Learning Robust Facial Landmark Detection via Hierarchical Structured Ensemble

Xu Zou[1,3], Sheng Zhong[2,3], Luxin Yan[2,3], Xiangyun Zhao[4], Jiahuan Zhou[4,*], Ying Wu[4]

[1] College of Life Science and Technology, Huazhong University of Science & Technology (HUST), Wuhan, China

[2] School of Aritificial Intelligence and Automation, HUST, Wuhan, China

[3] National Key Laboratory of Science & Technology on Multi-Spectral Information Processing, HUST, Wuhan, China

[4] Department of Electrical and Computer Engineering, Northwestern University, IL, USA

{xutsou,zhongsheng,yanluxin}@hust.edu.cn, zhaoxiangyun915@gmail.com, jzt011@eecs.northwestetn.edu, yingwu@northwestern.edu

Abstract

Heatmap regression-based models have significantly advanced the progress of facial landmark detection. However, the lack of structural constraints always generates inaccurate heatmaps resulting in poor landmark detection performance. While hierarchical structure modeling methods have been proposed to tackle this issue, they all heavily rely on manually designed tree structures. The designed hierarchical structure is likely to be completely corrupted due to the missing or inaccurate prediction of landmarks. To the best of our knowledge, in the context of deep learning, no work before has investigated how to automatically model proper structures for facial landmarks, by discovering their inherent relations. In this paper, we propose a novel Hierarchical Structured Landmark Ensemble (HSLE) model for learning robust facial landmark detection, by using it as the structural constraints. Different from existing approaches of manually designing structures, our proposed HSLE model is constructed automatically via discovering the most robust patterns so HSLE has the ability to robustly depict both local and holistic landmark structures simultaneously. Our proposed HSLE can be readily plugged into any existing facial landmark detection baselines for further performance improvement. Extensive experimental results demonstrate our approach significantly outperforms the baseline by a large margin to achieve a state-of-the-art performance.

1. Introduction

Facial landmark detection, known as face alignment, is essential to many facial analysis tasks including face recognition [35, 64, 30], face modeling [17, 24]. Due to the large variability of face shapes, head poses, lighting conditions, and background occlusions, facial landmark detection still remains challenging.

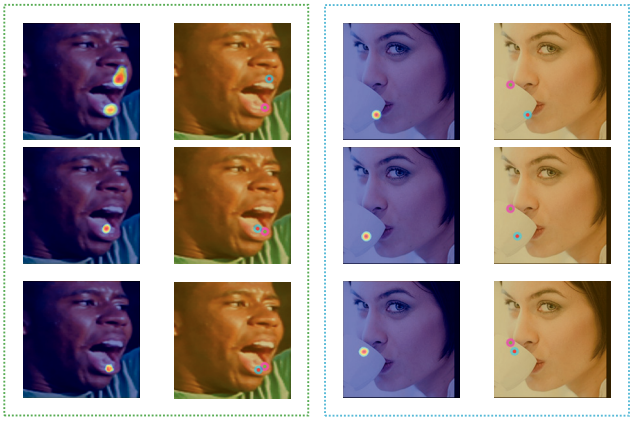


Figure 1: Two examples of abnormal situations. **Left**: a lip landmark in a facial image with exaggerated expression. **Right**: a cheek landmark in a facial image with occlusion. **Row.1** results are from a landmark detector without structural constraints, **Row.2** results are from a landmark detector with manually designed structural constraints and **Row.3** results are from a landmark detector with our proposed HSLE structural constraints. Red and Green dots represent predictions and groundtruth respectively.

Recently, the heatmap regression-based models [49, 57, 36, 52, 51] advance the progress of facial landmark detection. The success of heatmap regression-based models owes to the utilization of likelihood heatmaps to represent the probability distributions of landmark locations. However, inaccurate heatmaps (*e.g.* heatmaps with deviations or distractions) would be generated if abnormal situations occur (*e.g.* occlusion, illumination, noise, or unconstrained pose/expression variations, and etc.) which result in inaccurate or even incorrect localizations of facial landmarks (Figure 1.Row1) due to their low reliability or insufficient discrimination. To address this issue, structure modeling

*Jiahuan Zhou is the corresponding author.

in heatmap regression-based models has been proposed and achieved promising performance in facial landmark detection, since the aforementioned inaccurate/ambiguous landmarks could be amended and correctly reconstructed by utilizing structural constraints of facial landmarks. However, existing holistic structure modeling is sensitive to the landmark prediction quality, the constructed structure may be completely invalid due to the missing or inaccurate detection of landmarks caused by unconstrained abnormal situations as shown in Figure 1.Row2.

Therefore, by simultaneously modeling both holistic and local structures of landmarks, the localization of facial landmarks becomes more robust. Instead of using a holistic dense connected graph to simultaneously modeling the holistic and local structures of facial landmarks which is neither resource efficient nor feasible to inference, a hierarchical structure model is utilized for effective local structure modeling. In the community of facial landmark detection, few works [20, 9, 52] have been proposed to explore the hierarchical modeling of facial landmarks which all heavily rely on the manually designed tree-based hierarchical structure. However, their performance is not robust to the detection of facial landmarks since the manually designed tree structure will be totally corrupted because of the failure detection of landmarks. Therefore, in this work, we try to answer an important question: **can we automatically construct a more suitable hierarchical structure for learning robust facial landmark detection?**

In this paper, we propose a novel Hierarchical Structured Landmark Ensembles (HSLE) model to hierarchically represent both holistic and local structures for facial landmarks. In this work, we initially cluster landmarks into different groups, each of which shares the same landmarks that makes our model hierarchical. HSLE, a directed graph in essence, is constructed for each group automatically then. Each node in the HSLE denotes a predefined landmark and relationships from information passing between connected nodes are represented as edges in the HSLE. To construct the most reliable structure of the HSLE, a limited *Covering Set* model is utilized to discover the most robust connections of nodes. Due to the structural constraints propagated from the HSLE, a baseline facial landmark detector becomes more robust by trained jointly with the HSLE in an end-to-end fashion (Figure 1.Row3).

In this work, our contributions are four-fold: (1) We propose a novel Hierarchical Structured Landmark Ensembles (HSLE) model to hierarchically depict holistic and local structures for facial landmarks. Our proposed HSLE can be readily plugged into any existing facial landmark detection baselines for further performance improvement. (2) Due to the structural constraints propagated from the HSLE, the baseline facial landmark detector becomes more robust by trained jointly with the HSLE in an end-to-end fashion. (3) Compared with the aforementioned manual structure design-based methods, our automatically learned hierarchical structure is more reliable and robust to failure landmark detections, because structural constraints are automatically mined from data via discovering the most robust patterns. (4) Our approach significantly outperforms the baseline by a large margin to achieve a state-of-the-art result. The effectiveness of our model has been verified by extensive experiments on the 300W dataset [40, 41, 42] and the AFLW dataset [34].

2. Related Works

A large number of impressive achievements in the literature of facial landmark detection have been made since 1992. Because only landmarks with sufficient discrimination (*e.g.* corners of the ocular, corners of the mouse, and nose tip etc.) can be reliably located, structural constraints are usually adopted by former classic artworks, including Active Shape Models[13, 10, 37, 12], Active Appearance Models[11, 18, 43, 23, 46, 31], Constrained Local Models[14, 29, 44, 2, 28, 45], and Cascade Regression Models[6, 58, 65, 39, 7, 8, 25, 47, 66, 22, 55, 54, 56, 67, 19, 51]. Most of these methods start from an initial shape (*e.g.* mean facial shape[3]) or use the Point Distribution Model (*e.g.* [5, 60]) to enforce such a constraint.

Recently, Deep Convolutional Neural Networks (CNN) have advance the progress of facial landmark detection [48, 61, 62, 63, 59, 55, 50, 33]. Especially, state-of-the-art performance on facial landmark detection is achieved mostly by using heatmap regression models[49, 57, 36, 52, 51]. Merget *et al.* [36] introduced a fully-convolutional local-global context network, and a simple PCA-based 2D shape model is fitted as a holistic structural constraint. Wu *et al.* [52] proposed a boundary-aware face alignment model by utilizing boundary lines as the geometric structure.

Few existing works [20, 9] focus on hierarchically modeling both holistic and local structures. Ghiasi*et al.* [20] proposed a hierarchical deformable part model for localizing facial landmarks, which consists of a manually designed tree of parts. Each part is connected to a set of landmarks in a star topology. Their model would fail if the root node of the tree missed. Chu *et al.* [9] proposed a structured feature learning framework to reason the correlations among body joints in human pose estimation. A bi-directional tree structured model, which is also adopted by [52], is designated by hand for passing information between the neighbor joints. Since the number of facial landmarks is much more than the number of joints used in pose estimation, passing information between so many facial landmarks cannot be solved by a manually designed tree structure which is not robust enough either. Different from [20] and [9], our proposed HSLE model is constructed automatically via discovering the most robust patterns. Therefore, our HSLE is able to

robustly depict both local and holistic landmark structures simultaneously. To the best of our knowledge, our work is the first one to automatically model proper structures for landmarks, by discovering their inherent relations.

3. Our Method

As mentioned in the introduction, simultaneously modeling both holistic and local structures would be helpful for localizing facial landmarks more robustly. We propose a Hierarchical Structured Landmark Ensemble (HSLE) model for learning robust facial landmark detection by using it as structural constraints. The framework of the proposed approach is illustrated in Figure 2. The entire model can be jointly learned in an end-to-end fashion. The proposed HSLE model served as hierarchical structural constraints of facial landmarks.

Let $\mathcal{L} = \{l_1, l_2, ..., l_m\}$ denotes landmarks, $l_t \in \mathbb{R}^2$ indicates a landmark located at (x, y). The location of a landmark l_x in image $\boldsymbol{I} \in \mathbb{R}^{W \times H}$ is determined by:

$$l_t^* = \arg\max \phi_t(\boldsymbol{I}; \theta) + \tilde{\mathcal{H}}_t \tag{1}$$

where $\phi(\cdot)$ could be an arbitrary landmark detector, *e.g.* Stacked Hourglass[38] model. The output of $\phi_t(\cdot)$ is a heatmap $\mathcal{H}_t \in \mathbb{R}^{W \times H \times 1}$ of landmark t in $\mathcal{L}$. $\tilde{\mathcal{H}}_t \in \mathbb{R}^{W \times H \times 1}$, the output of HSLE, is expressed as the sum of a set of structural constraints of landmark t regularized by other landmarks. θ is the parameter of $\phi(\cdot)$.

In this section, we initially propose our Hierarchical Structured Landmark Ensemble (HSLE) model after recapping the traditional Covering Set model. Then we present the strategy for clustering landmarks into ensembles. Finally, we introduce the pattern discovery method for constructing the HSLE model, as well as some training issues.

3.1. Hierarchical Structured Landmark Ensemble

Based on the above concepts, both holistic and local structural constraints are useful for making facial landmark detection more robust. But since the number of facial landmarks is huge, it would be neither resource efficient nor feasible to inference, if a holistic dense connected graph were adopted for simultaneously modeling both holistic and local structures of landmarks. HSLE model is proposed to handle this issue.

HSLE means clustering landmarks into different groups, connecting these landmarks within each group on the basis of specific structures, and passing information from one landmark to another through these structures. It is noteworthy that different ensembles may share the same landmarks which makes this model hierarchical. Hierarchical structural constraints represented by the HSLE model would help make landmark detection more robust by propagating these constraints to the landmark detector when training.

3.1.1 Brief Overview of Covering Set

However, inappropriate landmark structures will make structural constraints useless. For example, there are two inappropriate landmark ensembles as shown in Figure 3(a) and Figure 3(b). In Figure 3(a), other nodes would not receive any information if node "C" missed. In Figure 3(b), other nodes would be misled if node "C" missed, since other nodes can only receive information from just one node. Dai *et al*. [15] introduce an idea named Covering Set for detector ensemble. A (n, t, m) Covering Set is a n-elements set composed of several m-elements substructures. For any t elements, there must exist at least one m-elements substructure whose elements are all belonged to those t elements. That is, there will exist at least one substructure if no more than $(n - t)$ nodes missed. Figure 3(c) and Figure 3(d) illustrate two different $(5, 4, 3)$ Covering Sets, and Figure 3(e) shows a fully connected graph which is also a $(5, 3, 3)$ Covering Set.

3.1.2 Formation of HSLE

Structure. Some extreme cases may be generated by the traditional Covering Set model. Figure 3(c) shows an example of a traditional $(5, 4, 3)$ Covering Set. In this extreme case, node "B" can not receive any information from other nodes, since node "B" doesn't belong to any substructure. To avoid this situation, owing to the property of the Covering Set model, a collection of limited Covering Set models is used to establish the structure of the HSLE, all elements in this collection together constitute the structure of the HSLE. The most robust structure $\mathcal{C}^*$ can be determined by:

$$\mathcal{C}^* = \arg\min_{\mathcal{C}} \sum_{i=1}^{N} \sum_{S_j^i \in \mathcal{C}_i} \kappa_{s_j^i}, \quad \left(\mathcal{C}_i^* = \arg\min_{\mathcal{C}_i} \sum_{S_j^i \in \mathcal{C}_i} \kappa_{s_j^i} \right)$$
$$s.t. \quad \begin{cases} \mathcal{K}(\mathcal{C}_i) \geq n_i - t_i \\ \mathcal{C}_i \subset T_i \\ \forall l_x \in \mathcal{C}_i, \sum_{j=1}^{m} \mathbf{1}\left(l_x \in S_j^i\right) \neq 0 \end{cases} \tag{2}$$

where $\mathcal{C} = \{\mathcal{C}_1, ..., \mathcal{C}_N\}$ illustrates a collection of limited Covering Sets from all ensembles, N is the total number of ensembles, $\mathcal{C}_i$ is one of the $\mathcal{C}$ consisted of substructures, $\mathcal{K}(\mathcal{C}_i)$ is the components missing tolerance number of $\mathcal{C}_i$, and $T_i = \left\{S_1^i, ..., S_m^i\right\}$ is a collection of all substructure candidates of ensemble i respectively, $\kappa_{s_j^i}$ is the error measure (*e.g.* mean inter-ocular normalized point-to-point Euclidean error of training set) of substructure S_j^i, n_i and t_i are model parameters as described above. $\mathbf{1}(\Delta) = 1$ if Δ is TRUE else $\mathbf{1}(\Delta) = 0$.

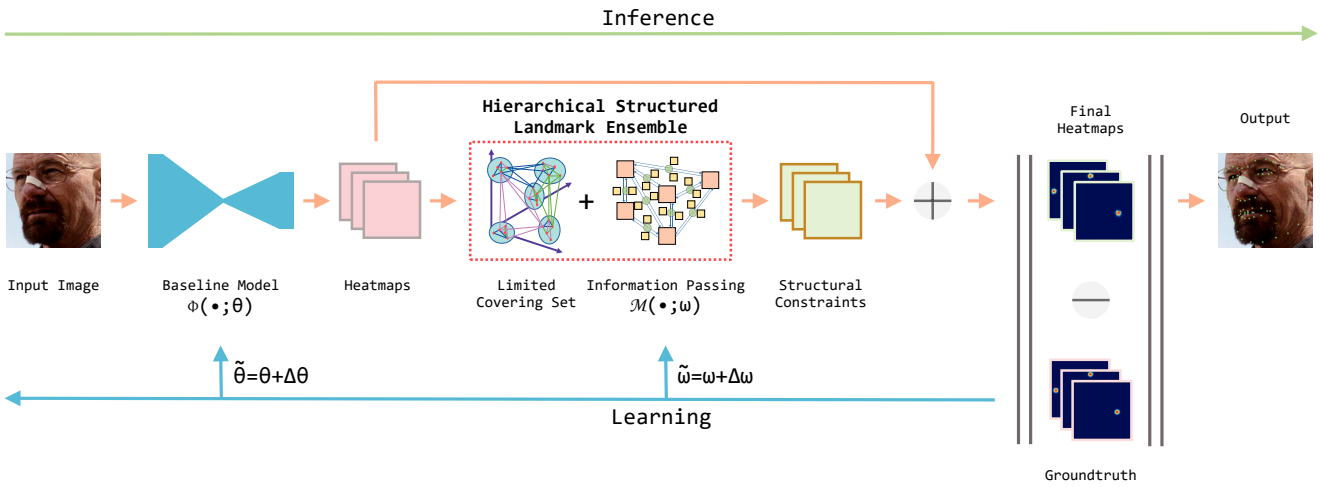


Figure 2: The framework of the proposed approach. The Hierarchical Structured Landmark Ensemble (HSLE) model is first constructed automatically by discovering the most robust patterns. The baseline facial landmark detector is trained jointly with the HSLE model in an end-to-end fashion then. The HSLE model is used to represent holistic and local structural constraints of facial landmarks. Structural constraints, outputs of the HSLE, are expressed as a set of feature maps have the same 2D shape as heatmaps generated by the baseline model. In inference, the output of the entire model is a set of landmark coordinates directly derived from final heatmaps according to Equation 1. Please note that, in the inference session, since parameters of the baseline model has been learned jointly at feature level under the supervision of structural constraints propagated from HSLE, heatmaps generated by the baseline model no longer indicate probability distributions of landmark locations.

Nodes & Edges. As described above, HSLE is a directed graph model in essence. Each node in HSLE denotes a predefined landmark. Relationships represented by information passing between connected nodes are denoted as edges in HSLE. We implement information passing as convolutional kernels following [9].

This limited Covering Set model enforces that each landmark should be included in at least one substructure. An example of a constructed Landmark Ensemble is illustrated in Figure 3(f).

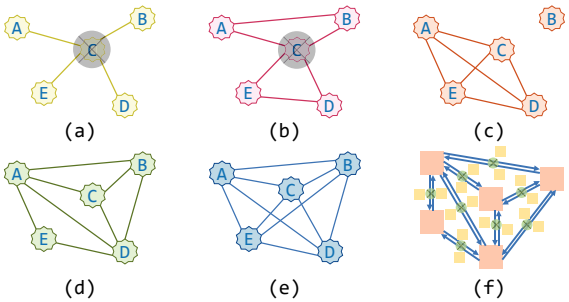


Figure 3: (a), (b), (c), (d), and (e): Illustration of different ideas about ensembles. (f): One example of a constructed landmark ensemble. Red squares illustrate heatmaps. Yellow squares illustrate convolutional kernels. Green circles illustrate convolution operations.

3.2. Landmark Clustering Strategy

Landmarks with stable relative relations (the relationship between a pair of landmarks should be invariant to head pose or facial expression to some extent) are preferred to be clustered into the same ensemble. The objective function of the landmark clustering operation could be written as:

$$E = \sum_{k=1}^{K} \sum_{i=1}^{M-1} \sum_{j=i+1}^{M} \mathbf{1}\left(\{i,j\},k\right) \cdot \left(\alpha s_{i,j} + \beta v_{i,j}^2\right) + \gamma \mathcal{V}^2,$$

$$\begin{cases} s_{i,j} = \sum_{p=1}^{P} \|l_{pi} - l_{pj}\|_2 \,/_P \\ v_{i,j}^2 = \sum_{p=1}^{P} (\|l_{pi} - l_{pj}\|_2 - s_{i,j})^2 \,/_P \\ \mathcal{V}^2 = \sum_{i=1}^{M} (e_i - \bar{e})^2 \,/M \end{cases}, \quad s.t. \quad \prod_{i=0}^{M} e_i \neq 0 \tag{3}$$

where K is the total number of ensembles. M is the total number of landmarks. P is the total number of training images (a subset of the entire training set). $\mathbf{1}\left(\{i,j\},k\right) = \begin{cases} 1, & \text{if } \{i,j\} \in \text{ensemble } k \\ 0, & \text{if } \{i,j\} \notin \text{ensemble } k \end{cases}$. $\|l_i - l_j\|_2$ is the point-to-point Euclidean distance between landmark i and j. e_i is the number of selected times of landmark i. $\bar{e}$ is the average selected times of all landmarks. α, β and γ are weights.

To solve this problem, we first randomly pick one of training images. Landmarks in that image are clustered into

different groups by leveraging K-means[32]. A landmark might be clustered into different groups at the same time if the difference were less than a threshold. Distances between landmarks and clustering centers would be refined by $\mathcal{V}^2$, to satisfy the constraints defined in Equation 3. E is calculated for the current clustering result. We run this entire procedure multiple times, the clustering result with the minimal E is selected as the final clustering result for constructing the HSLE.

3.3. Pattern Discovery for HSLE Construction

Since solving Equation 2 is a combinatorial optimization problem, randomized method inspired by [15] would be adopted to find the most robust structure for the HSLE.

The limited Covering Set of each ensemble is initialized into a fully connected graph. In each step we would randomly remove one edge from an arbitrary limited Covering Set if the constraints defined in Equation 2 were satisfied. Edges between landmarks with higher error measures within ensembles containing more edges have a larger probability to be removed. This procedure ends until a collection of minimal limited Covering Sets are obtained. All elements in this collection together constitute the structure of HSLE. To achieve the most robust structure, similar to the landmark clustering strategy, we run this entire procedure multiple times. The collection of minimal limited Covering Sets with the least error measure in total would be selected for constructing the structure of HSLE.

Structure construction procedure is summarized as Algorithm 1, $f(\cdot)$ is a function for counting the number of remaining edges.

3.4. Model Training

The loss of a baseline landmark detector could be simplified as:

$$\mathcal{L} = \|\phi(\boldsymbol{I};\theta) - \mathcal{G}\|_2^2 \tag{4}$$

where $\phi(\cdot) \in \mathbb{R}^{W*H*C}$ is the output of the baseline landmark detector. $\boldsymbol{I} \in \mathbb{R}^{W'*H'}$ is the input image. θ is parameters of the $\phi(\cdot)$. $\mathcal{G} \in \mathbb{R}^{W*H*C}$ is the groundtruth heatmap generated by groundtruth coordinates with a Gaussian distribution. C is the total number of landmarks.

To regularize the model with structural constraints determined by the HSLE, we can rewrite the loss as:

$$\mathcal{L}' = \sum_{t=1}^{C} \left\| \phi_t(\boldsymbol{I};\theta) + \sum_{p=1}^{P^t} \tilde{\mathcal{H}}_t^p - \mathcal{G}_t \right\|_2^2 \tag{5}$$

where $\tilde{\mathcal{H}}_t^p \in \mathbb{R}^{W*H}$ is the p^{th} structure constraint, expressed by the HSLE, regularized by another landmark. P^t is the total number of structural constraints of landmark t regularized by other landmarks.

The entire model is then trained in an end-to-end fashion by minimizing Equation 5.

Algorithm 1 Structure construction for HSLE.

Input:
The total number of ensembles: N
N limited Covering Sets: $\mathcal{C} = \{\mathcal{C}_1, ..., \mathcal{C}_N\}$
N fully connected graphs: $\mathcal{F} = \{\mathcal{F}_1, ..., \mathcal{F}_N\}$
The error measure of substructure S_j^i of ensemble i: $\kappa_{s_j^i}$, the error measure of edge i: $\varkappa_i$
The number of repetitions: R

Output:
The optimal collection of minimal limited Covering Sets: $\mathcal{C}^*$

$\mathcal{C}^* = \mathcal{F}, \kappa_{\min} = \infty$
for r **in** range(R) **do**
  for n **in** range(N) **do**
    $\mathcal{C}_n = \mathcal{F}_n$
  end for
  while $\exists \mathcal{C}_i \in \mathcal{C}, \mathcal{C}_i \neq \mathcal{C}_i^*$ **do**
    Find all removable edges $\{\mathcal{E}_{r1}, \mathcal{E}_{r2}, ..., \mathcal{E}_{rt}\}$ defined in Equation 2
    Calculate removal probabilities for each edge $\{\mathcal{P}_{r1}, \mathcal{P}_{r2}, ..., \mathcal{P}_{rt}\}$, $\mathcal{P}_{rj} = \mu^{\varkappa_{\mathcal{E}_{rj}}}/y + \lambda^{f(\mathcal{C}_i)}/z$, with y, z are normalization factors and μ, λ are constants
    Remove $\mathcal{E}_{rx}$ with probability $\mathcal{P}_{rx}$
  end while
  $\kappa = 0$
  for n **in** range(N) **do**
    $\kappa = \kappa + \sum_{S_j^n \in \mathcal{C}_n} \kappa_{s_j^n}$
  end for
  if $\kappa < \kappa_{\min}$ **then**
    $\kappa_{\min} = \kappa, \mathcal{C}^* = \mathcal{C}$
  end if
end for

4. Experiment

4.1. Implementation Details

We evaluate our model on two datasets to verify the effectiveness of our model. **300W**[40, 41, 42]: 3148 images for training and 689 images for testing. The testing dataset is split into three subsets: common subset (554 images), challenge subset (135 images) and the full set (689 images). Each image is annotated with 68 landmarks. **AFLW-Full**[67]: 20k images for training and 4386 images for testing. Each image is annotated with 19 landmarks.

For a trade-off between accuracy and efficiency, 17 ensembles are used to hierarchically depict the structure of 68 facial landmarks on 300W dataset, 4/5/8/11/14 ensembles are used to hierarchically depict the structure of 19 facial landmarks on AFLW dataset respectively. Each ensemble consists of a $(6, 4, 3)$ limited Covering Set com-

posed of 13 3-elements substructures. "max-pooling $\rightarrow$ 5×5-conv $\rightarrow$ nearest-upsampling" operations are adopted for information passing. All training images are cropped and resized to 256×256 according to the revised bounding boxes by enlarging about 14% ($\frac{256}{224}$) on the basis of provided groundtruth bounding boxes. All experiments have been carried out with the settings described in this section. The entire end-to-end model is trained from scratch. All our models are trained with Tensorflow[1].

4.2. Quantitative Results

We firstly compare our end-to-end trained model against the state-of-the-art methods on 300W[40, 41, 42] dataset. We report average point-to-point Euclidean errors normalized by both inter-pupil distance (ipd-norm) and inter-ocular distance (iod-norm), and median point-to-point Euclidean errors normalized by inter-ocular distance (iod-norm). For the comparison with all the other methods, we show the original results published in the literature.

The results are shown in Table 1. Experimental results demonstrate our approach consistently and significantly outperforms 3 different state-of-the-art baselines by a large margin to achieve a comparable result against the state-of-the-art methods. That is, due to the structural constraints propagated from the HSLE, the baseline facial landmark detectors become more robust by trained jointly with the HSLE. This phenomenon indicates that facial landmark detection can be more robust via learning from hierarchical structural constraints. We plot Cumulative Error Distributions curves of our proposed model against the 8-Stacked Hourglass[38] baseline model on the 300W dataset, as shown in Figure 4.

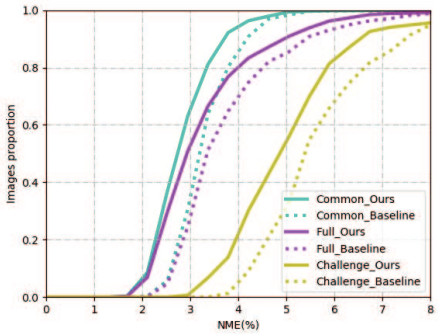


Figure 4: Cumulative Error Distributions curves of proposed model against the baseline model on the 300W dataset. Best viewed in color.

Since different facial landmarks have different discriminations, higher weights should be assigned to landmarks that are more discriminative. To this end, we report weighted mean iod-norm errors on 300W dataset. We divide 68 facial landmarks into three categories according to

	SC	Com.	Chal.	Full
		mean iod-norm error		
PCD-NN[27]	M	3.67	7.62	4.44
SAN[16]	N	3.34	6.60	3.98
DAN[26]	I	3.19	5.24	3.59
LAB[52]	M	2.98	5.19	3.49
DU-Net-BW-α[49]	N	3.00	5.36	3.46
DU-Net[49]	N	2.82	5.07	3.26
DCFE[51]	I	2.76	5.22	3.24
HG[38]*	N	3.30	5.69	3.77
HG-HSLE (ours)	A	**2.85**	**5.03**	**3.28**
DU-NET[49]*	N	3.07	5.13	3.47
DU-NET*-HSLE(ours)	A	**2.88**	**5.01**	**3.30**
Merget *et al.* [36]*	N	3.76	6.32	4.26
Merget*-HSLE(ours)	A	**3.21**	**5.69**	**3.70**
		mean ipd-norm error		
MDM[50]	I	4.83	10.14	5.88
TCDCN[63]	N	4.80	8.60	5.54
CFSS[66]	I	4.73	9.98	5.76
RCN[21]	A	4.67	8.44	5.41
DAN[26]	I	4.42	7.57	5.03
TSR[33]	M	4.36	7.56	4.99
RAR[55]	A	4.12	8.35	4.94
SHN[57]	N	4.12	7.00	4.68
DVLN[53]	N	3.94	7.62	4.66
DCFE[51]	I	3.83	7.54	4.55
LAB[52]	M	3.42	6.98	4.12
HG[38]*	N	4.56	8.18	5.27
HG-HSLE (ours)	A	**3.94**	**7.24**	**4.59**
		median iod-norm error		
TCDCN[63]	N	4.11	6.87	-
CLNF[4]	I	3.47	6.37	-
CFSS[66]	I	3.20	5.97	-
CE-CLM[60]	I	3.13	5.62	-
Merget *et al.* [36]	N	3.04	5.55	-
Merget *et al.* [36](fit)	I	2.86	5.29	-
HG[38]*	N	3.17	5.42	-
HG-HSLE (ours)	A	**2.72**	**4.86**	-
DU-Net[49]*	N	2.79	4.75	-
DU-NET*-HSLE(ours)	A	**2.73**	**4.67**	-
Merget *et al.* [36]*	N	3.10	5.54	-
Merget*-HSLE(ours)	A	**2.87**	**5.07**	-

Table 1: Mean/Median point-to-point Euclidean errors (%) normalized by ipd or iod on 300W dataset. HG[38]*, DU-Net[49]* and Merget *et al.* [36]* are selected as the baselines. **OURS** and the BEST performance (besides ours) are highlighted in **bold** and underlined respectively. "*" indicates results re-implemented by ourselves with the code provided by the authors. SC indicates the adopted structure constraint (N for none or not mentioned, M for manually designed, I for initial shape, A for automatically constructed).

their discriminations (Figure 5). Category 𝔞 contains landmarks with lowest discriminations (*e.g.* outlines), category 𝔠 contains landmarks with highest discriminations (*e.g.* corners of ocular), and category 𝔟 contains all other remaining landmarks. We assign different weights to landmarks according to their category. Results are shown in Table 2. Numbers in the first column indicate the relevant weights. What could be learned from Table 2 is that facial landmarks with higher discriminations could achieve more improvements by our HSLE model, makes our proposed idea more sense for most applications.

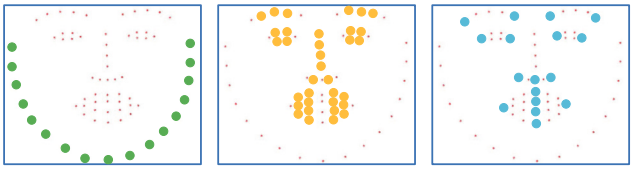


Figure 5: Divide 68 facial landmarks into three categories (𝔞, 𝔟 and 𝔠) according to their discriminations.

	common	improvement	challenge	improvement
BASE-0-0-1	2.90	12.76%	5.25	15.05%
HSLE-0-0-1	2.53		4.46	
BASE-0-1-0	3.05	15.41%	5.33	13.32%
HSLE-0-1-0	2.58		4.62	
BASE-1-0-0	4.12	10.68%	6.81	6.46%
HSLE-1-0-0	3.68		6.37	

Table 2: Weighted mean iod-norm Error (%) of proposed model against the baseline model. "BASE" indicates the baseline model and "HSLE" indicates the proposed model. BASE/HSLE-a-b-c denote the weights a, b and c assigned to categories 𝔞, 𝔟 and 𝔠 respectively. 8-stacked Hourglass model is selected as the baseline model in this experiment.

4.3. Qualitative Results

Some qualitative results on the 300W dataset are presented in Figure 6. Images with different color borders are results derived from one baseline and the proposed HSLE model respectively. The results from images with unconstrained situations demonstrate that the baseline facial landmark detector becomes much more robust by trained jointly with HSLE in an end-to-end fashion, due to the structural constraints propagated from the HSLE model.

5. Discussion

In order to make a clearer study of the impact of the HSLE model on the overall performance, we further conducted supplementary experiments on AFLW dataset [34]. For evaluation, the AFLW-Full protocol has been used [67]. As shown in Table 3, our method can achieve consistent improvements.

Line	Model	MEAN IOD-NORM ERROR
BASELINE: 4-stack hourglass network		
1	BASE-S4	7.84
2	TREE-S4	7.64
3	HSLE-S4-E5	**7.52**
BASELINE: 8-stack hourglass network		
4	BASE-S8	7.64
5	TREE-S8	7.56
6	HSLE-S8-E17	**7.24**
HSLE with different settings		
7	HSLE-S4-E4	7.60
8	HSLE-S4-E5	7.52
9	HSLE-S4-E8	7.47
10	HSLE-S4-E11	7.51
11	HSLE-S4-E14	7.52

Table 3: Quantitative results on AFLW dataset. "BASE" is the baseline model Stacked Hourglass. "TREE" is a manually designed bi-directional tree structured model proposed by [9]. "HSLE" is the proposed model. "S4/S8" mean stacking 4/8 hourglass modules respectively. " EX" means "X" ensembles are used to hierarchically depict the structural constraints.

Experiments with different baselines. We have reported results with 3 different state-of-the-art baselines in Table 1. Experimental results verify the effectiveness of our method. We further conducted experiments using the Stacked Hourglass as the baseline but stacking different number of hourglass modules. Line 1/3 and Line 4/6 in Table 3 show the proposed HSLE model consistently outperforms the baselines stacked 4/8 hourglass modules respectively, which also verifies that explicitly applying structural constraints as proposed outperforms implicitly incorporating it by stacking multiple hourglasses.

Compare with manually designed structural constraints methods. The bottleneck of manually structural constraints is the difficulty to be applied to a large number of landmarks, therefore manually designed structural constraints, such as [9], are not suitable for the 300W dataset (annotated with 68 landmarks per image). As for the AFLW-Full dataset with 19 landmarks per image, we re-implemented a manually designed 19-nodes bi-directional tree structured model(Figure 7(b)) refer to [9]. Line 2/3 and Line 5/6 in Table 3 and Figure 7(a) show our proposed model consistently outperforms the manually designed structural constraints, which demonstrates the auto-

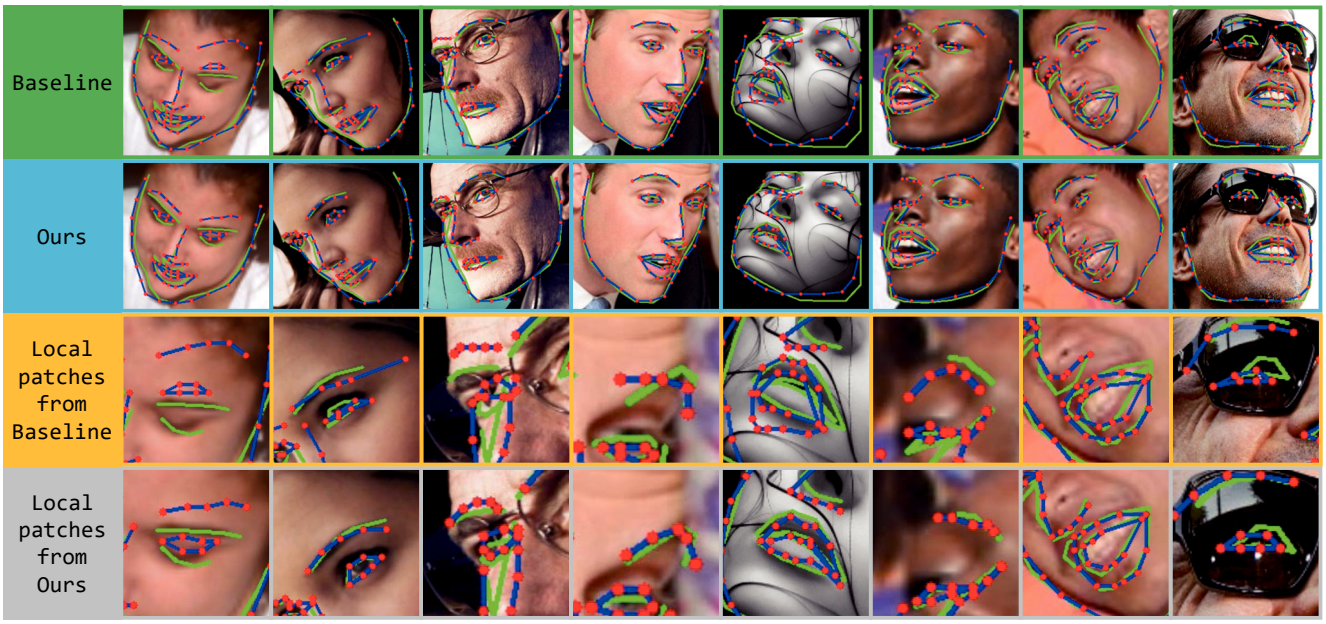
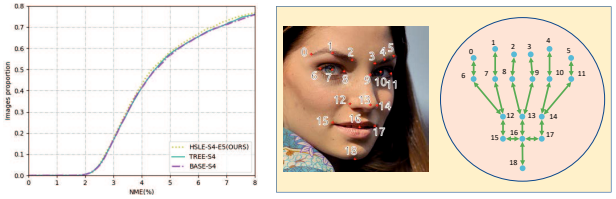


Figure 6: Qualitative results on the 300W. Images with green borders (row 1) are results derived from a baseline model directly. Images with blue borders (row 2) are results derived from our HSLE model. Images with yellow borders (row 3) and images with gray borders (row 4) are local enlarged image patches from the baseline and ours respectively. BLUE curves and RED dots are predictions. GREEN curves are groundtruth.

matically learned hierarchical structural constraints by our method are not only much more suitable for a large number of landmarks, but also more robust than the manually designed constraints.



(a) (b)

Figure 7: (a) CED curves on AFLW-Full dataset. (b) The diagram of the structure and the information flow of the re-implemented manually structured model.

HSLE with different settings. We evaluated our model using different $x - (n, t, m)$ settings $(4/5/8/11/14 - (6, 4, 3))$ on AFLW dataset, and the results show our model always improves the baseline. The complexity of our model grows with the increase of x, n and the decrease of t, and the continuing increase of parameter size would not further improve the performance, as shown in Line 7∼11, Table 3. Moreover, to depict the structural constraints for 68 landmarks, when applying the $17 - (6, 4, 3)$ setting, the number of parameters has increased by 35802, there are 221 substructures for passing information. If a fully connected graph, or the so called $1 - (68, 3, 3)$ setting, is applied, the number of parameters will increase by 8118792, and there will be 50116 substructures for passing information. For a trade-off between accuracy and efficiency, we mainly reported the results with $5 - (6, 4, 3)$ setting for 19 landmarks and $17 - (6, 4, 3)$ setting for 68 landmarks.

6. Conclusion

In this paper, we present a Hierarchical Structured Landmark Ensemble (HSLE) model for learning robust facial landmark detection. Due to the structural constraints propagated from the HSLE, the baseline facial landmark detectors consistently become more robust by trained jointly with the HSLE in an end-to-end fashion. The effectiveness of our idea has been verified by extensive experiments, indicates that facial landmark detection can be more robust via learning from hierarchical structural constraints.

Compared with the baseline model, the runtime of the proposed model for inference (68 landmarks) has increased by about 36ms on Intel i7-9700K (3.60GHz ×8) CPU and Nvidia GeForce GTX 1080Ti (11GB) GPU.

Acknowledgment

This work was supported in part by National Science Foundation grant IIS-1619078, IIS-1815561, the Army Research Office ARO W911NF-16-1-0138, the National Science Foundation for Young Scientists of China grant 61806081 and the China Postdoctoral Science Foundation grant 2018M632858.

References

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org. 6

[2] Akshay Asthana, Stefanos Zafeiriou, Shiyang Cheng, and Maja Pantic. Robust discriminative response map fitting with constrained local models. In *CVPR*, 2013. 2

[3] Akshay Asthana, Stefanos Zafeiriou, Shiyang Cheng, and Maja Pantic. Incremental face alignment in the wild. In *CVPR*, 2014. 2

[4] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. Continuous conditional neural fields for structured regression. In *ECCV*, 2014. 6

[5] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. Openface 2.0: Facial behavior analysis toolkit. In *FG*, 2018. 2

[6] Xavier P Burgos-Artizzu, Pietro Perona, and Piotr Dollár. Robust face landmark estimation under occlusion. In *ICCV*, 2013. 2

[7] Xudong Cao, Yichen Wei, Fang Wen, and Jian Sun. Face alignment by explicit shape regression. *International Journal of Computer Vision*, 107(2):177–190, 2014. 2

[8] Dong Chen, Shaoqing Ren, Yichen Wei, Xudong Cao, and Jian Sun. Joint cascade face detection and alignment. In *ECCV*, 2014. 2

[9] Xiao Chu, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. Structured feature learning for pose estimation. In *CVPR*, 2016. 2, 4, 7

[10] Timothy F Cootes, ER Baldock, and J Graham. An introduction to active shape models. *Image Processing and Analysis*, pages 223–248, 2000. 2

[11] Timothy F Cootes, Gareth J Edwards, and Christopher J Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (6):681–685, 2001. 2

[12] Timothy F Cootes and Christopher J Taylor. Active shape models—'smart snakes'. In *BMVC*. 1992. 2

[13] Timothy F Cootes, Christopher J Taylor, David H Cooper, and Jim Graham. Active shape models-their training and application. *Computer Vision and Image Understanding*, 61(1):38–59, 1995. 2

[14] David Cristinacce and Timothy F Cootes. Feature detection and tracking with constrained local models. In *BMVC*, 2006. 2

[15] Shengyang Dai, Ming Yang, Ying Wu, and Aggelos Katsaggelos. Detector ensemble. In *CVPR*, 2007. 3, 5

[16] Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi Yang. Style aggregated network for facial landmark detection. In *CVPR*, 2018. 6

[17] Pengfei Dou, Shishir K Shah, and Ioannis A Kakadiaris. End-to-end 3d face reconstruction with deep neural networks. In *CVPR*, 2017. 1

[18] Gareth J Edwards, Timothy F Cootes, and Christopher J Taylor. Face recognition using active appearance models. In *ECCV*, 1998. 2

[19] Zhen-Hua Feng, Josef Kittler, William Christmas, Patrik Huber, and Xiao-Jun Wu. Dynamic attention-controlled cascaded shape regression exploiting training data augmentation and fuzzy-set sample weighting. In *CVPR*, 2017. 2

[20] Golnaz Ghiasi and Charless C Fowlkes. Occlusion coherence: Localizing occluded faces with a hierarchical deformable part model. In *CVPR*, 2014. 2

[21] Sina Honari, Jason Yosinski, Pascal Vincent, and Christopher Pal. Recombinator networks: Learning coarse-to-fine feature aggregation. In *CVPR*, 2016. 6

[22] Gee-Sern Hsu, Kai-Hsiang Chang, and Shih-Chieh Huang. Regressive tree structured model for facial landmark localization. In *ICCV*, 2015. 2

[23] Fatih Kahraman, Muhittin Gokmen, Sune Darkner, and Rasmus Larsen. An active illumination and appearance (aia) model for face alignment. In *CVPR*, 2007. 2

[24] Michal Kawulok, Emre Celebi, and Bogdan Smolka. *Advances in face detection and facial image analysis*. Springer, 2016. 1

[25] Vahid Kazemi and Josephine Sullivan. One millisecond face alignment with an ensemble of regression trees. In *CVPR*, 2014. 2

[26] Marek Kowalski, Jacek Naruniec, and Tomasz Trzcinski. Deep alignment network: A convolutional neural network for robust face alignment. In *CVPR Workshops*, 2017. 6

[27] Amit Kumar and Rama Chellappa. Disentangling 3d pose in a dendritic cnn for unconstrained 2d face alignment. In *CVPR*, 2018. 6

[28] Neeraj Kumar, Peter Belhumeur, and Shree Nayar. Facetracer: A search engine for large collections of images with faces. In *ECCV*, 2008. 2

[29] Lin Liang, Rong Xiao, Fang Wen, and Jian Sun. Face alignment via component-based discriminative search. In *ECCV*, 2008. 2

[30] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *CVPR*, 2017. 1

[31] Xiaoming Liu. Generic face alignment using boosted appearance model. In *CVPR*, 2007. 2

[32] Stuart Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982. 5

[33] Jiang-Jing Lv, Xiaohu Shao, Junliang Xing, Cheng Cheng, Xi Zhou, et al. A deep regression architecture with two-stage re-initialization for high performance facial landmark detection. In *CVPR*, 2017. 2, 6

[34] Peter M. Roth Martin Koestinger, Paul Wohlhart and Horst Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization.

In *Proc. First IEEE International Workshop on Benchmarking Facial Image Analysis Technologies*, 2011. 2, 7

[35] Iacopo Masi, Stephen Rawls, Gérard Medioni, and Prem Natarajan. Pose-aware face recognition in the wild. In *CVPR*, 2016. 1

[36] Daniel Merget, Matthias Rock, and Gerhard Rigoll. Robust facial landmark detection via a fully-convolutional local-global context network. In *CVPR*, 2018. 1, 2, 6

[37] Stephen Milborrow and Fred Nicolls. Locating facial features with an extended active shape model. In *ECCV*, 2008. 2

[38] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016. 3, 6

[39] Shaoqing Ren, Xudong Cao, Yichen Wei, and Jian Sun. Face alignment at 3000 fps via regressing local binary features. In *CVPR*, 2014. 2

[40] Christos Sagonas, Epameinondas Antonakos, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: Database and results. *Image and Vision Computing*, 47:3–18, 2016. 2, 5, 6

[41] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *ICCV Workshops*, 2013. 2, 5, 6

[42] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. A semi-automatic methodology for facial landmark annotation. In *CVPR Workshops*, 2013. 2, 5, 6

[43] Jason Saragih and Roland Goecke. A nonlinear discriminative approach to aam fitting. In *ICCV*, 2007. 2

[44] Jason M Saragih, Simon Lucey, and Jeffrey F Cohn. Face alignment through subspace constrained mean-shifts. In *ICCV*, 2009. 2

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

[46] Patrick Sauer, Timothy F Cootes, and Christopher J Taylor. Accurate regression procedures for active appearance models. In *BMVC*, 2011. 2

[47] Brandon M Smith, Jonathan Brandt, Zhe Lin, and Li Zhang. Nonparametric context modeling of local appearance for pose-and expression-robust facial landmark localization. In *CVPR*, 2014. 2

[48] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deep convolutional network cascade for facial point detection. In *CVPR*, 2013. 2

[49] Zhiqiang Tang, Xi Peng, Shijie Geng, Lingfei Wu, Shaoting Zhang, and Dimitris Metaxas. Quantized densely connected u-nets for efficient landmark localization. In *ECCV*, 2018. 1, 2, 6

[50] George Trigeorgis, Patrick Snape, Mihalis A Nicolaou, Epameinondas Antonakos, and Stefanos Zafeiriou. Mnemonic descent method: A recurrent process applied for end-to-end face alignment. In *CVPR*, 2016. 2, 6

[51] Roberto Valle and M José. A deeply-initialized coarse-to-fine ensemble of regression trees for face alignment. In *ECCV*, 2018. 1, 2, 6

[52] Wayne Wu, Chen Qian, Shuo Yang, Quan Wang, Yici Cai, and Qiang Zhou. Look at boundary: A boundary-aware face alignment algorithm. In *CVPR*, 2018. 1, 2, 6

[53] Wenyan Wu and Shuo Yang. Leveraging intra and inter-dataset variations for robust face alignment. In *CVPR Workshops*, 2017. 6

[54] Yue Wu and Qiang Ji. Constrained joint cascade regression framework for simultaneous facial action unit recognition and facial landmark detection. In *CVPR*, 2016. 2

[55] Shengtao Xiao, Jiashi Feng, Junliang Xing, Hanjiang Lai, Shuicheng Yan, and Ashraf Kassim. Robust facial landmark detection via recurrent attentive-refinement networks. In *ECCV*, 2016. 2, 6

[56] Xuehan Xiong and Fernando De la Torre. Global supervised descent method. In *CVPR*, 2015. 2

[57] Jing Yang, Qingshan Liu, and Kaihua Zhang. Stacked hourglass network for robust facial landmark localisation. In *CVPR Workshops*, 2017. 1, 2, 6

[58] Xiang Yu, Junzhou Huang, Shaoting Zhang, Wang Yan, and Dimitris N Metaxas. Pose-free facial landmark fitting via optimized part mixtures and cascaded deformable shape model. In *ICCV*, 2013. 2

[59] Xiang Yu, Feng Zhou, and Manmohan Chandraker. Deep deformation network for object landmark localization. In *ECCV*, 2016. 2

[60] Amir Zadeh, Yao Chong Lim, Tadas Baltrusaitis, and Louis-Philippe Morency. Convolutional experts constrained local model for 3d facial landmark detection. In *ICCV Workshops*, 2017. 2, 6

[61] Jie Zhang, Shiguang Shan, Meina Kan, and Xilin Chen. Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment. In *ECCV*, 2014. 2

[62] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Facial landmark detection by deep multi-task learning. In *ECCV*, 2014. 2

[63] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Learning deep representation for face alignment with auxiliary attributes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(5):918–930, 2016. 2, 6

[64] Xiangyun Zhao, Xiaodan Liang, Luoqi Liu, Teng Li, Yugang Han, Nuno Vasconcelos, and Shuicheng Yan. Peak-piloted deep network for facial expression recognition. In *ECCV*, 2016. 1

[65] Feng Zhou, Jonathan Brandt, and Zhe Lin. Exemplar-based graph matching for robust facial landmark localization. In *ICCV*, 2013. 2

[66] Shizhan Zhu, Cheng Li, Chen Change Loy, and Xiaoou Tang. Face alignment by coarse-to-fine shape searching. In *CVPR*, 2015. 2, 6

[67] Shizhan Zhu, Cheng Li, Chen-Change Loy, and Xiaoou Tang. Unconstrained face alignment via cascaded compositional learning. In *CVPR*, 2016. 2, 5, 7