tasks are performed by cascaded subnetworks, that allow an iterative refinement of the localization accuracy. To the best of our knowledge, this is the first such formulation of the face localization problem;
- the proposed CCNN framework is experimentally shown to outperform contemporary state-of-the-art approaches.

A review of the major contributions in literature regarding facial landmarks localization is provided in Section 2. The proposed CCNN architecture is then detailed in Section 3, and its effectiveness is outlined through the experimental tests discussed in Section 4. Conclusions are finally drawn in Section 5.

2. State-of-the-Art: Facial Landmark Point Localization

The localization of facial landmark points is a fundamental computer vision task that has been studied in a multitude of works, dating back to the seminal works on active shape models (ASMs) (Cootes and Taylor, 1992), active appearance models (AAMs) (Cootes et al., 2001), and constrained local models (CLMs) (Cristinacce and Cootes, 2006), which have paved the way for recent localization schemes. Classical face localization schemes utilize either parametric (Cristinacce and Cootes, 2006) or non-parametric (Belhumeur et al., 2013) models to learn the statistical distribution of face landmark points, in order to provide the actual position of the interested locations in the treated images, trying to deal with significant appearance and pose variations.

More recently, state-of-the art results on facial landmarks localization have been achieved by resorting to regression- and heatmap-based methods, and exploiting deep learning strategies to learn high-level facial features. The following subsections provide an overview of the most relevant works resorting to such approaches.

2.1. Cascaded Shape Regression Schemes

Cascaded Shape Regression (CSR) (Trigeorgis et al., 2016) schemes localize landmark points through an iterative process, where regression estimates are progressively refined using local image features, computed at the selected landmarks' locations, as inputs. Such schemes are commonly initialized with an estimate of the landmarks based on an average face template, and a bounding box of the face provided by a detector such as Viola-Jones (Viola and Jones, 2001). CSR-based approaches have been shown to be computationally efficient by applying fast regression cascades, yet their convergence, and the resulting localization accuracy, might be susceptible to an inaccurate initialization. Actually, as the common initialization of the face landmarks corresponds to frontal head poses, CSR schemes are typically limited to yaw, pitch, and head roll angles of less than 30°.

Notable examples of employed local features include scale-invariant feature transform (SIFT) characteristics (Lowe, 2004), used in the supervised descent method (SDM) proposed by Xiong and De la Torre (2013). Local binary features (LBFs) have been instead employed to estimate facial landmarks' locations through regression trees in Ren et al. (2014), with a low

required computational complexity allowing to process images at 3000 fps. Computationally efficiency has been also sought in Chen et al. (2014), where random forests regression has been applied to Haar-like local image features.

An explicit shape regression (ESR) is performed in Cao et al. (2012), where a vectorial regression function inferring the whole set of facial landmarks is directly learned from the input image, designing a two-level boosted regression using shape-indexed and correlation-based features. A unified model for face detection, pose estimation, and landmark localization has been instead suggested in Zhu and Ramanan (2012). The facial features and their geometrical relations are there encoded by the vertices of a graph, while the regression inference is obtained resorting to a mixture of trees, learned using a training set, with a shared pool of parts. An iterative coarse-to-fine shape searching (CFSS) refinement method has been introduced in Zhu et al. (2015), where the initial coarse solution allows to constrain the search space of the finer shapes, thus allowing to avoid suboptimal local minima, and improving the estimation of large pose variations.

2.2. Deep Learning-based Schemes

Deep learning techniques have been applied to face landmark localization in Zhang et al. (2016), where a multi-task estimation of facial attributes, such as gender, expression, and appearance attributes has been performed using a task-constrained deep convolutional network (TCDCN), guaranteeing robust and accurate estimates. The authors of Chen et al. (2017) have instead proposed a 4-stage coarse-to-fine framework (CTFF) for landmark localization, with the interested points first coarsely predicted, and then refined by extracting multi-scale patches. An attention gate network applied to fuse all results.

A joint usage of deep learning techniques and regression-based schemes has been proposed in Trigeorgis et al. (2016) with the mnemonic descent method (MDM), where feature learned through a CNN are processed through regression, thus yielding an end-to-end trainable scheme.

A framework based on cascaded CNN regressions, which progressively refine localization, has been introduced in Xiao et al. (2016). The landmark locations are there sequentially improved at each stage, allowing the more reliable ones to be processed earlier. The proposed recurrent attentive-refinement (RAR) approach uses long short-term memory (LSTM) networks to identify reliable landmarks, and refine their localization. Recurrent networks have been also employed in Lai et al. (2017), where a deep recurrent regression (DRR) approach has been designed, by leveraging on deep shape-indexed features and recurrent shape features to learn the connections between the regressions. A sequential linear regression has been used to learn and update the shapes.

In (He et al., 2017b), the authors have combined cascaded shape regression with a CNN, applying a fully end-to-end cascaded CNN (FEC-CNN) (He et al., 2017a) as a backbone network. Differently from the SDM approach in Xiong and De la Torre (2013), which learns a cascaded linear regression using projection matrices of SIFT descriptors, the FEC-CNN approach extracts differentiable shape-indexed patches, and feeds them into the subnetworks to predict the shape residual for the next step of the regression refinement. A deep alignment network (DAN) based on a cascaded CNN, where each stage refines the landmark positions estimated by the preceding one, has been also presented in (Kowalski et al., 2017). In order to extract features from the entire face, instead of relying on local patches, heatmaps are employed at the initialization stage of the DAN method to provide visual information about landmark locations. A landmark heatmap is an image with high intensity values around the interested landmark locations, and intensity decreasing proportionally to the distance from the nearest landmark. Differently from the approaches mentioned in the following subsections, and also from our approach, the heatmpas employed in the DAN method are not estimated through CNNs, but used solely as a mean for transferring information between stages (Kowalski et al., 2017).

2.3. Heatmap-based Schemes

Exploiting multiple convolution layers and nonlinear activation functions, deep learning approaches have been also employed to deal with landmarks localization by creating heatmaps significantly different than the ones obtained through classical approaches (Coifman and Lafon, 2006), thus providing the means for achieving high robustness. Facial landmark localization and human pose estimation have been first performed resorting to heatmaps and CNNs to detect the Synovial joints (Pfister et al., 2015), with the optical flow used to fuse heatmap predictions from consecutive frames. This approach has been extended (Belagiannis and Zisserman, 2017) by deriving a cascaded heatmap estimation subnetwork, consisting of multiple heatmap regression units, where the heatmap is estimated progressively such that each heatmap regression unit is given its predecessor's output. The obtained localization estimates are characterized by a high *robustness*. However, their accuracy is inherently limited by the coarse spatial resolution of the generated heatmaps, typically in the order of $1/4^{th}$ with respect to the input image. This kind of approach is of particular interest in our work, which is based on heatmap estimation refined through a cascaded regression subnetwork.

A two-stage architecture (TSA), where heatmaps are first estimated through a basic landmark prediction stage, and then refined using a whole landmark regression stage made of a set of shape regression subnetworks, each adapted and trained for a particular pose, has been proposed in Shao et al. (2017). As the varying appearances of the face images might reduce the localization accuracy, Dong et al. (2018) have proposed a style aggregated network (SAN), that exploits a generative adversarial network (GAN) to compute an appearance-invariant face representation. This aggregates varying face appearances, such as dark and light faces images, and improves the localization accuracy. A conditional GAN (CGAN) has been instead employed by Chen et al. (2020) to induce geometric priors on the face landmarks, by introducing a discriminator that classifies real vs. erroneous ("fake") localizations.

The aforementioned approaches, representing the state of the art on facial landmark localization using either regression- or heatmap-based methods, are taken into account in Section 4 to compare the performance achievable with the proposed framework, with the results currently available in literature.

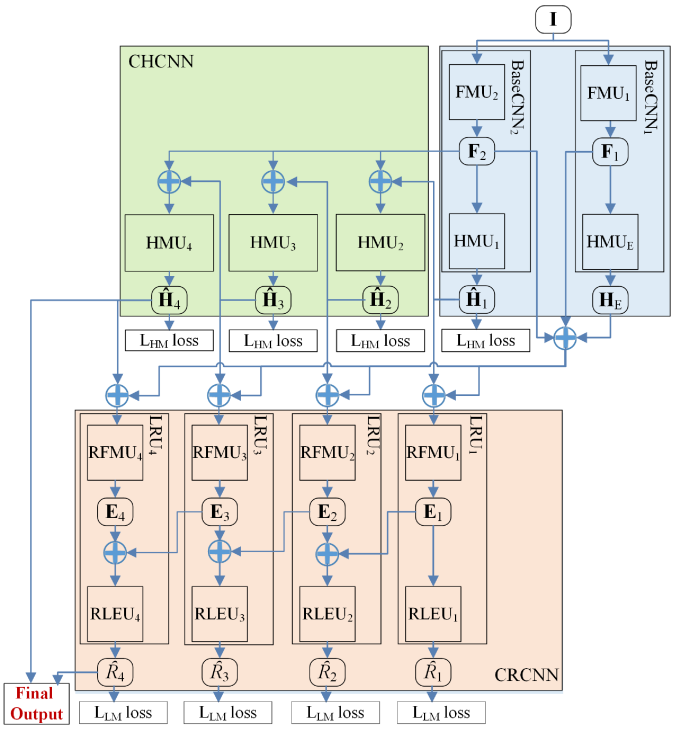


Fig. 3. Outline of the proposed CCNN localization network. The input image is encoded by two Base subnetworks, BaseCNN$_1$ and BaseCNN$_2$. Their outputs are processed by the Cascaded Heatmap CNN (CHCNN), made of heatmap estimation units (HMUs), and refined by the Cascaded Regression CNN (CRCNN), consisting of landmark regression units (LRUs).

3. Proposed Face Localization using a Cascaded CNN

A face landmarks localization task consists in searching for a set of points $P = \left\{\mathbf{p}^{(i)}\right\}_1^N$, such that $\mathbf{p}^{(i)}=[x^{(i)}, y^{(i)}]^T$, in an image $I \in \mathbb{R}^{w\times h\times 3}$, corresponding to pre-established facial fiducial points. The number N of data to be estimated relates to the used annotation convention. Figures 2 and 3 depict the general and detailed architecture of the proposed CCNN, respectively, comprising a total of three sub-networks.

The heatmap is used as a state variable initiated by the base subnetwork (Section 3.1), and is iteratively refined by applying two complementary losses:

- the heatmap-based loss, (Section 3.2), that induces the graph structure of the detected landmarks, and
- the coordinates-based representation, refined by pointwise regression (Section 3.3).

The first part of our network is a pseudo-siamese (non-weight-sharing) sub-network consisting of two subnetworks $\{BaseCNN_1, BaseCNN_2\}$, which compute the corresponding feature maps $\{\mathbf{F}_1, \mathbf{F}_2\}$ of the input image, and an initial heatmap estimate.

The second subnetwork is a cascaded heatmap CNN (CHCNN) that robustly estimates a single 2D heatmap per facial feature location. Examples of the generated heatmaps are depicted in Figure 4. The CHCNN consists of K cascaded heatmap estimation units, each estimating a 3D heatmap $\hat{\mathbf{H}}_k \in \mathbb{R}^{\frac{w}{s}\times\frac{h}{s}\times N}$, with $k = 1,\ldots,K$. The size of the obtained heatmaps depends on the scale factor s. The cascaded formulation implies that the k^{th} heatmap unit (HMU) receives as inputs the heatmap $\hat{\mathbf{H}}_{k-1}$ estimated by its predecessor, along with the feature map $\mathbf{F}_2$. The HMUs are non-weight-sharing, as each unit refines a different estimate of the heatmaps, thus learning a different regressor. Our scheme therefore differs from the heatmap-based pose estimation of Belagiannis and Zisserman (2017), which employs weight-sharing cascaded units. Each HMU is trained using the loss:

$$L_{HM} = \frac{s^2}{whN}\sum_{y,x,i=1}^{\frac{w}{s},\frac{h}{s},N}\left[\mathbf{H}_k(x,y,i) - \hat{\mathbf{H}}_k(x,y,i)\right]^2, \quad (1)$$

where $\mathbf{H}_k \in \mathbb{R}^{\frac{w}{s}\times\frac{h}{s}\times N}$, $k = 1,\ldots,K$, represent the ground-truth heatmaps, derived from the ground-truth set of points P by applying to each landmark a 2D symmetric Gaussian filter with standard deviation σ.

The landmark estimates $\hat{P}_k = \{\hat{\mathbf{p}}_k^{(i)}\}_1^N$ of each HMU are obtained as the locations of the maxima of $\hat{\mathbf{H}}_k$, that is, $\hat{\mathbf{p}}_k^{(i)} = \text{argmax}_{x,y}\,\hat{\mathbf{H}}_k(x,y,i)$. Such points, computed on a coarse grid, are then refined by the third subnetwork, i.e., the cascaded regression CNN (CRCNN). This latter consists of K cascaded landmark regression units (LRUs), with the k^{th} LRU receiving as input the output of its preceding LRU, that is, $\mathbf{E}_{k-1}$, the output $\mathbf{H}_k$ of the corresponding HMU, and the outputs of the base subnetworks, namely $\mathbf{F}_1$, $\mathbf{F}_2$, and $\mathbf{H}_E$. Each LRU applies a regression loss to improve the heatmap-based landmark estimate, by computing the refinement $\hat{R}_k$,

$$\hat{R}_k = vec\,(P) - vec\left(\hat{P}_k\right), \quad (2)$$

where $vec\,(\cdot)$ is a vectorized replica of the N points in a set. Equation 2 is optimized using the landmark loss function

$$L_{LM} = \frac{1}{2N}\sum_{i=1}^{2N}\left(R_k(i) - \hat{R}_k(i)\right)^2, \quad (3)$$

where $R_k \in \mathbb{R}^{2N}$ are the distances (separately for the horizontal and vertical directions) between the ground-truth landmarks and the estimates computed by the heatmaps.

The details of the designed subnetworks are reported in the following. All convolutional layers of the proposed CCNN are implemented with a subsequent batch normalization layer.

3.1. Base Subnetwork

The proposed Base subnetwork consists of two pseudo-siamese (non-weight-sharing) units, indicated as BaseCNN$_1$ and BaseCNN$_2$ and detailed in Table 1. The input of the network is a face image $I \in \mathbb{R}^{256\times256\times3}$, having set $w = h = 256$.

The first part of both BaseCNN$_1$ and BaseCNN$_2$, comprising layers A1-A7 from Table 1, computes feature maps from the input image using 3×3 filters, and it is indicated as feature map unit (FMU) in Table 1 and Figure 3. The size of the produced feature maps is the same as that of the employed heatmaps, and it is set by using $s = 4$ in the designed architecture. Following the vast majority of contemporary works on facial landmark localization, and in order to adhere to the 300-W competition

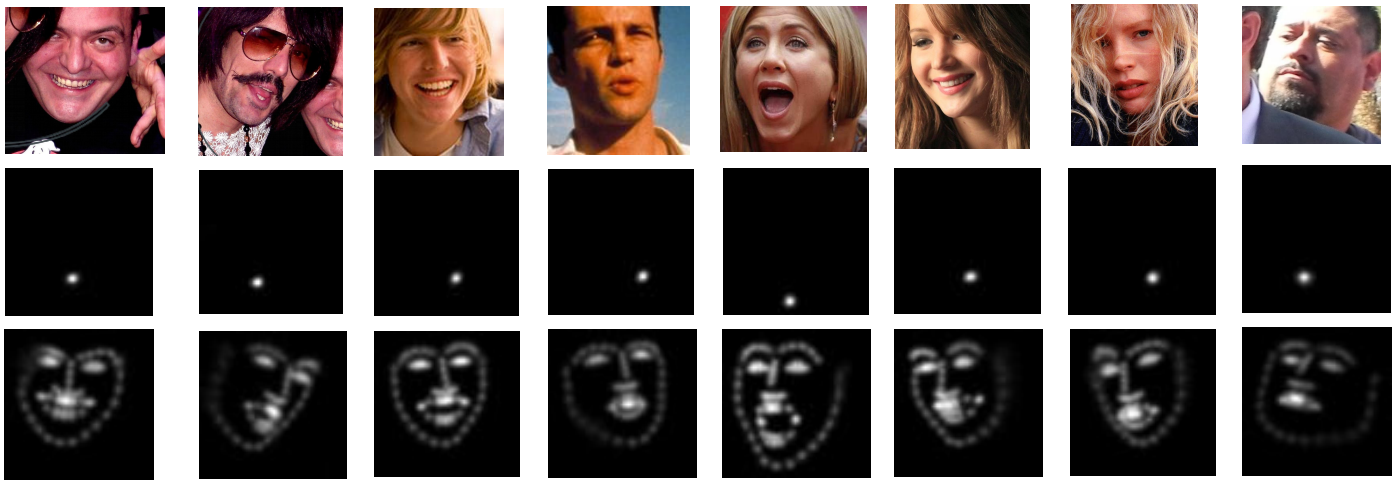


Fig. 4. Visualizations of facial landmarks localization heatmaps. The first row shows the face images, while the second row depicts a corresponding single heatmap of a particular facial feature. The third row shows the corresponding $N = 68$ points of all heatmap.

Table 1. Base subnetwork architecture. Two such non-weight-sharing units are used over here as depicted in Figure 3.

(a): $\text{FMU}_1/\text{FMU}_2$

Layer	Feature Map	F_{Size}	Stride	Pad
Input: I	256 x 256 x 3	-	-	-
A1-Conv	256 x 256 x 3	3 x 3	1 x 1	2 x 2
A1-ReLu	256 x 256 x 64	-	-	-
A2-Conv	256 x 256 x 64	3 x 3	1 x 1	2 x 2
A2-ReLu	256 x 256 x 64	-	-	-
A2-Pool	256 x 256 x 64	2 x 2	2 x 2	0 x 0
A3-Conv	128 x 128 x 64	3 x 3	1 x 1	2 x 2
A3-ReLu	128 x 128 x 64	-	-	-
A4-Conv	128 x 128 x 64	3 x 3	1 x 1	2 x 2
A4-ReLu	128 x 128 x 128	-	-	-
A4-Pool	128 x 128 x 128	2 x 2	2 x 2	0 x 0
A5-Conv	64 x 64 x 128	3 x 3	1 x 1	2 x 2
A5-ReLu	64 x 64 x 128	-	-	-
A6-Conv	64 x 64 x 128	3 x 3	1 x 1	2 x 2
A6-ReLu	64 x 64 x 128	-	-	-
A7-Conv	64 x 64 x 128	1 x 1	1 x 1	-
Output: F_1/F_2	64 x 64 x 68	-	-	-

(b): $\text{HMU}_E/\text{HMU}_1$

Layer	Feature Map	F_{Size}	Stride	Pad
Input: F_1/F_2	64 x 64 x 68	-	-	-
A8-Conv	64 x 64 x 68	9 x 9	1 x 1	8 x 8
A8-ReLu	64 x 64 x 68	-	-	-
A9-Conv	64 x 64 x 128	9 x 9	1 x 1	8 x 8
A9-ReLu	64 x 64 x 128	-	-	-
A10-Conv	64 x 64 x 128	1 x 1	1 x 1	0 x 0
A10-ReLu	64 x 64 x 256	-	-	-
A11-Conv	64 x 64 x 256	1 x 1	1 x 1	0 x 0
A11-ReLu	64 x 64 x 256	-	-	-
A11-Dropout(0.5)	64 x 64 x 256	-	-	-
A12-Conv	64 x 64 x 256	1 x 1	1 x 1	0 x 0
A12-ReLu	64 x 64 x 68	-	-	-
Output:$\text{H}_\text{E}/\hat{\text{H}}_1$	64 x 64 x 68	-	-	-

guidelines (Trigeorgis et al., 2016; Tuzel et al., 2016), $N = 68$ is employed in the designed architecture, thus resulting in feature maps $\mathbf{F}_1, \mathbf{F}_2 \in \mathbb{R}^{64\times64\times68}$. The subsequent layers A8-A12 make up the heatmap units HMU_E and HMU_1, which estimate the N heatmaps (one per facial feature) by applying 9×9 filters to encode the relations between neighboring facial features.

Table 2. Heatmap estimation unit HMU_k, $k = 2, \ldots, 4$. The input to each HMU is the output of the previous one, combined with the feature map F_2.

Layer	Feature Map	F_{Size}	Stride	Pad
Input: $\text{F}_2 \oplus \hat{\text{H}}_{k-1}$	64 x 64 x 136	-	-	-
B1-Conv	64 x 64 x 136	7 x 7	1 x 1	6 x 6
B1-ReLu	64 x 64 x 64	-	-	-
B2-Conv	64 x 64 x 64	13 x 13	1 x 1	12 x 12
B2-ReLu	64 x 64 x 64	-	-	-
B3-Conv	64 x 64 x 64	1 x 1	1 x 1	0 x 0
B3-ReLu	64 x 64 x 128	-	-	-
B4-Conv	64 x 64 x 128	1 x 1	1 x 1	0 x 0
B4-ReLu	64 x 64 x 68	-	-	-
Output: $\hat{\text{H}}_k$	64 x 64 x 68	-	-	-
L_{HM} regression loss				

Table 3. Landmark regression unit LRU_k, $k = 1, \ldots, 4$. The input to each LRU is the output of the previous one, the output of corresponding HMU, and the feature maps F_1 and F_2.

(a): RFMU_k

Layer	Feature Map	F_{Size}	Stride	Pad
Input: $\text{F}_1 \oplus \text{F}_2 \oplus \hat{\text{H}}_k \oplus \text{H}_E$	64 x 64 x 272	-	-	-
C1-Conv	64 x 64 x 272	7 x 7	2 x 2	5 x 5
C1-Pool	32 x 32 x 64	2 x 2	1 x 1	1 x 1
C2-Conv	32 x 32 x 64	5 x 5	2 x 2	3 x 3
C2-Pool	16 x 16 x 128	2 x 2	1 x 1	1 x 1
C3-Conv	16 x 16 x 128	3 x 3	2 x 2	1 x 1
C3-Pool	8 x 8 x 256	2 x 2	1 x 1	1 x 1
Output: E_k	8 x 8 x 256	-	-	-

(b): RLEU_k

Layer	Feature Map	F_{Size}	Stride	Pad
Input : $\text{E}_k \oplus \text{E}_{k-1}$	8 x 8 x 512	-	-	-
C4-Conv	8 x 8 x 512	3 x 3	2 x 2	1 x 1
C4-Pool	4 x 4 x 512	2 x 2	1 x 1	1 x 1
C5-Conv	4 x 4 x 512	3 x 3	2 x 2	1 x 1
C5-Pool	2 x 2 x 1024	2 x 2	1 x 1	1 x 1
C6-Conv	2 x 2 x 1024	1 x 1	1 x 1	0 x 0
Output : $\hat{R}_k$	1 x 1 x 136	-	-	-
L_{LM} regression loss				

BaseCNN_1 and BaseCNN_2 are trained using different losses and backpropagation paths, as depicted in Figure 3. Specifically, the first Base subnetwork, together with its outputs H_E

and $\mathbf{F}_1$, is connected to the CRCNN subnetwork, and is therefore trained by backpropagating its losses L_{LM}. Thus, $\mathbf{H}_\mathrm{E}$ is adapted to the regression task. On the other hand, BaseCNN$_2$ has two outputs, that is, the initial heatmap estimation $\hat{\mathbf{H}}_1$ and the feature map $\mathbf{F}_2$, which are forwarded to both the CHCNN and the CRCNN subnetworks, with both the losses L_{HM} and L_{LM} involved in the backpropagation process.

3.2. Cascaded Heatmap Estimation CNN (CHCNN)

The heatmaps $\hat{\mathbf{H}}_k$, $k = 1, \ldots, K$, are estimated in a coarse resolution of $1/s$, with $s = 4$, of the input image resolution. As shown in Figure 3, $K = 4$ HMUs are employed in the implementation of the proposed CCCN localization network. The structure of the HMUs used in the designed CHCNN is detailed in Table 2.

The cascaded architecture of the CHCNN implies that each HMU estimates a heatmap $\hat{\mathbf{H}}_k$ using the heatmap $\hat{\mathbf{H}}_{k-1}$ received from the preceding unit, and the feature map $\mathbf{F}_2$ estimated by the base subnetwork BaseCNN$_2$. Such joint input is shown in Table 2 as $\mathbf{F}_2 \oplus \hat{\mathbf{H}}_{k-1}$, with $\oplus$ denoting the concatenation of variables along their third dimension. The HMU architecture comprises wide filters of sizes $[7 \times 7]$ and $[13 \times 13]$, corresponding to the B1-Conv and B2-Conv layers, respectively. These layers encode the geometric relationships between relatively distant landmarks. Each heatmap is trained using L_{HM} (Eq. 1), with the locations of the facial landmarks labeled by narrow Gaussians with $\sigma = 1.3$ to improve training convergence.

3.3. Cascaded Regression CNN (CRCNN)

The CRCNN is applied to refine the robust, but coarse, landmark estimates of the CHCNN. The CRCNN comprises $K = 4$ LRUs, whose framework is detailed in Table 3. Specifically, the input to the k^{th} regression unit is obtained as the concatenation $\mathbf{F}_1 \oplus \mathbf{F}_2 \oplus \hat{\mathbf{H}}_k \oplus \mathbf{H}_E$ between the feature maps $\mathbf{F}_1$ and $\mathbf{F}_2$ computed by the base CNNs, the corresponding heatmap estimate $\hat{\mathbf{H}}_k$, and the activation map $\mathbf{H}_\mathrm{E}$ computed by BaseCNN$_1$.

Specifically, each LRUs consists of two succeeding parts:

- a regression feature map unit (RFMU), consisting of layers C1-C3 in Table 3, which computes the activation map $\mathbf{E}_k$;
- a residual localization error unit (RLEU), comprising layers C4-C6 in Table 3, that estimates the residual localization error $\hat{R}_k$.

The output of the k^{th} regression unit is the refinement term $\hat{R}_k$ as in Eq. 2 and Table 3. The network is trained using L_{LM} (Eq. 3), and the final residual localization estimate is given by the last regression unit.

4. Experimental Results

The proposed CCNN scheme has been experimentally evaluated on the image datasets typically considered in contemporary state-of-the-art works, in order to take into account distinct appearance and acquisition conditions of the considered face images. Specifically, we have performed tests on the 300-W competition dataset (Sagonas et al., 2016), a state-of-the-art face localization dataset comprising 3,837 near-frontal face images, and on the Caltech occluded faces in the wild (COFW) dataset (Burgos-Artizzu et al., 2013), a challenging dataset consisting of 1,007 faces depicting a wide range of occlusion patterns. The results obtained on the two considered datasets are respectively shown in Section 4.1 and 4.2. An ablation study, whose results are outlined in Section 4.3, has been also performed to analyze the specific contribution of each component of the proposed CCNN architecture.

All considered RGB images have been resized to 256×256 pixels, with their values normalized to the range $[-0.5, 0.5]$. Training images have been augmented using color variations, rotation by small angles, scaling, and translations. The learning rate has been changed gradually, starting with 10^{-5} for the initial 30 epochs, followed by 5×10^{-6} for the following five epochs, and then set to 10^{-6} for the remaining training epochs. The CCNN has been trained for 2,500 epochs in total.

The localization accuracy *per single face image* has been quantified through the normalized localization error (NLE) between the localized and ground-truth landmarks, that is,

$$NLE = \frac{1}{N \cdot d} \sum_{i=1}^{N} \left\| \hat{\mathbf{p}}^{(i)} - \mathbf{p}^{(i)} \right\|_2 , \tag{4}$$

where $\hat{\mathbf{p}}^{(i)}$ and $\mathbf{p}^{(i)}$ are the estimated and ground-truth coordinates of a particular facial landmark, respectively. The normalization factor d is either the inter-ocular distance (the distance between the outer corners of the eyes) (Ren et al., 2014; Zafeiriou et al., 2017; Zhu et al., 2015), or the inter-pupil distance (the distance between the eye centers) (Trigeorgis et al., 2016).

The localization accuracy *for a set of images* is quantified by the average localization error and the failures rate. We have considered as failures the estimates having a NLE greater than $\alpha = 0.08$ (Trigeorgis et al., 2016). We also report results in terms of area under the cumulative error distribution curve (AUC$_\alpha$) (Trigeorgis et al., 2016; Tuzel et al., 2016), summing up the obtained error distributions up to α. The proposed CCNN scheme has been implemented in Matlab and the MatConvNet-1.0-beta23 deep learning framework (Vedaldi and Lenc, 2015) using a NVIDIA Titan XP GPU.

4.1. 300-W Results

The 300-W competition dataset comprises images from five dabases, namely LFPW, HELEN, AFW, IBUG, and 300-W *private*[1]. Each image in the 300-W is annotated with 68 landmarks, and accompanied by a bounding box generated by a face detector. In the performed tests, the provided bounding boxes have been expanded by 20% on all sides, with the resulting region of interest resized to 256×256 pixels.

As in the most established approaches (Kowalski et al., 2017), the available data is divided into training and testing parts (Lee et al., 2015). The former set consists of the AFW dataset and subsets from the LFPW and HELEN datasets, for a total of 3148 images. The proposed CCNN has been trained using the 300-W training set, together with the frontal face images

[1]The "300-W *private* test set" dataset was originally a private and proprietary dataset used for the evaluation of the 300W challenge submissions.

Table 4. Facial landmarks localization results, in terms of NLE (%), on the 300-W *common*, *challenging*, LFPW, HELEN, *public*, and *private* datasets. Best results on each dataset are marked in bold.

Category	Paper	Method	Inter-ocular normalization						Inter-pupil normalization					
			Common	*Challenging*	LFPW	HELEN	*Public*	*Private*	*Common*	*Challenging*	LFPW	HELEN	*Public*	*Private*
CSR	Ren et al. (2014)	LBF	-	-	-	-	-	-	4.95	11.98	-	-	6.32	-
	Zhu et al. (2015)	CFSS	-	-	5.75	6.35	-	7.51	4.73	9.98	4.87	4.63	5.76	6.27
	Kowalski and Naruniec (2016)	*k*-cluster	3.34	6.56	-	-	3.97	-	-	-	-	-	-	-
DL	Zhang et al. (2016)	TCDCN	-	-	4.59	4.85	-	3.52	4.80	8.60	6.24	4.60	5.54	10.28
	Chen et al. (2017)	CTFF	-	-	-	-	-	-	**3.73**	7.12	-	-	**4.47**	-
DL+CSR	Trigeorgis et al. (2016)	MDM	-	-	-	-	4.05	5.05	-	-	-	-	-	-
	Xiao et al. (2016)	RAR	-	-	3.99	4.30	-	-	4.12	8.35	-	-	4.94	-
	Lai et al. (2017)	DRR	-	-	-	-	-	-	4.07	8.29	**4.49**	**4.02**	4.90	-
	He et al. (2017b)	FEC-CNN	-	-	-	-	-	-	4.98	6.56	-	-	5.14	-
DL+HM	Shao et al. (2017)	TSA	-	-	-	-	-	-	4.45	8.03	-	-	5.15	-
	Dong et al. (2018)	SAN	3.34	6.60	-	-	3.98	-	-	-	-	-	-	-
	Chen et al. (2020)	CGAN	-	-	-	-	-	3.96	-	-	-	-	-	-
DL+HM+CSR	Kowalski et al. (2017)	DAN	3.19	5.24	3.17	3.20	3.59	4.30	4.42	7.57	-	-	5.03	-
		DAN-Menpo	**3.09**	4.88	**3.05**	**3.12**	**3.44**	3.97	4.29	7.05	-	-	4.83	-
DL+HM+CSR	Proposed approach	CCNN	3.23	**3.99**	3.30	3.20	**3.44**	**3.33**	4.55	**5.67**	4.63	4.51	4.85	**4.74**

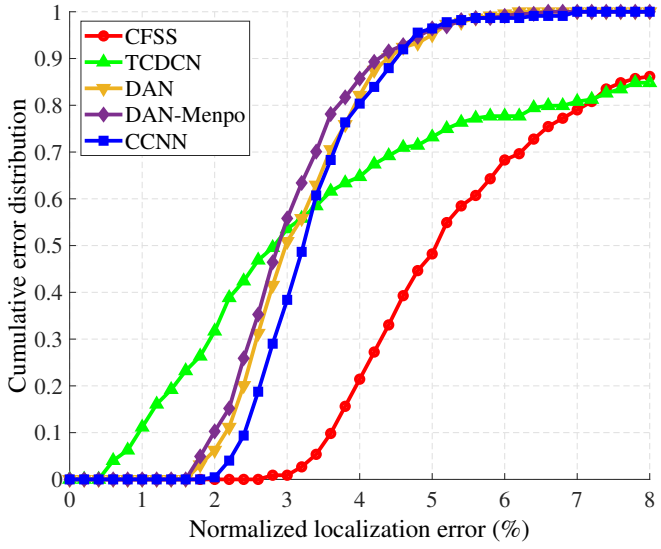


Fig. 5. CEDs *vs.* NLE results for the LFPW test set.

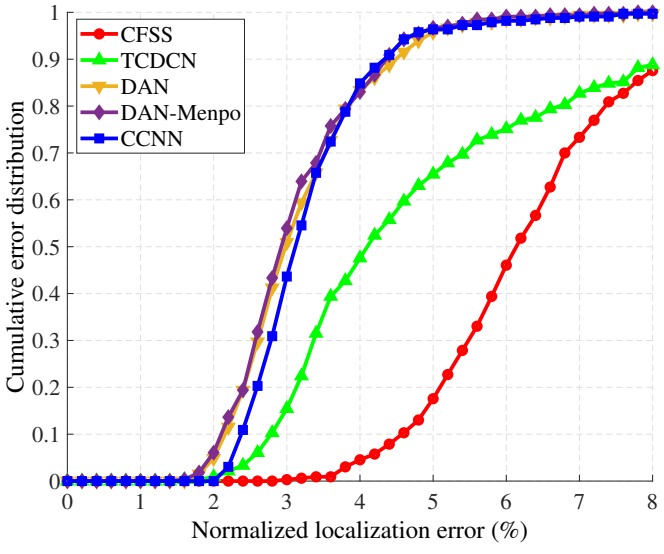


Fig. 6. CEDs *vs.* NLE results for the HELEN test set.

from the Menpo dataset (Zafeiriou et al., 2017), also annotated with 68 landmark points. The profile faces of the Menpo dataset have been instead annotated with 39 landmark points, and thus could not have been used in our evaluation. The overall training set consists of 11,007 images, out of which 2,500 samples have been randomly chosen and employed as validation set.

The test data consists of the remaining images from the 300-W datasets, comprising samples from IBUG, 300-W *private*, and the test sets from the LFPW and HELEN databases. In order to facilitate comparisons with state-of-the-art methods, the available test data is organized into a *common* subset, comprising test samples from the LFPW and HELEN datasets (554 images), a *challenging* subset, given by the IBUG dataset (135 images), a 300-W *public* subset, consisting of all the test samples from the LFPW, HELEN, and IBUG datasets (689 images), and the 300-W *private* test set (600 images).

In order to evaluate the effectiveness of the proposed approach, the results achieved with our CCNN architecture are compared with the performance reported in most of the state-of-the-art works mentioned in Section 2. All the approaches considered for comparison have been trained on the 300-W training dataset, while the methods in Chen et al. (2017), He et al. (2017b), and Shao et al. (2017) have added the Menpo database for training, as also done for some tests of the DAN approach Kowalski et al. (2017), whose results are therefore mentioned in the following as either DAN or DAN-Menpo, depending on the considered training dataset.

Table 5. Facial landmarks localization results, in terms of AUC and failure rate (FR, in %) for inter-ocular normalization, on the 300-W *public* and *private* datasets. Best results on each dataset are marked in bold.

Category	Paper	Method	*Public*		*Private*	
			$AUC_{0.08}$	FR	$AUC_{0.08}$	FR
CSR	Cao et al. (2012)	ESR	43.12	10.45	32.35	17.00
	Xiong and De la Torre (2013)	SDM	42.94	10.89	-	-
	Zhu et al. (2015)	CFSS	49.87	5.08	39.81	12.30
DL+CSR	Trigeorgis et al. (2016)	MDM	52.12	4.21	45.32	6.80
DL+HM	Chen et al. (2020)	CGAN	-	-	53.64	2.50
DL+HM+CSR	Kowalski et al. (2017)	DAN	55.33	1.16	47.00	2.67
		DAN-Menpo	57.07	0.58	50.84	1.83
DL+HM+CSR	Proposed approach	CCNN	**57.88**	**0.58**	**58.67**	**0.83**

Table 4 reports the NLE performance achieved on the *common* and *challenging* subsets, as well as on the LFPW, HELEN, *public* and *private* test datasets. The considered state-of-the-art approaches are grouped into categories relying on CSR, DL, and HM schemes, as mentioned in Section 2. The localization results of the comparison schemes are quoted as reported in the original papers. The obtained scores testify that our CCNN approach compares favorably with respect to all the other schemes. In particular, the proposed framework outperforms all previous approaches when applied to the *challenging*

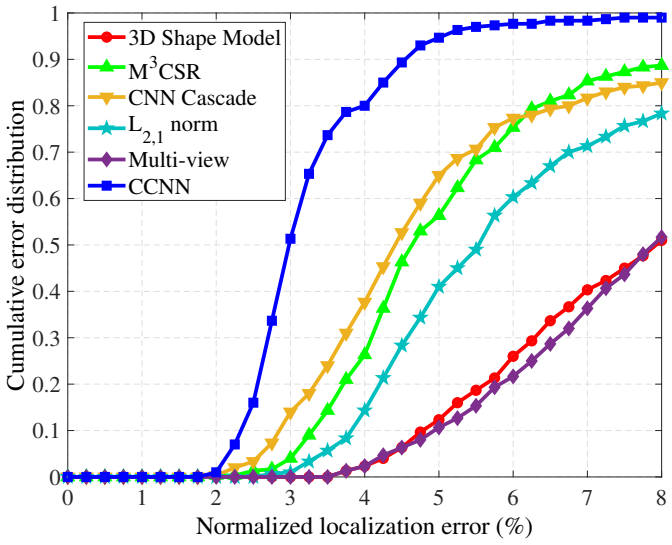


Fig. 7. CEDs *vs.* NLE results for the 300-W *indoor* test set.

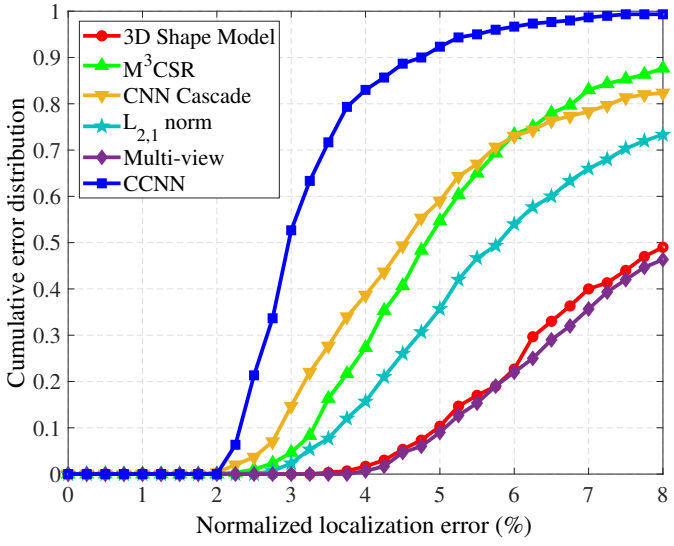


Fig. 8. CEDs *vs.* NLE results for the 300-W *outdoor* test set.

set, which comprises face images among the hardest to be processed. This capability can be attributed to the use of a combination of cascaded heatmap estimation and regression units.

When applied to the LFPW and HELEN dataset, the proposed CCNN is on par with existing state-of-the-art techniques, as these datasets mostly consist of easy-to-localize frontal face images. Figures 5 and 6 better detail the observed behaviors, respectively reporting the cumulative error distributions (CEDs) achieved on LFPW and HELEN datasets, when considering inter-ocular normalization.

On the other hand, the proposed CCNN outperforms state-of-the-art techniques on the 300-W *private* dataset, when considering both the inter-pupil and inter-ocular normalization. The obtained results point out that our method is less sensitive to low image quality, larger yaw angles, and facial deformations, with respect to the comparison approaches. In fact, a consistent accuracy over all considered image classes is achieved when employing the proposed CCNN. Conversely, the other approaches show notable dependency on the considered scenario, performing well on frontal face images such as those in the HELEN and LFPW datasets, yet tending to fail in challenging conditions such as those in the 300-W *private* set. For instance, applying the DAN and DAN-Menpo schemes (Kowalski et al., 2017) to the *challenging* dataset, mean errors at 5.23% and 4.88% are respectively obtained, whereas our method achieves a smaller error equal to 3.99%. The ability in producing consistent results irrespective of the considered facial conditions is therefore a core advantage of the proposed approach over contemporary state-of-the-art schemes.

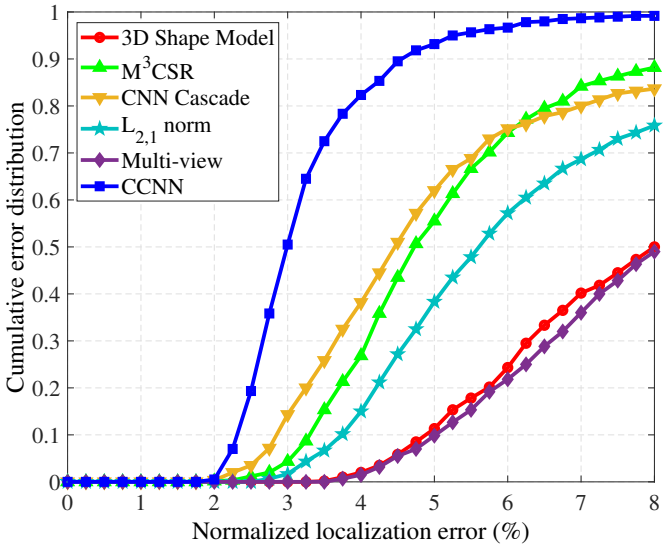


Fig. 9. CEDs *vs.* NLE results for the 300-W *private* test set.

Further results are provided in Table 5, where the $AUC_{0.08}$ measures and the localization failure rates of the proposed approach are compared against state-of-the-art schemes on the 300-W *public* and *private* datasets, for inter-ocular normalization. The proposed CCNN outperforms all the other schemes when applied to both test sets.

We have also performed tests on the *indoor* and *outdoor* subsets of the 300-W *private* test set, as done during the 300-W challenge (Sagonas et al., 2016). The results obtained in these scenarios for inter-ocular normalization are reported in terms of CEDs in Figures 7 and 8. The behaviors achieved with state-of-the-art approaches are quoted from the results of the 300-W challenge[2] (Sagonas et al., 2016). Specifically, the 3D Shape Model (Čech et al., 2016), M^{3}CSR (Deng et al., 2016), CNN Cascade (Fan and Zhou, 2016), $L_{2,1}$Norm (Martinez and Valstar, 2016), and Multi-view (Uřičář et al., 2016) approaches have been taken into account.

The proposed CCNN scheme significantly outperforms all contemporary schemes in both *indoor* and *outdoor* conditions. For the sake of completeness, Figure 9 depicts the CEDs achieved over the whole 300-W *private* test set.

Figure 10 shows some results applying the proposed CCNN scheme to images in the 300-W *indoor* and *outdoor* test sets. Red and green dots depict the ground-truth and the estimated

[2] https://ibug.doc.ic.ac.uk/media/uploads/competitions/300w_results.zip

Fig. 10. Facial landmarks localization examples, with images taken from the 300-W test set. Red and green dots depict the ground-truth and estimated landmark points by the proposed CCNN scheme, respectively.

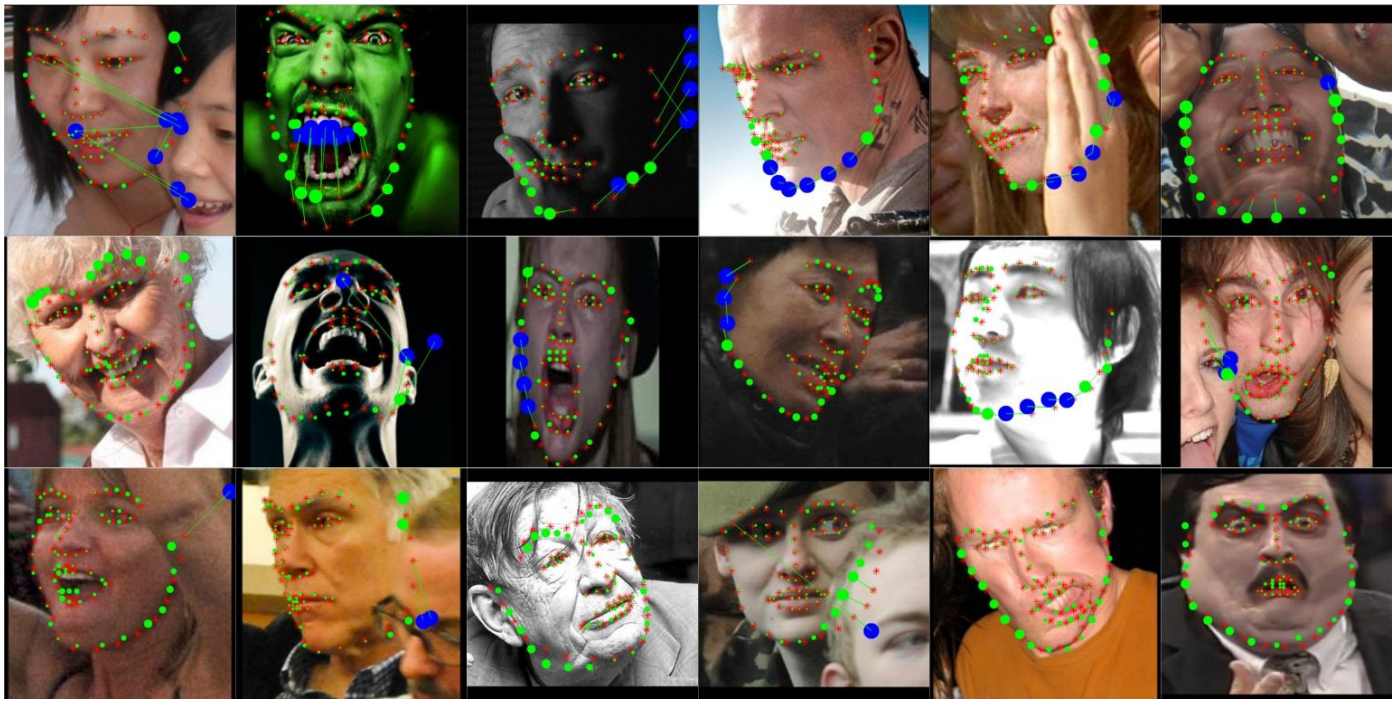


Fig. 11. Facial landmarks localization examples for extremely difficult images, with images taken from the 300-W *challenging* test set. Red and green dots depict the ground-truth and estimated landmark points by the proposed CCNN scheme, respectively. The sizes of the green dots increase according to the error of the estimated landmarks. When the error is more than 20 pixels, a line connects the ground-truth and the estimated landmarks. When the error is greater than 30 pixels then the landmark point dots are painted in blue.

landmark points, respectively. In particular, we show face images with relevant yaw angles, and facial expressions that significantly differ from frontal faces. The localization of landmarks on such images exemplifies the effectiveness of the proposed CCNN framework.

On the other hand, in Figure 11 we report examples of the landmarks detected in the 300-W *challenging* test set. There, the size of green dots increases according to the error of the estimated landmarks. When the error is greater than 20 pixels, a line connecting the ground-truth with the estimated landmarks is shown. When the error is greater than 30 pixels, the estimated landmark dots are painted in blue.

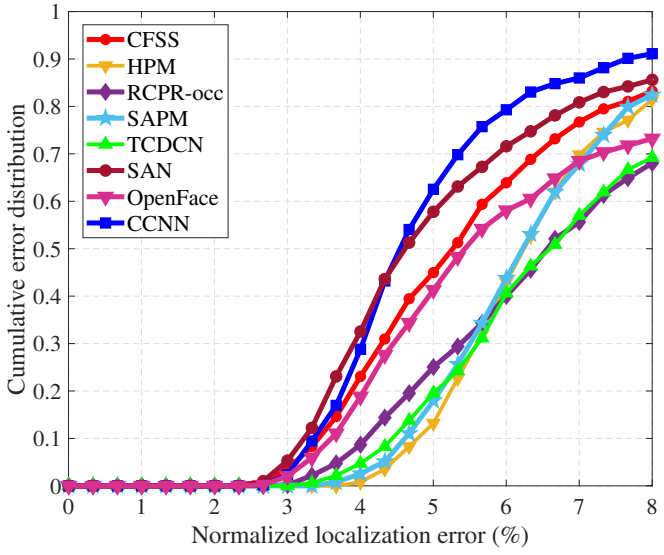


Fig. 12. CEDs *vs.* NLE results for the COFW test set.

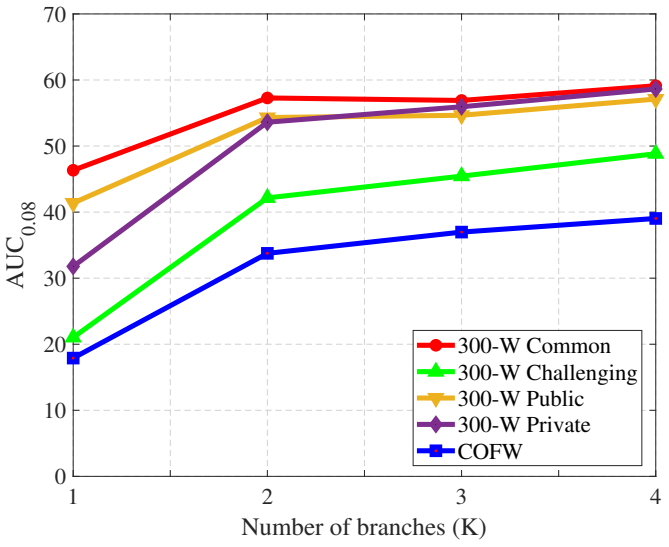


Fig. 13. Ablation study results for the proposed CCNN scheme: performance in terms of $AUC_{0.08}$ when varying the number of cascades in both the CHCNN (heatmap) and CRCNN (regression) subnetworks.

4.2. COFW Results

Like the 300-W database, also the COFW dataset has been annotated with 68 landmark points (Ghiasi and Fowlkes, 2015). Following the procedure employed in previous works (Burgos-Artizzu et al., 2013), we have used 500 images for training and 507 for testing the proposed network. The obtained accuracy has been compared with the publicly available[3] performance achieved with state-of-the-art localization schemes (Ghiasi and Fowlkes, 2015). In more detail, such results have been obtained training the CFSS (Zhu et al., 2015) and TCDCN (Zhang et al., 2016) schemes using the *HELEN*, *LFPW*, and *AFW* datasets. The RCPR-occ scheme (Burgos-Artizzu et al., 2013) has been trained using the same training sets as our CCNN model, while the HPM, SAPM (Ghiasi and Fowlkes, 2015), SAN[4](Dong et al., 2018), and OpenFace[5] (Zadeh et al., 2017) appraches have been trained using the *HELEN* and *LFPW* datasets.

The comparative results are depicted in Figure 12 in terms of CEDs, showing that the proposed CCNN approach significanlty outperforms all the other considered schemes when employed on the COFW dataset, furtherly testifying the effectiveness of the proposed method in challenging conditions.

4.3. Ablation Study

In order to investigate the contribution of each element in the proposed CCCN architecture, a detailed ablation study has been performed focusing on several aspects characterizing the proposed approach.

The influence on the achievable performance of the number K of cascaded units employed in the CHCNN (heatmaps) and CRCNN (regression) subnetworks has been first evaluated. For this purpose, we have trained the proposed CCNN using $K = \{1, 2, 3, 4\}$ cascades, with the same setup and training sets outlined in Section 4.1, using the same test sets described in Sections 4.1 and 4.2. The obtained results are depicted in Fig. 13 in terms of $AUC_{0.08}$, showing that increasing the number of employed cascaded units improves the achievable accuracy. The most significant improvement is attained when using two units instead of a single one, with additional cascades providing relatively small gains.

Further studies have been conducted to assess the effects of specific choices in the proposed CCNN design. The obtained results are reported in Table 6 in terms of NLE achieved over the 300-W *Public*, 300-W *Private*, and COFW datasets. As baseline reference, the error rates obtained with the proposed CCNN when trained for 2500 epochs, as in the tests described in Section4.1 and 4.2, and for 300 additional epochs, are reported in Table 6, showing that the proposed network has reached a plateau of performance.

The usefulness of the CRCNN regression subnetwork in refining the received outputs has been evaluated by performing test relying only on the CHCNN heatmap subnetwork. To do this, the BaseCNN_1 and CRCNN subnetworks have been disconnected from the rest of the CCNN network trained for 2800 epochs, and an additional training has been carried out for a further 20 epochs. A performance worsening is observed when carrying out such tests, especially in terms of $AUC_{0.08}$.

The importance of using two base subnetworks has been investigated in great detail, by first testing the performance of a network in which both the $\mathbf{F}_1$ and $\mathbf{F}_2$ have been disconnected from the CRCNN subnetwork. In order to do this, the input layer of LRUs, that is, C1-Conv in Table 3, has been resized and re-initialized. As in the previous case, the modified network has been trained for several additional epochs. The same has been done excluding only the $\mathbf{F}_1$ feature map from the CRCNN subnetwork. Furthermore, the whole BaseCNN_1 subnetwork has been removed from the rest of the architecture, thus letting the landmark regression layers getting their information only from $\mathbf{F}_2$ and $\hat{\mathbf{H}}_k$, $k = 1, \ldots, K$.

The relevance of the $\mathbf{F}_2$ feature map has been instead ad-

[3] https://github.com/golnazghiasi/cofw68-benchmark
[4] https://github.com/D-X-Y/SAN
[5] https://github.com/TadasBaltrusaitis/OpenFace

Table 6. Ablation studies performed on the proposed CCNN architecture. Tests have been done on the 300-W *Public*, 300-W *Private*, and COFW datasets. Results are given in terms of NLE (%), where the best performance marked as bold.

Tested condition	Additional epochs	300-W *Public*		300-W *Private*		COFW	
		NLE	$AUC_{0.08}$	NLE	$AUC_{0.08}$	NLE	$AUC_{0.08}$
Baseline CCNN	0	**3.44**	**57.88**	3.33	58.67	5.26	39.42
	300	3.45	57.03	**3.25**	59.55	5.23	39.78
CRCNN removed (only CHCNN)	300 + 20	3.47	56.61	3.28	59.28	5.20	39.62
$\mathbf{F}_1$ and $\mathbf{F}_2$ removed from CRCNN	300 + 20	3.57	55.94	3.91	51.96	5.24	40.07
	300+100	3.52	56.02	3.90	51.79	**5.10**	41.00
$\mathbf{F}_1$ removed from CRCNN	300 + 20	3.60	55.78	3.96	51.87	5.29	40.05
	300+100	3.54	55.95	3.91	51.70	5.16	40.99
BaseCNN_1 removed	300+20	3.54	55.92	3.90	51.99	5.24	40.43
	300+100	3.55	55.80	3.93	51.74	5.14	**41.14**
$\mathbf{F}_2$ removed from CRCNN	300+20	3.62	55.54	3.99	51.30	5.27	40.33
	300 + 100	3.54	55.94	3.93	51.48	5.13	40.90
$\mathbf{F}_2$ replaced by $\hat{\mathbf{H}}_k$ in CRCNN	300+20	3.45	56.98	3.38	57.86	5.16	40.52
	300+100	3.57	55.73	3.64	55.23	5.68	37.99
CHCNN and CRCNN with weight sharing	300+20	3.69	54.08	3.77	53.49	5.23	39.45
	300+100	3.78	53.41	3.94	52.51	5.71	37.98
Only $\hat{\mathbf{H}}_4$ as input for the CRCNN	300	3.53	55.92	3.91	51.50	5.12	41.08
K=6 cascaded units in CHCNN and CRCNN	300+300	3.45	57.02	3.26	**59.67**	5.13	40.41
	300+500	3.45	57.02	3.27	59.62	5.13	40.44

dressed considering two distinct scenarios: in a first case, $\mathbf{F}_2$ is removed as input from the CRCNN subnetwork, and layer C1-Conv is resized and re-initialized similarly to when $\mathbf{F}_1$ is excluded. In a second case, $\mathbf{F}_2$ is replaced by $\hat{\mathbf{H}}_k$, $k = 1, \ldots, K$, as input to each LRU_k (with $\hat{\mathbf{H}}_k$ therefore included twice in each k^{th} unit), so that C1-Conv can be kept with its original dimensions, without the need for a re-initialization.

The need for training the proposed CCNN without resorting to weight-sharing has been proven by testing a network where units in the CHCNN and CRCNN subnetworks are instead trained to share the same parameters, as in Belagiannis and Zisserman (2017). The results reported in Table 6 testify that such alternative degrades the achievable accuracy. The better behavior of the proposed CCNN is attributed to the capability of learning different regression functions per cascade.

We have also studied the possibility of using only the output of the CHCNN layer, that is, the highest quality heatmap $\hat{\mathbf{H}}_4$, as input to all the LRUs in the CRCNN subnetwork. The obtained results show that the proposed CCNN architecture, with paired HMUs and LRUs, can achieve better performance.

Eventually, we have tested an architecture with two additional cascaded units, initialized with the weights from the 4^{th} cascade, in each of the CHCNN and CRCNN subnetworks. A slight improvement has been in this case achieved over the COFW test set, while the performance on the 300-W dataset has not changed notably.

5. Conclusions

In this work, we have introduced a deep-learning-based cascaded formulation for coarse-to-fine localization of facial landmarks. The proposed cascaded CNN (CCNN) exploits two paired cascaded subnetworks: the heatmap subnetwork (CHCNN) estimates a coarse but robust heatmap corresponding to the facial landmarks, while the cascaded regression subnetwork (CRCNN) refines the accuracy of the CHCNN-generated landmarks via regression. The two cascaded subnetworks are aligned such that the output of each CHCNN unit is used as an input to the corresponding CRCNN unit. This allows the iterative refinement of the localization points. The CCNN is a face localization scheme that is fully data-driven and end-to-end trainable. It extends previous results on heatmap-based localization (Belagiannis and Zisserman, 2017), and it is experimentally shown to be robust to large variations in head poses. Moreover, it compares favorably with contemporary face localization schemes when evaluated using state-of-the-art face alignment datasets. The proposed CCNN scheme does not utilize any particular appearance attribute of faces, and can be applied to the localization of other classes of objects. Such an approach might pave the way for other localization solutions, such as those dedicated to sensor localization (Gepshtein and Keller, 2015; Keller and Gur, 2011), where the initial estimate of the heatmap is given by a graph algorithm, rather than an image domain CNN. The succeeding CNN architecture could be designed similarly to the proposed CHCNN and CRCNN subnetworks, thus offering an opportunity for further extensions in future works.

Acknowledgment

This work has been partially supported by COST Action 1206 {"De-identification for privacy protection in multimedia content" }. We gratefully acknowledge the support of NVIDIA Corporation for providing the Titan X Pascal GPU for this research work.

References

Belagiannis, V., Zisserman, A., 2017. Recurrent human pose estimation, in: 2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017), pp. 468–475. doi:10.1109/FG.2017.64.

Belhumeur, P.N., Jacobs, D.W., Kriegman, D.J., Kumar, N., 2013. Localizing parts of faces using a consensus of exemplars. IEEE Transactions on Pattern Analysis and Machine Intelligence 35, 2930–2940. doi:10.1109/TPAMI.2013.23.

Burgos-Artizzu, X.P., Perona, P., Dollár, P., 2013. Robust face landmark estimation under occlusion, in: 2013 IEEE International Conference on Computer Vision, pp. 1513–1520. doi:10.1109/ICCV.2013.191.

Cao, X., Wei, Y., Wen, F., Sun, J., 2012. Face alignment by explicit shape regression, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 2887–2894. doi:10.1109/CVPR.2012.6248015.

Chen, D., Ren, S., Wei, Y., Cao, X., Sun, J., 2014. Joint cascade face detection and alignment, in: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (Eds.), Computer Vision – ECCV 2014, Springer International Publishing, Cham. pp. 109–122.

Chen, X., Zhou, E., Mo, Y., Liu, J., Cao, Z., 2017. Delving deep into coarse-to-fine framework for facial landmark localization, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 2088–2095. doi:10.1109/CVPRW.2017.260.

Chen, Y., Shen, C., Chen, H., Wei, X., Liu, L., Yang, J., 2020. Adversarial learning of structure-aware fully convolutional networks for landmark localization. IEEE Transactions on Pattern Analysis and Machine Intelligence 42, 1654–1669.

Coifman, R.R., Lafon, S., 2006. Diffusion maps. Applied and Computational Harmonic Analysis 21, 5 – 30. URL: http://www.sciencedirect.com/science/article/pii/S1063520306000546, doi:https://doi.org/10.1016/j.acha.2006.04.006.

Cootes, T.F., Edwards, G.J., Taylor, C.J., 2001. Active appearance models. IEEE Transactions on Pattern Analysis and Machine Intelligence 23, 681–685. doi:10.1109/34.927467.

Cootes, T.F., Taylor, C.J., 1992. Active shape models — 'smart snakes', in: Hogg, D., Boyle, R. (Eds.), BMVC92, Springer London, London. pp. 266–275.

Cristinacce, D., Cootes, T., 2006. Feature detection and tracking with constrained local models, in: British Machine Vision Conference, pp. 929–938.

Deng, J., Liu, Q., Yang, J., Tao, D., 2016. M^3 CSR: multi-view, multi-scale and multi-component cascade shape regression. Image Vision Comput. 47, 19–26. doi:10.1016/j.imavis.2015.11.005.

Dong, X., Yan, Y., Ouyang, W., Yang, Y., 2018. Style aggregated network for facial landmark detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 379–388.

Fan, H., Zhou, E., 2016. Approaching human level facial landmark localization by deep learning. Image Vision Comput. 47, 27–35. doi:10.1016/j.imavis.2015.11.004.

Gepshtein, S., Keller, Y., 2015. Sensor network localization by augmented dual embedding. IEEE Transactions on Signal Processing 63, 2420–2431. doi:10.1109/TSP.2015.2411211.

Ghiasi, G., Fowlkes, C.C., 2015. Occlusion coherence: Detecting and localizing occluded faces. arXiv: 1506.08347. URL: https://dblp.org/rec/bib/journals/corr/GhiasiF15.

He, Z., Kan, M., Zhang, J., Chen, X., Shan, S., 2017a. A fully end-to-end cascaded CNN for facial landmark detection, in: 2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017), pp. 200–207. doi:10.1109/FG.2017.33.

He, Z., Zhang, J., Kan, M., Shan, S., Chen, X., 2017b. Robust FEC-CNN: A high accuracy facial landmark detection system, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 2044–2050. doi:10.1109/CVPRW.2017.255.

Huang, Z., Zhao, X., Shan, S., Wang, R., Chen, X., 2013. Coupling alignments with recognition for still-to-video face recognition, in: 2013 IEEE International Conference on Computer Vision, pp. 3296–3303. doi:10.1109/ICCV.2013.409.

Keller, Y., Gur, Y., 2011. A diffusion approach to network localization. IEEE Transactions on Signal Processing 59, 2642–2654.

Kowalski, M., Naruniec, J., 2016. Face alignment using k-cluster regression forests with weighted splitting. IEEE Signal Processing Letters 23, 1567–1571. doi:10.1109/LSP.2016.2608139.

Kowalski, M., Naruniec, J., Trzcinski, T., 2017. Deep alignment network: A convolutional neural network for robust face alignment, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 2034–2043. doi:10.1109/CVPRW.2017.254.

Kumar, N., Belhumeur, P., Nayar, S., 2008. Facetracer: A search engine for large collections of images with faces, in: Forsyth, D., Torr, P., Zisserman, A. (Eds.), Computer Vision – ECCV 2008, Springer Berlin Heidelberg, Berlin, Heidelberg. pp. 340–353.

Lai, H., Xiao, S., Pan, Y., Cui, Z., Feng, J., Xu, C., Yin, J., Yan, S., 2017. Deep recurrent regression for facial landmark detection. IEEE Transactions on Circuits and Systems for Video Technology PP, 1–1. doi:10.1109/TCSVT.2016.2645723.

Le, V., Brandt, J., Lin, Z., Bourdev, L., Huang, T.S., 2012. Interactive facial feature localization. Computer Vision – ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part III , 679–692doi:10.1007/978-3-642-33712-3_49.

Lee, D., Park, H., Yoo, C.D., 2015. Face alignment using cascade gaussian process regression trees, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4204–4212. doi:10.1109/CVPR.2015.7299048.

Lowe, D.G., 2004. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60, 91–110. doi:10.1023/B:VISI.0000029664.99615.94.

Mahpod, S., Keller, Y., 2018. Kinship verification using multiview hybrid distance learning. Computer Vision and Image Understanding 167, 28 – 36.

Martinez, B., Valstar, M.F., 2016. L2,1-based regression and prediction accumulation across views for robust facial landmark detection. Image and Vision Computing 47, 36 – 44. doi:https://doi.org/10.1016/j.imavis.2015.09.003.

Messer, K., Kittler, J., Sadeghi, M., Marcel, S., Marcel, C., Bengio, S., Cardinaux, F., Sanderson, C., Czyz, J., Vandendorpe, L., Srisuk, S., Petrou, M., Kurutach, W., Kadyrov, A., Paredes, R., Kepenekci, B., Tek, F.B., Akar, G.B., Deravi, F., Mavity, N., 2003. Face verification competition on the xm2vts database. Audio- and Video-Based Biometric Person Authentication: 4th International Conference, AVBPA 2003 Guildford, UK, June 9–11, 2003 Proceedings , 964–974doi:10.1007/3-540-44887-X_112.

Pfister, T., Charles, J., Zisserman, A., 2015. Flowing convnets for human pose estimation in videos, in: International Conference on Computer Vision (ICCV).

Ren, S., Cao, X., Wei, Y., Sun, J., 2014. Face alignment at 3000 fps via regressing local binary features, in: 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1685–1692. doi:10.1109/CVPR.2014.218.

Sagonas, C., Antonakos, E., Tzimiropoulos, G., Zafeiriou, S., Pantic, M., 2016. 300 faces in-the-wild challenge: database and results. Image and Vision Computing 47, 3 – 18. doi:https://doi.org/10.1016/j.imavis.2016.01.002.

Shao, X., Xing, J., Lv, J., Xiao, C., Liu, P., Feng, Y., Cheng, C., 2017. Unconstrained face alignment without face detection, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 2069–2077. doi:10.1109/CVPRW.2017.258.

Trigeorgis, G., Snape, P., Nicolaou, M.A., Antonakos, E., Zafeiriou, S., 2016. Mnemonic descent method: A recurrent process applied for end-to-end face alignment, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4177–4187. doi:10.1109/CVPR.2016.453.

Tuzel, O., Marks, T.K., Tambe, S., 2016. Robust face alignment using a mixture of invariant experts, in: Leibe, B., Matas, J., Sebe, N., Welling, M. (Eds.), Computer Vision – ECCV 2016, Springer International Publishing, Cham. pp. 825–841.

Uřičář, M., Franc, V., Thomas, D., Sugimoto, A., Hlaváč, V., 2016. Multi-view facial landmark detector learned by the Structured Output SVM. Image and Vision Computing 47, 45 – 59. doi:https://doi.org/10.1016/j.imavis.2016.02.004.

Vedaldi, A., Lenc, K., 2015. Matconvnet – convolutional neural networks for matlab, in: Proceeding of the ACM Int. Conf. on Multimedia.

Viola, P., Jones, M., 2001. Rapid object detection using a boosted cascade of simple features, in: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001, pp. I–511–I–518 vol.1. doi:10.1109/CVPR.2001.990517.

Xiao, S., Feng, J., Xing, J., Lai, H., Yan, S., Kassim, A., 2016. Robust facial landmark detection via recurrent attentive-refinement networks, in: Leibe, B., Matas, J., Sebe, N., Welling, M. (Eds.), Computer Vision – ECCV 2016, Springer International Publishing, Cham. pp. 57–72.

Xiong, X., De la Torre, F., 2013. Supervised descent method and its applications to face alignment, in: 2013 IEEE Conference on Computer Vision and Pattern Recognition, pp. 532–539. doi:10.1109/CVPR.2013.75.

Zadeh, A., Baltrusaitis, T., Morency, L., 2017. Convolutional experts constrained local model for facial landmark detection, in: CVPR Workshops, IEEE Computer Society. pp. 2051–2059.

Zafeiriou, S., Trigeorgis, G., Chrysos, G., Deng, J., Shen, J., 2017. The Menpo facial landmark localisation challenge: A step towards the solution, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 2116–2125. doi:10.1109/CVPRW.2017.263.

Zhang, Z., Luo, P., Loy, C.C., Tang, X., 2016. Learning deep representation for face alignment with auxiliary attributes. IEEE Transactions on Pattern Analysis and Machine Intelligence 38, 918–930. doi:10.1109/TPAMI.2015.2469286.

Zhu, S., Li, C., Loy, C.C., Tang, X., 2015. Face alignment by coarse-to-fine shape searching, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4998–5006. doi:10.1109/CVPR.2015.7299134.

Zhu, X., Ramanan, D., 2012. Face detection, pose estimation, and landmark localization in the wild, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 2879–2886. doi:10.1109/CVPR.2012.6248014.

Čech, J., Franc, V., Uřičář, M., Matas, J., 2016. Multi-view facial landmark detection by using a 3D shape model. Image and Vision Computing 47, 60 – 70. doi:https://doi.org/10.1016/j.imavis.2015.11.003.