UGLLI Face Alignment: Estimating Uncertainty with Gaussian Log-Likelihood Loss*

Abhinav Kumar[†,1], Tim K. Marks[†,2], Wenxuan Mou[†,3], Chen Feng[4], Xiaoming Liu[5]
abhinav.kumar@utah.edu, tmarks@merl.com, wenxuan.mou@manchester.ac.uk, cfeng@nyu.edu, liuxm@cse.msu.edu
[1]University of Utah, [2]Mitsubishi Electric Research Labs (MERL), [3]University of Manchester, [4]New York University, [5]Michigan State University

Abstract

Modern face alignment methods have become quite accurate at predicting the locations of facial landmarks, but they do not typically estimate the uncertainty of their predicted locations. In this paper, we present a novel framework for jointly predicting facial landmark locations and the associated uncertainties, modeled as 2D Gaussian distributions, using Gaussian log-likelihood loss. Not only does our joint estimation of uncertainty and landmark locations yield state-of-the-art estimates of the uncertainty of predicted landmark locations, but it also yields state-of-the-art estimates for the landmark locations (face alignment). Our method's estimates of the uncertainty of landmarks' predicted locations could be used to automatically identify input images on which face alignment fails, which can be critical for downstream tasks.

1. Introduction

Face alignment is the task of estimating the pixel locations of a set of predefined facial landmark points (e.g., eye and mouth corners) in an input face image. Most methods for face alignment focus on accurately estimating the facial landmark locations [35, 37] without estimating the uncertainty of these location estimates. Estimating uncertainty not only enables the identification of failure cases in real-world scenarios, but also allows downstream tasks to be adjusted either automatically or manually based upon the estimated uncertainty. Therefore, while it is certainly important to improve the accuracy of face alignment systems, it is equally important to predict their uncertainty.

Our contributions can be summarized as follows. This is the first work to introduce the concept of parametric uncertainty estimation for image-based landmark estimation (and for face alignment in particular). To estimate landmark locations in a differentiable manner, we do not select the location of the maximum (argmax) of each landmark's heatmap, but instead propose to use the spatial mean of the positive elements of each heatmap. To estimate uncertainty, we add a Cholesky Estimator Network (CEN) branch to estimate the covariance matrix of a Gaussian uncertainty distribution. We combine these estimates using a Gaussian log-likelihood loss that enables simultaneous estimation of landmark locations and their uncertainty. This joint estimation, which we call Uncertainty with Gaussian Log-LIkelihood (UGLLI), yields state-of-the-art results for both uncertainty estimation and facial landmark localization. Moreover, we find that the choice of methods for calculating mean and covariance is crucial. Landmark positions are best obtained by taking a spatial mean over the heatmaps, rather than by direct regression. In contrast, the uncertainty covariance matrices are best obtained by direct regression, not from the heatmaps.

Figure 1: Results of our joint face alignment and uncertainty estimation on three test images. Ground truth (green) and predicted (yellow) landmark locations are shown. The estimated uncertainty of the predicted location of each landmark is shown in blue (Gaussian error ellipse for Mahalanobis distance 1). Landmarks that are occluded (e.g., by the hand in center image) tend to have larger uncertainty.

2. Related Work

Early methods for face alignment were based on Active Shape Models (ASM) and Active Appearance Models (AAM) [9, 34], as well as their multi-view and multi-camera variations [10, 1]. Subsequently, direct regression methods (which map directly from the features extracted at facial landmark locations to the face shape or landmark locations) became popular due to their excellent performance.

*The first workshop on Statistical Deep Learning for Computer Vision, in *Seoul, Korea*, 2019. Copyright by Author(s).

[†]Equal Contributions

Of these, tree-based regression methods [24, 14] proved particularly fast, and the subsequent cascaded regression methods [11, 33, 32] improved accuracy.

Recent approaches [41, 42, 38, 5, 30] are based on deep learning. The currently most successful deep methods, such as stacked hourglass networks [38, 5] and densely connected U-nets (DU-Net) [30], use a cascade of deep networks, an architecture that was originally developed for human body pose estimation [20]. These models [20, 5, 29, 30] are trained using the ℓ_2 distance between the predicted heatmap for each landmark and a proxy ground-truth heatmap that is generated by placing a symmetric Gaussian distribution with small fixed variance at the ground-truth landmark location. They then infer landmark locations using the argmax of each predicted heatmap. Indirect inference through a predicted heatmap offers several advantages over direct prediction [2].

However, this approach has at least two disadvantages. First, it introduces quantization errors during inference, since the heatmap's argmax can only be determined to the nearest pixel [18, 21, 28]. To achieve sub-pixel localization for body pose estimation, [18] replaces the argmax with a spatial mean over the softmax. In a different approach to sub-pixel localization, which is applied to videos, [28] samples two additional points adjacent to the max of the heatmap to estimate a local peak. Second, using a symmetric Gaussian proxy ground-truth heatmap makes it difficult to infer uncertainties [7]. To estimate face alignment uncertainty, [7] uses a non-parametric approach: a kernel density network obtained by convolving the heatmaps with a fixed symmetric Gaussian kernel.

Finally, there are other methods for regression with uncertainty that have not been applied to landmark regression. The mixture density network (MDN) [4] estimates parameters of Gaussian distributions in a mixture, though typically such Gaussians are one-dimensional or have diagonal covariance matrices. Also for 1-D regression, [16] uses ensembles and adversarial training to produce two outputs, one for prediction and one for uncertainty.

3. Proposed Method

Figure 2 shows an overview of our UGLLI Face Alignment. The input RGB face image is passed through a DU-Net [30] architecture, to which we add two additional components branching from each hourglass (each U-net). The first new component is a *mean estimator*, which computes the estimated location of each landmark as the weighted spatial mean of the positive elements of the corresponding heatmap. The second new component, the *Cholesky Estimator Network* (CEN), emerges from the bottleneck layer of each hourglass (CEN weights are shared across all hourglasses). The CEN estimates the Cholesky coefficients of the covariance matrix of a 2D Gaussian prob-

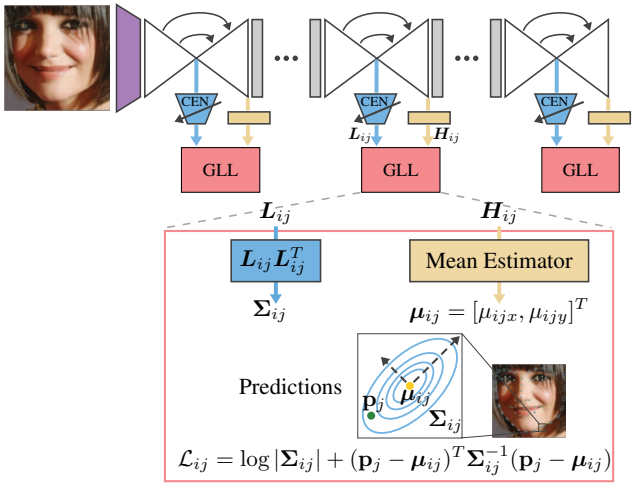

Figure 2: An UGLLI overview. From each hourglass, we propose a shared Cholesky Estimator Network (CEN) that is appended to the bottleneck layer and a mean estimator that is applied to the heatmap. The figure shows the joint landmark prediction and uncertainty estimation performed for each hourglass i and each landmark j. The ground-truth (labeled) landmark location is represented by $\mathbf{p}_j$.

ability distribution for each landmark location. For each hourglass i and each landmark j, the landmark's location estimate $\boldsymbol{\mu}_{ij}$ and estimated covariance matrix $\boldsymbol{\Sigma}_{ij}$ are tied together by a Gaussian log-likelihood (GLL) loss function $\mathcal{L}_{ij}$, which enables end-to-end optimization of the entire face alignment and uncertainty estimation framework. Rather than the argmax of the heatmap, we choose a mean estimator for the heatmap that is differentiable and enables sub-pixel accuracy: the weighted spatial mean of the heatmap's positive elements. Unlike the non-parametric model of [7], our uncertainty prediction method is parametric: we directly estimate the parameters of a single Gaussian distribution. Furthermore, our method does not constrain the Gaussian covariance matrix to be diagonal.

3.1. Mean Estimator

Let $\boldsymbol{H}_{ij}(x, y)$ denote the value at pixel location (x, y) of the jth landmark's heatmap from the ith hourglass. Then the landmark's location estimate $\boldsymbol{\mu}_{ij} = [\mu_{ijx}, \mu_{ijy}]^T$ is given by first post-processing the pixels of the heatmap $\boldsymbol{H}_{ij}$ with a function σ, then taking the weighted spatial mean of the result. We considered three different functions for σ: the ReLU function (eliminates the negative values), the softmax function (making the mean estimator a soft-argmax of the heatmap [6, 39, 18, 13]), and a temperature-controlled softmax function (which, depending on the temperature setting, provides a continuum of softmax functions that range from a "hard" argmax to the uniform distribution). As explained in Section 5, choosing σ to be the ReLU function yields the simplest and best mean estimator. Estimating the

landmark location from the positive heatmap by taking the spatial mean can be considered as the maximum likelihood estimate (MLE) of the mean of a 2D Gaussian distribution that is sampled on a regular grid, where the heatmap values represent the frequency of samples at each grid location.

3.2. Gaussian Log-Likelihood Loss

UGLLI uses heatmaps for estimating landmarks' locations, but not for estimating their uncertainty. We experimented with several methods for computing a covariance matrix directly from a heatmap, but none was accurate enough. We believe the reason is that many images have some landmarks that can be located very accurately, and thus the uncertainty of these locations is very small (a fraction of a pixel) in at least one direction. In current heatmap-based networks, however, the resolution of the heatmap is too low to accurately represent such small uncertainties.

Cholesky Estimator Network (CEN) We represent the uncertainty of each landmark location as a Gaussian distribution with covariance matrix $\mathbf{\Sigma}_{ij}$, a 2×2 symmetric positive definite matrix. $\mathbf{\Sigma}_{ij}$ has three degrees of freedom that are captured by its Cholesky decomposition: a lower-triangular matrix $\boldsymbol{L}_{ij}$ such that $\mathbf{\Sigma}_{ij} = \boldsymbol{L}_{ij}\boldsymbol{L}_{ij}^T$. To estimate the elements of $\boldsymbol{L}_{ij}$, we append a Cholesky Estimator Network (CEN) to the bottleneck of each hourglass. The CEN is a fully connected linear layer whose input is the bottleneck of the hourglass ($128 \times 4 \times 4 = 2048$ dimensions) and output is a vector of $68 \times 3 = 224$ dimensions. $\boldsymbol{L}_{ij}$ must have positive diagonal elements to be the Cholesky decomposition of a covariance matrix, so we pass the corresponding entries of the output through an ELU activation function [8] to which we add a constant to ensure the output is always positive (asymptote is negative x-axis).

Given the predicted Gaussian distribution for a landmark of an input image at each hourglass i, the likelihood that the landmark j is at image location $\mathbf{p}_j$ is given by:

$$P(\mathbf{p}_j|\boldsymbol{\mu}_{ij},\boldsymbol{L}_{ij}) = \frac{\exp\left(-\frac{1}{2}(\mathbf{p}_j - \boldsymbol{\mu}_{ij})^T \mathbf{\Sigma}_{ij}^{-1}(\mathbf{p}_j - \boldsymbol{\mu}_{ij})\right)}{2\pi\sqrt{|\mathbf{\Sigma}_{ij}|}} \quad (1)$$

where $\mathbf{\Sigma}_{ij} = \boldsymbol{L}_{ij}\boldsymbol{L}_{ij}^T$. Thus, for each landmark in every input image, the network outputs a Gaussian distribution (parameterized by $\boldsymbol{\mu}_{ij}$ and $\boldsymbol{L}_{ij}$). The goal of training is for the network to learn a mapping from input images to Gaussian distributions, such that the likelihood of the groundtruth landmark locations (over all landmarks and all training images) is as large as possible. Maximizing the likelihood (1) is equivalent to minimizing the negative log likelihood, so we use the negative log likelihood as our loss function. Our loss function $\mathcal{L}_{ij}$ at each hourglass i for the landmark j can be expressed as the sum of two terms, T_1 and T_2:

$$\mathcal{L}_{ij} = \underbrace{\log|\mathbf{\Sigma}_{ij}|}_{T_1} + \underbrace{(\mathbf{p}_j - \boldsymbol{\mu}_{ij})^T \mathbf{\Sigma}_{ij}^{-1}(\mathbf{p}_j - \boldsymbol{\mu}_{ij})}_{T_2}. \quad (2)$$

In (2), T_2 is the squared Mahalanobis distance, while T_1 serves as a regularization or prior term that ensures that the Gaussian uncertainty distribution does not get too large. Note that if $\mathbf{\Sigma}_{ij}$ is the identity matrix, (2) reduces to the standard ℓ_2 distance. The objective for a single hourglass is obtained by averaging the losses across all the landmarks $j = 1, ..., N_p$, and the total loss $\mathcal{L}$ for each input image is a weighted sum of the losses of every hourglass i:

$$\mathcal{L} = \sum_{i=1}^{K} \lambda_i \mathcal{L}_i\,, \quad \text{where} \quad \mathcal{L}_i = \frac{1}{N_p}\sum_{j=1}^{N_p} \mathcal{L}_{ij}\,. \quad (3)$$

At test time, each landmark's estimated mean and covariance matrix are derived from the final hourglass K.

4. Experiments

4.1. Data Splits

We use the 300-W [26, 25, 27] dataset for training, and the 300-W and Menpo [40, 31] datasets for evaluation. Every face in each dataset is labeled with the locations of 68 landmarks. The images are cropped using the detector bounding boxes provided by the 300-W challenge and resized to 256×256. Images with no detector bounding box are initialized by adding 5% uniform noise to the location of each edge of the tight bounding box around the landmarks, as in [5]. There are two commonly used train/test splits of the 300-W dataset; we evaluate our method on both.

Split 1 The train set includes 3148 images: the training images from Helen [17] and LFPW [3], and all AFW [23] images. The full test set has 689 images: the test images from Helen and LFPW (*common* subset), and all IBUG images (*challenge* subset). As in [30], training images are augmented randomly using scaling, rotation, and color jittering.

Split 2 The train set includes all 3837 training and test images from Helen, LFPW, AFW, and IBUG. The test set has 600 images, known as 300-W Indoor and Outdoor. Training images are augmented randomly using scaling, rotation, color jittering, and random occlusion, as in [5].

4.2. Training

We modified the PyTorch [22] code for DU-Net [30] provided by the authors and initialized using their pre-trained model (Split 1). The RMSprop optimizer is used as in [5, 30], with batch size 24. We train for 40 epochs: 20 with learning rate 10^{-4}, followed by 20 with learning rate 2×10^{-5}. All hourglasses have equal weights $\lambda_i = 1$ in (3). On a 12 GB GeForce GTX Titan-X GPU, training takes ∼2 hours (Split 1), and inference time per image is 17ms.

4.3. Evaluation metrics

Normalized Mean Error (NME). The NME for a single face image is defined as:

$$\text{NME}\ (\%) = \frac{1}{N_p}\sum_{j=1}^{N_p} \frac{||\mathbf{p}_j - \boldsymbol{\mu}_{Kj}||_2}{d} \times 100, \quad (4)$$

Table 1: NME comparison between our proposed method and the state-of-the-art methods on the 300-W Common, Challenge, and Full datasets (Split 1).

	$\text{NME}_{\text{inter-ocular}}$ (%) (↓)		
	Common	Challenge	Full
DSRN [19]	4.12	9.68	5.21
CPM+SBR [13]	3.28	7.58	4.10
SAN [12]	3.34	6.60	3.98
DAN [15]	3.19	5.24	3.59
DU-Net [30] (Public code)	2.97	5.53	3.47
UGLLI (Ours)	**2.78**	**5.08**	**3.23**

Table 2: NME and AUC comparison between our proposed method and the state-of-the-art methods on the 300-W Test (Split 2) and Menpo datasets.

	NME_{box} (%) (↓)		$\text{AUC}^{7}_{\text{box}}$ (%) (↑)	
	300-W	Menpo	300-W	Menpo
2D-FAN [5]	2.56	2.32	66.90	67.40
KDN-Gaussian [7]	2.49	2.26	67.30	68.40
UGLLI (Ours)	**2.24**	**2.20**	**68.27**	**69.85**

where $\mathbf{p}_j$ and $\boldsymbol{\mu}_{Kj}$ respectively denote the ground truth and predicted location of landmark j from the final hourglass K. Several variations of the normalizing term d have been used in the literature. $\text{NME}_{\text{inter-ocular}}$ [26, 15, 30] sets d to the distance between the outer corners of the two eyes, while NME_{box} [40, 5, 7] sets d to the geometric mean of the width and height of the provided ground-truth bounding box $\left(\sqrt{w_{\text{bbox}} \cdot h_{\text{bbox}}}\right)$. If a ground-truth bounding box is not provided, the tight bounding box of the landmarks is used [5, 7]. In each table, we report our results using the same error metric as the previous methods compared.

Area Under the Curve (AUC). To compute the AUC, first plot the cumulative distribution of the fraction of test images whose NME (%) is less than or equal to the value on the horizontal axis. The AUC for a test set is then computed as the area under that curve, up to a cutoff NME value. We report AUC with cutoff 7%, as in [5, 7].

5. Results

Evaluation of Landmark Location Prediction The face alignment results for 300-W Split 1 and Split 2 are summarized in Table 1 and Table 2, respectively. Table 2 also shows the results of our model (trained on Split 2) on the Menpo dataset (6679 frontal training images with 68 landmarks), as in [5, 7]. The results in both tables show that UGLLI significantly outperforms the state of the art.

Evaluation of Uncertainty Prediction We use the fourth root of the determinant of the covariance matrix as a scalar measure of estimated uncertainty: $|\boldsymbol{\Sigma}_{Kj}|^{1/4}$. Figure 3 plots the normalized estimated uncertainty vs. the normalized landmark location error on all 300-W Test images (Split 2). On the left, each point represents one landmark in one image. On the right, each point represents the average across all landmarks in one image. The Pearson correlation coefficients over each plot show that our predicted uncertainties are strongly correlated with the actual errors, and ours outperform the uncertainty estimates of Chen et al. [7].

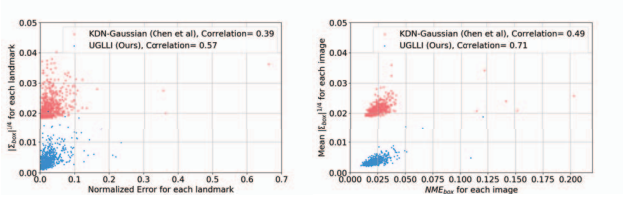

Figure 3: *Left:* Predicted uncertainty vs. actual error for each landmark. *Right:* Predicted uncertainty vs. actual error for each image (averaged across all landmarks in image).

Ablation Studies Table 3 compares modifications of our approach on Split 2. Table 3 shows that computing the loss only on the last hourglass (HG) performs worse than computing loss on all hourglasses, because of the vanishing gradient problem [36]. Moreover, UGLLI's Gaussian Log-Likelihood (GLL) loss outperforms using MSE loss on the landmark locations (equivalent to setting all $\boldsymbol{\Sigma}_{ij} = \mathbf{I}$). Regarding the mean estimator, direct regression (Direct) from each hourglass bottleneck to output the mean (rather than using the heatmap) is ineffective, consistent with previous observations that neural networks have difficulty predicting continuous real values [2, 21]. As described in Section 3.1, in addition to the ReLU function, we compared two other functions for σ: soft-argmax (s-amax), and a temperature-scaled soft-argmax (τ-s-amax). Results for temperature-scaled soft-argmax and ReLU are essentially tied, but the former is more complicated and requires tuning the temperature parameter, so we chose ReLU for our UGLLI model.

Table 3: Ablation studies on 300-W Test and Menpo datasets using our method trained on 300-W Split 2.

Change from UGLLI model:		NME_{box} (%)		$\text{AUC}^{7}_{\text{box}}$ (%)	
Changed	From → To	300-W	Menpo	300-W	Menpo
Supervision	All HGs → Last HG	2.47	2.34	65.07	68.06
Loss	GLL → MSE	2.40	2.28	66.05	68.70
Mean Estimator	Heatmap → Direct	4.95	4.60	34.45	42.05
	ReLU → s-amax	3.01	2.81	57.44	61.30
	ReLU → τ-s-amax	2.26	**2.19**	67.97	**69.94**
—	UGLLI (Ours)	**2.24**	2.20	**68.27**	69.85

6. Conclusion

In this paper, we present UGLLI, a novel end-to-end trainable framework for face alignment and uncertainty estimation using a Gaussian log-likelihood loss. The uncertainty of each predicted landmark location is estimated as a 2D Gaussian distribution, and the determinant of this covariance matrix is used as a scalar measure of uncertainty. The joint estimation of landmark location and uncertainty not only provides state-of-the-art uncertainty measures but also yields state-of-the-art estimates for the landmark locations. Future work includes application of this framework to uncertainty estimation in other landmark regression tasks, such as human body pose estimation, as well as using estimated uncertainties to selectively improve the predictions.

References

[1] A. Asthana, T. Marks, M. Jones, K. Tieu, and R. M.V. Fully automatic pose-invariant face recognition via 3d pose normalization. In *ICCV*, 2011.

[2] V. Belagiannis and A. Zisserman. Recurrent human pose estimation. In *FG*, 2017.

[3] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar. Localizing parts of faces using a consensus of exemplars. In *CVPR*, 2011.

[4] C. M. Bishop. Mixture density networks. Technical report, Citeseer, 1994.

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

[6] O. Chapelle and M. Wu. Gradient descent optimization of smoothed information retrieval metrics. *Information retrieval*, 2010.

[7] L. Chen and Q. Ji. Kernel density network for quantifying regression uncertainty in face alignment. In *NeurIPS Workshops*, 2018.

[8] D.-A. Clevert, T. Unterthiner, and S. Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). In *ICLR*, 2016.

[9] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *TPAMI*, 2001.

[10] T. F. Cootes, G. V. Wheeler, K. N. Walker, and C. J. Taylor. View-based active appearance models. *Image and Vision Computing*, 2002.

[11] P. Dollár, P. Welinder, and P. Perona. Cascaded pose regression. In *CVPR*, 2010.

[12] X. Dong, Y. Yan, W. Ouyang, and Y. Yang. Style aggregated network for facial landmark detection. In *CVPR*, 2018.

[13] X. Dong, S.-I. Yu, X. Weng, S.-E. Wei, Y. Yang, and Y. Sheikh. Supervision-by-registration: An unsupervised approach to improve the precision of facial landmark detectors. In *CVPR*, 2018.

[14] V. Kazemi and J. Sullivan. One millisecond face alignment with an ensemble of regression trees. In *CVPR*, 2014.

[15] M. Kowalski, J. Naruniec, and T. Trzcinski. Deep alignment network: A convolutional neural network for robust face alignment. In *CVPR Workshops*, 2017.

[16] B. Lakshminarayanan, A. Pritzel, and C. Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *NeurIPS*, 2017.

[17] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang. Interactive facial feature localization. In *ECCV*, 2012.

[18] D. C. Luvizon, D. Picard, and H. Tabia. 2d/3d pose estimation and action recognition using multitask deep learning. In *CVPR*, 2018.

[19] X. Miao, X. Zhen, X. Liu, C. Deng, V. Athitsos, and H. Huang. Direct shape regression networks for end-to-end face alignment. In *CVPR*, 2018.

[20] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016.

[21] A. Nibali, Z. He, S. Morgan, and L. Prendergast. Numerical coordinate regression with convolutional neural networks. *arXiv preprint arXiv:1801.07372*, 2018.

[22] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. In *NeurIPS Workshops*, 2017.

[23] D. Ramanan and X. Zhu. Face detection, pose estimation, and landmark localization in the wild. In *CVPR*, 2012.

[24] S. Ren, X. Cao, Y. Wei, and J. Sun. Face alignment at 3000 fps via regressing local binary features. In *CVPR*, 2014.

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

[26] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *CVPR Workshops*, 2013.

[27] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. A semi-automatic methodology for facial landmark annotation. In *CVPR workshops*, 2013.

[28] Y. Tai, Y. Liang, X. Liu, L. Duan, J. Li, C. Wang, F. Huang, and Y. Chen. Towards highly accurate and stable face alignment for high-resolution videos. In *AAAI*, 2019.

[29] Z. Tang, X. Peng, S. Geng, L. Wu, S. Zhang, and D. Metaxas. Quantized densely connected u-nets for efficient landmark localization. In *ECCV*, 2018.

[30] Z. Tang, X. Peng, K. Li, and D. Metaxas. Towards efficient u-nets: A coupled and quantized approach. *TPAMI*, 2019.

[31] G. Trigeorgis, P. Snape, M. A. Nicolaou, E. Antonakos, and S. Zafeiriou. Mnemonic descent method: A recurrent process applied for end-to-end face alignment. In *CVPR*, 2016.

[32] O. Tuzel, T. Marks, and S. Tambe. Robust face alignment using a mixture of invariant experts. In *ECCV*, 2016.

[33] G. Tzimiropoulos. Project-out cascaded regression with an application to face alignment. In *CVPR*, 2015.

[34] G. Tzimiropoulos and M. Pantic. Optimization problems for fast aam fitting in-the-wild. In *ICCV*, 2013.

[35] N. Wang, X. Gao, D. Tao, H. Yang, and X. Li. Facial feature point detection: A comprehensive survey. *Neurocomputing*, 2018.

[36] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *CVPR*, 2016.

[37] Y. Wu and Q. Ji. Facial landmark detection: A literature survey. *IJCV*, 2018.

[38] J. Yang, Q. Liu, and K. Zhang. Stacked hourglass network for robust facial landmark localisation. In *CVPR Workshops*, 2017.

[39] K. M. Yi, E. Trulls, V. Lepetit, and P. Fua. Lift: Learned invariant feature transform. In *ECCV*, 2016.

[40] S. Zafeiriou, G. Trigeorgis, G. Chrysos, J. Deng, and J. Shen. The menpo facial landmark localisation challenge: A step towards the solution. In *CVPR Workshops*, 2017.

[41] J. Zhang, S. Shan, M. Kan, and X. Chen. Coarse-to-fine auto-encoder networks (CFAN) for real-time face alignment. In *ECCV*, 2014.

[42] S. Zhu, C. Li, C. Change Loy, and X. Tang. Face alignment by coarse-to-fine shape searching. In *CVPR*, 2015.