LUVLi Face Alignment: Estimating Landmarks' Location, Uncertainty, and Visibility Likelihood

Abhinav Kumar[*,1], Tim K. Marks[*,2], Wenxuan Mou[*,3],
Ye Wang[2], Michael Jones[2], Anoop Cherian[2], Toshiaki Koike-Akino[2], Xiaoming Liu[4], Chen Feng[5]

abhinav3663@gmail.com, tmarks@merl.com, wenxuanmou@gmail.com,
[ywang, mjones, cherian, koike]@merl.com, liuxm@cse.msu.edu, cfeng@nyu.edu

[1]University of Utah, [2]Mitsubishi Electric Research Labs (MERL), [3]University of Manchester, [4]Michigan State University, [5]New York 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 nor predict whether landmarks are visible. In this paper, we present a novel framework for jointly predicting landmark locations, associated uncertainties of these predicted locations, and landmark visibilities. We model these as mixed random variables and estimate them using a deep network trained with our proposed Location, Uncertainty, and Visibility Likelihood (LUVLi) loss. In addition, we release an entirely new labeling of a large face alignment dataset with over 19,000 face images in a full range of head poses. Each face is manually labeled with the ground-truth locations of 68 landmarks, with the additional information of whether each landmark is unoccluded, self-occluded (due to extreme head poses), or externally occluded. Not only does our joint estimation yield accurate estimates of the uncertainty of predicted landmark locations, but it also yields state-of-the-art estimates for the landmark locations themselves on multiple standard face alignment datasets. Our method's estimates of the uncertainty of predicted landmark locations could be used to automatically identify input images on which face alignment fails, which can be critical for downstream tasks.

1. Introduction

Modern methods for face alignment (facial landmark localization) perform quite well most of the time, but all of them fail some percentage of the time. Unfortunately, almost all of the state-of-the-art (SOTA) methods simply output predicted landmark locations, with no assessment of whether (or how much) downstream tasks should *trust* these landmark locations. This is concerning, as face alignment is a key pre-processing step in numerous safety-critical applications, including advanced driver assistance systems (ADAS), driver monitoring, and remote measurement of vital signs [57]. As deep neural networks are notorious for producing overconfident predictions [33], similar concerns have been raised for other neural network technologies [46], and they become even more acute in the era of adversarial machine learning where adversarial images may pose a great threat to a system [14]. However, previous work in face alignment (and landmark localization in general) has largely ignored the area of uncertainty estimation.

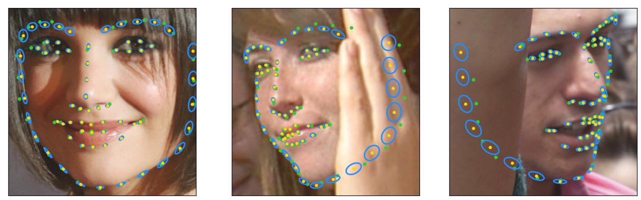


Figure 1: Results of our joint face alignment and uncertainty estimation on three test images. Ground-truth (green) and predicted (yellow) landmark locations are shown. The estimated uncertainty of the predicted location of each landmark is shown in blue (Error ellipse for Mahalanobis distance 1). Landmarks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

To address this need, we propose a method to jointly estimate facial landmark locations and a parametric probability distribution representing the uncertainty of each estimated location. Our model also jointly estimates the visibility of landmarks, which predicts whether each landmark is occluded due to extreme head pose.

We find that the choice of methods for calculating mean and covariance is crucial. Landmark locations are best obtained using heatmaps, rather than by direct regression. To estimate landmark locations in a differentiable manner using heatmaps, 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

*Equal Contributions

heatmap. Unlike landmark locations, uncertainty distribution parameters are best obtained by direct regression rather than from heatmaps. To estimate the uncertainty of the predicted locations, we add a Cholesky Estimator Network (CEN) branch to estimate the covariance matrix of a multivariate Gaussian or Laplacian probability distribution. To estimate visibility of each landmark, we add a Visibility Estimator Network (VEN). We combine these estimates using a joint loss function that we call the Location, Uncertainty and Visibility Likelihood (LUVLi) loss. Our primary goal in designing this model was to estimate uncertainty in landmark localization. In the process, not only does our method yields accurate uncertainty estimation, but it also produces SOTA landmark localization results on several face alignment datasets.

Uncertainty can be broadly classified into two categories [41]: *epistemic* uncertainty is related to a lack of knowledge about the model that generated the observed data, and *aleatoric* uncertainty is related to the noise inherent in the observations, e.g., sensor or labelling noise. The ground-truth landmark locations marked on an image by human labelers would vary across multiple labelings of an image by different human labelers (or even by the same human labeler). Furthermore, this variation will itself vary across different images and landmarks (e.g., it will vary more for occluded landmarks and poorly lit images). The goal of our method is to estimate this aleatoric uncertainty.

The fact that each image only has one ground-truth labeled location per landmark makes estimating this uncertainty distribution difficult, but not impossible. To do so, we use a parametric model for the uncertainty distribution. We train a neural network to estimate the parameters of the model for each landmark of each input face image so as to maximize the likelihood under the model of the ground-truth location of that landmark (summed across all landmarks of all training faces).

The main contributions of this work are as follows:

- This is the first work to introduce the concept of parametric uncertainty estimation for face alignment.
- We propose an end-to-end trainable model for the joint estimation of landmark location, uncertainty, and visibility likelihood (LUVLi), modeled as a mixed random variable.
- We compare our model using multivariate Gaussian and multivariate Laplacian probability distributions.
- Our algorithm yields accurate uncertainty estimation and state-of-the-art landmark localization results on several face alignment datasets.
- We are releasing a new dataset with manual labels of the locations of 68 landmarks on over $19{,}000$ face images in a wide variety of poses, where each landmark is also labeled with one of three visibility categories.

2. Related Work

2.1. Face Alignment

Early methods for face alignment were based on Active Shape Models (ASM) and Active Appearance Models (AAM) [16, 18, 66, 69, 78] as well as their variations [1, 19, 36, 49, 50, 62]. Subsequently, direct regression methods became popular due to their excellent performance. Of these, tree-based regression methods [9, 17, 40, 60, 76] proved particularly fast, and the subsequent cascaded regression methods [2, 22, 75, 77, 83] improved accuracy.

Recent approaches [7, 72, 73, 79, 81, 84, 87, 88] are all based on deep learning and can be classified into two subcategories: direct regression [10, 73] and heatmap-based approaches. The SOTA deep methods, e.g., stacked hourglass networks [7, 84] and densely connected U-nets (DU-Net) [72], use a cascade of deep networks, originally developed for human body 2D pose estimation [55]. These models [7, 55, 71, 72] 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. [48] uses a larger variance for early hourglasses and a smaller variance for later hourglasses. [79] employs different variations of MSE for different pixels of the proxy ground-truth heatmap. Recent works also infer facial boundary maps to improve alignment [79, 81]. In heatmap-based methods, landmarks are estimated by the argmax of each predicted heatmap. Indirect inference through a predicted heatmap offers several advantages over direct prediction [4].

Disadvantages of Heatmap-Based Approaches. These heatmap-based methods have at least two disadvantages. First, since the goal of training is to mimic a proxy ground-truth heatmap containing a fixed symmetric Gaussian, the predicted heatmaps are poorly suited to uncertainty prediction [13, 14]. Second, they suffer from quantization errors since the heatmap's argmax is only determined to the nearest pixel [51, 56, 70]. To achieve sub-pixel localization for body pose estimation, [51] replaces the argmax with a spatial mean over the softmax. Alternatively, for sub-pixel localization in videos, [70] samples two additional points adjacent to the max of the heatmap to estimate a local peak.

Landmark Regression with Uncertainty. We have only found two other methods that estimate uncertainty of landmark regression, both developed concurrently with our approach. The first method [13, 14] estimates face alignment uncertainty using a non-parametric approach: a kernel density network obtained by convolving the heatmaps with a fixed symmetric Gaussian kernel. The second [32] performs body pose estimation with uncertainty using direct regression method (no heatmaps) to directly predict the mean and precision matrix of a Gaussian distribution.

2.2. Uncertainty Estimation in Neural Networks

Uncertainty estimation broadly uses two types of approaches [46]: sampling-based and sampling-free. Sampling-based methods include Bayesian neural networks [67], Monte Carlo dropout [29], and bootstrap ensembles [45]. They rely on multiple evaluations of the input to estimate uncertainty [46], and bootstrap ensembles also need to store several sets of weights [37]. Thus, sampling-based methods work for small 1D regression problems but might not be feasible for higher-dimensional problems [37].

Sampling-free methods produce two outputs, one for the estimate and the other for the uncertainty, and optimize Gaussian log-likelihood (GLL) instead of classification and regression losses [41, 45, 46]. [45] combines the benefits of sampling-free and sampling-based methods.

Recent object detection methods have used uncertainty estimation [3, 34, 35, 38, 46, 47, 53]. Sampling-free methods [35, 46, 47] jointly estimate the four parameters of the bounding box using Gaussian log-likelihood [47], Laplacian log-likelihood [46], or both [35]. However, these methods assume the four parameters of the bounding box are independent (assume a diagonal covariance matrix). Sampling-based approaches use Monte Carlo dropout [53] and network ensembles [45] for object detection. Uncertainty estimation has also been applied to pixelwise depth regression [41], optical flow [37], pedestrian detection [5, 6, 54] and 3D vehicle detection [26].

3. Proposed Method

Figure 2 shows an overview of our LUVLi Face Alignment. The input RGB face image is passed through a DU-Net [72] architecture, to which we add three additional components branching from 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 and the third new component, the *Cholesky Estimator Network* (CEN) and the *Visibility Estimator Network* (VEN), emerge from the bottleneck layer of each U-net. CEN and VEN weights are shared across all U-nets. The CEN estimates the Cholesky coefficients of the covariance matrix for each landmark location. The VEN estimates the probability of visibility of each landmark in the image, 1 meaning visible and 0 meaning not visible. For each U-net i and each landmark j, the landmark's location estimate $\boldsymbol{\mu}_{ij}$, estimated covariance matrix $\boldsymbol{\Sigma}_{ij}$, and estimated visibility $\widehat{v}_{ij}$ are tied together by the LUVLi loss function $\mathcal{L}_{ij}$, which enables end-to-end optimization of the entire 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 [13, 14], our uncertainty prediction method is parametric: we directly estimate the parameters of a single multivariate Laplacian or Gaussian distribution. Furthermore, our method does not constrain the Laplacian or Gaussian covariance matrix to be diagonal.

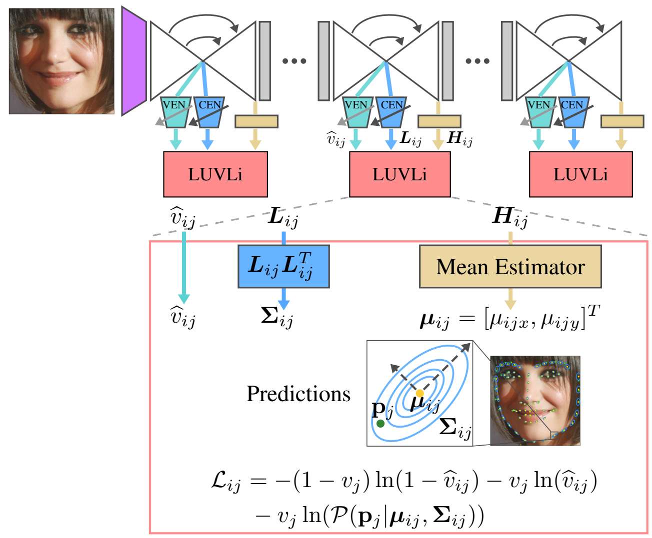


Figure 2: Overview of our LUVLi method. From each U-net of a DU-Net, we append a shared Cholesky Estimator Network (CEN) and Visibility Estimator Network (VEN) to the bottleneck layer and apply a mean estimator to the heatmap. The figure shows the joint estimation of location, uncertainty, and visibility of the landmarks performed for each U-net i and landmark j. The landmark has ground-truth (labeled) location $\mathbf{p}_j$ and visibility $v_j \in \{0, 1\}$.

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 U-net. 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 (See (16) in the supplementary material). We considered three different functions for σ: the ReLU function (eliminates the negative values), the softmax function (makes the mean estimator a soft-argmax of the heatmap [12, 25, 51, 85]), 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). The ablation studies (Section 5.5) show that choosing σ to be the ReLU function yields the simplest and best mean estimator.

3.2. LUVLi Loss

Occluded landmarks, e.g., landmarks on the far side of a profile-pose face, are common in real data. To explicitly represent visibility, we model the probability distributions of landmark locations using mixed random variables. For each landmark j in an image, we denote the ground-truth (labeled) visibility by the binary variable $v_j \in \{0, 1\}$, where 1 denotes *visible*, and the ground-truth location by $\mathbf{p}_j$. By convention, if the landmark is not visible ($v_j = 0$), then $\mathbf{p}_j = \emptyset$, a special symbol indicating non-existence. Together, these variables are distributed according to an unknown distribution $p(v_j, \mathbf{p}_j)$. The marginal Bernoulli distribution $p(v_j)$ captures the probability of visibility, $p(\mathbf{p}_j|v_j{=}1)$ denotes the distribution of the landmark location when it is visible, and $p(\mathbf{p}_j|v_j = 0) = \mathbf{1}_\emptyset(\mathbf{p}_j)$, where $\mathbf{1}_\emptyset$ denotes the PMF that assigns probability one to the symbol $\emptyset$.

After each U-net i, we estimate the joint distribution of the visibility v and location $\mathbf{z}$ of each landmark j via

$$q(v, \mathbf{z}) = q_v(v) q_z(\mathbf{z}|v), \tag{1}$$

where $q_v(v)$ is a Bernoulli distribution with

$$q_v(v = 1) = \widehat{v}_{ij}, \qquad q_v(v = 0) = 1 - \widehat{v}_{ij}, \tag{2}$$

where $\widehat{v}_{ij}$ is the predicted probability of visibility, and

$$q_z(\mathbf{z}|v = 1) = \mathcal{P}(\mathbf{z}|\boldsymbol{\mu}_{ij}, \boldsymbol{\Sigma}_{ij}), \tag{3}$$

$$q_z(\mathbf{z}|v = 0) = \emptyset, \tag{4}$$

where $\mathcal{P}$ denotes the likelihood of the landmark being at location $\mathbf{z}$ given the estimated mean $\boldsymbol{\mu}_{ij}$ and covariance $\boldsymbol{\Sigma}_{ij}$.

The LUVLi loss is the negative log-likelihood with respect to $q(v, \mathbf{z})$, as given by

$$\begin{aligned}\mathcal{L}_{ij} &= -\ln q(v_j, \mathbf{p}_j) \\ &= -\ln q_v(v_j) - \ln q_z(\mathbf{p}_j|v_j) \\ &= -(1 - v_j)\ln(1 - \widehat{v}_{ij}) - v_j \ln(\widehat{v}_{ij}) \\ &\quad - v_j \ln\big(\mathcal{P}(\mathbf{p}_j|\boldsymbol{\mu}_{ij}, \boldsymbol{\Sigma}_{ij})\big),\end{aligned} \tag{5}$$

and thus minimizing the loss is equivalent to maximum likelihood estimation.

The terms of (5) are a binary cross entropy plus v_j times the negative log-likelihood of $\mathbf{p}_j$ with respect to $\mathcal{P}$. This can be seen as an instance of multi-task learning [11], since we are predicting three things about each landmark: its location, uncertainty, and visibility. The first two terms on the right hand side of (5) can be seen as a classification loss for visibility, while the last term corresponds to a regression loss of location estimation. The sum of classification and regression losses is also widely used in object detection [39].

Minimization of negative log-likelihood also corresponds to minimizing KL-divergence, since

$$\mathbb{E}[-\ln q(v_j, \mathbf{p}_j)] = \mathbb{E}\left[\ln\frac{p(v_j, \mathbf{p}_j)}{q(v_j, \mathbf{p}_j)} - \ln p(v_j, \mathbf{p}_j)\right] \tag{6}$$

$$= \mathrm{D_{KL}}(p(v_j, \mathbf{p}_j) \| q(v_j, \mathbf{p}_j)) + \mathbb{E}[-\ln p(v_j, \mathbf{p}_j)], \tag{7}$$

where expectations are with respect to $(v_j, \mathbf{p}_j) \sim p(v_j, \mathbf{p}_j)$, and the entropy term $\mathbb{E}[-\ln p(v_j, \mathbf{p}_j)]$ is constant with respect to the estimate $q(v_j, \mathbf{p}_j)$. Further, since

$$\begin{aligned}\mathbb{E}[-\ln q(v_j, \mathbf{p}_j)] &= \mathbb{E}_{v_j \sim p(v_j)}[-\ln q(v_j)] \\ &\quad + p_v \mathbb{E}_{\mathbf{p}_j \sim p(\mathbf{p}_j|v_j=1)}[-\ln \mathcal{P}(\mathbf{p}_j|\boldsymbol{\mu}_{ij}, \boldsymbol{\Sigma}_{ij})],\end{aligned} \tag{8}$$

where $p_v := p(v_j = 1)$ for brevity, minimizing the negative log-likelihood (LUVLi loss) is also equivalent to minimizing the combination of KL-divergences given by

$$\mathrm{D_{KL}}\big(p(v_j)\|q(v)\big) + p_v \mathrm{D_{KL}}\big(p(\mathbf{p}_j|v_j{=}1)\|\mathcal{P}(\mathbf{z}|\boldsymbol{\mu}_{ij}, \boldsymbol{\Sigma}_{ij})\big) \tag{9}$$

3.2.1 Models for Location Likelihood

For the multivariate location distribution $\mathcal{P}$, we consider two different models: Gaussian and Laplacian.

Gaussian Likelihood. The 2D Gaussian likelihood is:

$$\mathcal{P}(\mathbf{z}|\boldsymbol{\mu}_{ij}, \boldsymbol{\Sigma}_{ij}) = \frac{\exp\big(-\frac{1}{2}(\mathbf{z}-\boldsymbol{\mu}_{ij})^T \boldsymbol{\Sigma}_{ij}^{-1}(\mathbf{z}-\boldsymbol{\mu}_{ij})\big)}{2\pi\sqrt{|\boldsymbol{\Sigma}_{ij}|}}. \tag{10}$$

Substituting (10) into (5), we have

$$\begin{aligned}\mathcal{L}_{ij} = &-(1-v_j)\ln(1-\widehat{v}_{ij}) - v_j \ln(\widehat{v}_{ij}) + v_j \underbrace{\frac{1}{2}\log|\boldsymbol{\Sigma}_{ij}|}_{T_1} \\ &+ v_j \underbrace{\frac{1}{2}(\mathbf{p}_j - \boldsymbol{\mu}_{ij})^T \boldsymbol{\Sigma}_{ij}^{-1}(\mathbf{p}_j - \boldsymbol{\mu}_{ij})}_{T_2}.\end{aligned} \tag{11}$$

In (11), 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.

Laplacian Likelihood. We use a 2D Laplacian likelihood [43] given by:

$$P(\mathbf{z}|\boldsymbol{\mu}_{ij}, \boldsymbol{\Sigma}_{ij}) = \frac{e^{-\sqrt{3(\mathbf{z}-\boldsymbol{\mu}_{ij})^T \boldsymbol{\Sigma}_{ij}^{-1}(\mathbf{z}-\boldsymbol{\mu}_{ij})}}}{\frac{2\pi}{3}\sqrt{|\boldsymbol{\Sigma}_{ij}|}}. \tag{12}$$

Substituting (12) in (5), we have

$$\begin{aligned}\mathcal{L}_{ij} = &-(1-v_j)\ln(1-\widehat{v}_{ij}) - v_j \ln(\widehat{v}_{ij}) + v_j \underbrace{\frac{1}{2}\log|\boldsymbol{\Sigma}_{ij}|}_{T_1} \\ &+ v_j \underbrace{\sqrt{3(\mathbf{p}_j - \boldsymbol{\mu}_{ij})^T \boldsymbol{\Sigma}_{ij}^{-1}(\mathbf{p}_j - \boldsymbol{\mu}_{ij})}}_{T_2}.\end{aligned} \tag{13}$$

In (13), T_2 is a scaled Mahalanobis distance, while T_1 serves as a regularization or prior term that ensures that the Laplacian uncertainty distribution does not get too large.

Note that if $\boldsymbol{\Sigma}_{ij}$ is the identity matrix and if all landmarks are assumed to be visible, then (11) simply reduces to the squared ℓ_2 distance, and (13) just minimizes the ℓ_2 distance.

3.3. Uncertainty and Visibility Estimation

Our proposed method uses heatmaps for estimating landmarks' locations, but not for estimating their uncertainty and visibility. We experimented with several methods for computing a covariance matrix directly from a heatmap, but none were accurate enough. We discuss this in Section 5.1.

Cholesky Estimator Network (CEN). We represent the uncertainty of each landmark location using a 2×2 covariance matrix $\boldsymbol{\Sigma}_{ij}$, which is symmetric positive definite. The three degrees of freedom of $\boldsymbol{\Sigma}_{ij}$ are captured by its Cholesky decomposition: a lower-triangular matrix $\boldsymbol{L}_{ij}$ such that $\boldsymbol{\Sigma}_{ij} = \boldsymbol{L}_{ij}\boldsymbol{L}_{ij}^T$. To estimate the elements of $\boldsymbol{L}_{ij}$, we append a Cholesky Estimator Network (CEN) to the bottleneck of each U-net. The CEN is a fully connected linear layer whose input is the bottleneck of the U-net ($128 \times 4 \times 4 = 2{,}048$ dimensions) and output is an $N_p \times 3$-dimensional vector, where N_p is the number of landmarks (e.g., 68). As the Cholesky decomposition $\boldsymbol{L}_{ij}$ of a covariance matrix must have positive diagonal elements, we pass the corresponding entries of the output through an ELU activation function [15], to which we add a constant to ensure the output is always positive (asymptote is negative x-axis).

Visibility Estimator Network (VEN). To estimate the visibility of the landmark v_e, we add another fully connected linear layer whose input is the bottleneck of the U-net ($128 \times 4 \times 4 = 2{,}048$ dimensions) and output is an N_p-dimensional vector. This is passed through a sigmoid activation so the predicted visibility $\widehat{v}_{ij}$ is between 0 and 1.

The addition of these two fully connected layers only slightly increases the size of the original model. The loss for a single U-net is the averaged $\mathcal{L}_{ij}$ 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 all K of the U-nets:

$$\mathcal{L} = \sum_{i=1}^{K} \lambda_i \mathcal{L}_i \,, \quad \text{where} \quad \mathcal{L}_i = \frac{1}{N_p} \sum_{j=1}^{N_p} \mathcal{L}_{ij} \,. \tag{14}$$

At test time, each landmark's mean and Cholesky coefficients are derived from the Kth (final) U-net. The covariance matrix is calculated from the Cholesky coefficients.

4. New Dataset: MERL-RAV

To promote future research in face alignment with uncertainty, we now introduce a new dataset with entirely new, manual labels of over 19,000 face images from the AFLW [42] dataset. In addition to landmark locations, every landmark is labeled with one of three visibility classes. We call the new dataset *MERL Reannotation of AFLW with Visibility* (MERL-RAV).

Visibility Classification. Each landmark of every face is classified as either *unoccluded*, *self-occluded*, or *externally occluded*, as illustrated in Figure 3. *Unoccluded* denotes landmarks that can be seen directly in the image, with no obstructions. *Self-occluded* denotes landmarks that are occluded because of extreme head pose—they are occluded by another part of the face (e.g., landmarks on the far side of a profile-view face). *Externally occluded* denotes landmarks that are occluded by hair or an intervening object such as a cap, hand, microphone, or goggles. Human labelers are generally very bad at localizing self-occluded landmarks, so we do not provide ground-truth locations for these. We do provide ground-truth (labeled) locations for both unoccluded and externally occluded landmarks.

Relationship to Visibility in LUVLi. In Section 3, visible landmarks ($v_j = 1$) are landmarks for which ground-truth location information is available, while invisible landmarks ($v_j = 0$) are landmarks for which no ground-truth location information is available ($\mathbf{p}_j = \emptyset$). Thus, invisible ($v_j = 0$) in the model is equivalent to the *self-occluded* landmarks in our dataset. In contrast, both *unoccluded* and *externally occluded* landmarks are considered visible ($v_j = 1$) in our model. We choose this because human labelers are generally good at estimating the locations of externally occluded landmarks but poor at estimating the locations of self-occluded landmarks.

Table 1: Overview of face alignment datasets. [Key: Self Occ= Self-Occlusions, Ext Occ= External Occlusions]

Dataset	**#train**	**#test**	**#marks**	**Profile Images**	**Self Occ**	**Ext Occ**
COFW [8]	1,345	507	29	×	×	✓
COFW-68 [30]	-	507	68	×	×	✓
300-W [63–65]	3,837	600	68	×	×	×
Menpo 2D [21, 74, 86]	7,564	7,281	68/39	✓	F/P	×
300W-LP-2D [90]	61,225	-	68	✓	T	×
WFLW [81]	7,500	2,500	**98**	✓	×	×
AFLW [42]	20,000	**4,386**	21	✓	✓	×
AFLW-19 [89]	20,000	**4,386**	19	✓	×	×
AFLW-68 [59]	20,000	**4,386**	68	✓	×	×
MERL-RAV (Ours)	15,449	3,865	68	✓	✓	✓

Existing Datasets. The most commonly used publicly available datasets for evaluation of 2D face alignment are summarized in Table 1. The 300-W dataset [63–65] uses a 68-landmark system that was originally used for Multi-PIE [31]. Menpo 2D [21, 74, 86] makes a hard distinction (denoted F/P) between nearly frontal faces (F) and profile faces (P). Menpo 2D uses the same landmarks as 300-W for frontal faces, but for profile faces it uses a different set of 39 landmarks that do not all correspond to the 68 landmarks in the frontal images. 300W-LP-2D [7, 90] is a synthetic dataset created by automatically reposing 300-W faces, so it has a large number of labels, but they are noisy. The 3D model locations of self-occluded landmarks are projected onto the visible part of the face as if the face were transparent (denoted by T). The WFLW [81] and AFLW-68 [59] datasets do not identify which landmarks are self-occluded, but instead label self-occluded landmarks as if they were located on the visible boundary of the noseless face.

Differences from Existing Datasets. Our MERL-RAV dataset is the only one that labels every landmark using both types of occlusion (self-occlusion and external occlusion). Only one other dataset, AFLW, indicates which individual landmarks are self-occluded, but it has far fewer landmarks and does not label external occlusions. COFW and COFW-68 indicate which landmarks are externally occluded but do not have self-occlusions. Menpo 2D categorizes faces as frontal or profile, but landmarks of the two classes are incompatible. Unlike Menpo 2D, our dataset smoothly transitions from frontal to profile, with gradually more and more landmarks labeled as self-occluded.

Our dataset uses the widely adopted 68 landmarks used by 300-W, to allow for evaluation and cross-dataset comparison. Since it uses images from AFLW, our dataset has pose variation up to $\pm 120°$ yaw and $\pm 90°$ pitch. Focusing on yaw, we group the images into five pose classes: frontal,

Pose	Side	#Train	#Test
Frontal	-	8,778	2,195
Half-Profile	Left half	1,180	295
	Right half	1,221	306
Profile	Left	2,080	521
	Right	2,190	548
Total	-	15,449	3,865

Table 2: Statistics of our new dataset for face alignment.

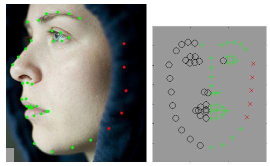


Figure 3: **Unoccluded**, **externally occluded**, and self-occluded landmarks.

left and right half-profile, and left and right profile. The train/test split is in the ratio of $4:1$. Table 2 provides the statistics of our MERL-RAV dataset. A sample image from the dataset is shown in Figure 3. In the figure, unoccluded landmarks are green, externally occluded landmarks are red, and self-occluded landmarks are indicated by black circles in the face schematic on the right.

5. Experiments

Our experiments use the datasets 300-W [63–65], 300W-LP-2D [90], Menpo 2D [21, 74, 86], COFW-68 [8, 30], AFLW-19 [42], WFLW [81], and our MERL-RAV dataset. Training and testing protocols are described in the supplementary material. On a 12 GB GeForce GTX Titan-X GPU, the inference time per image is 17 ms.

Evaluation Metrics. We use the standard metrics NME, AUC, and FR [14, 72, 79]. In each table, we report results using the same metric adopted in respective baselines.

Normalized Mean Error (NME). The NME is defined as:

$$\text{NME }(\%) = \frac{1}{N_p}\sum_{j=1}^{N_p} v_j \frac{\|\mathbf{p}_j - \boldsymbol{\mu}_{Kj}\|_2}{d} \times 100, \qquad (15)$$

where v_j, $\mathbf{p}_j$ and $\boldsymbol{\mu}_{Kj}$ respectively denote the visibility, ground-truth and predicted location of landmark j from the Kth (final) U-net. The factor of v_j is there because we cannot compute an error value for points without ground-truth location labels. Several variations of the normalizing term d are used. NME_{box} [7, 14, 86] sets d to the geometric mean of the width and height of the ground-truth bounding box $\left(\sqrt{w_{\text{bbox}} \cdot h_{\text{bbox}}}\right)$, while $\text{NME}_{\text{inter-ocular}}$ [44, 64, 72] sets d to the distance between the outer corners of the two eyes. If a ground-truth box is not provided, the tight bounding box of the landmarks is used [7, 14]. NME_{diag} [68, 81] sets d as the diagonal of the bounding box.

Area Under the Curve (AUC). To compute the AUC, the cumulative distribution of the fraction of test images whose NME $(\%)$ is less than or equal to the value on the horizontal axis is first plotted. The AUC for a test set is then computed as the area under that curve, up to the cutoff NME value.

Failure Rate (FR). FR refers to the percentage of images in the test set whose NME is larger than a certain threshold.

5.1. 300-W Face Alignment

We train on the 300-W [63–65], and test on 300-W, Menpo 2D [21, 74, 86], and COFW-68 [8, 30]. Some of the models are pre-trained on the 300W-LP-2D [90].

Table 3: $\text{NME}_{\text{inter-ocular}}$ on 300-W Common, Challenge, and Full datasets (Split 1). [Key: **Best**, **Second best**]

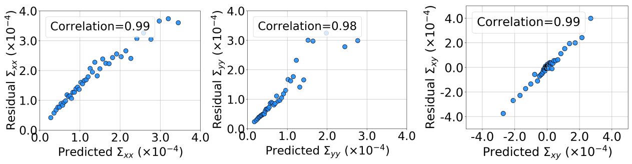
	$\text{NME}_{\text{inter-ocular}}$ (%)(↓)		
	Common	Challenge	Full
SAN [23]	3.34	6.60	3.98
AVS [59]	3.21	6.49	3.86
DAN [44]	3.19	5.24	3.59
LAB (w/B) [81]	2.98	5.19	3.49
Teacher [24]	2.91	5.91	3.49
DU-Net (Public code) [72]	2.97	5.53	3.47
DeCaFa (More data) [20]	2.93	5.26	3.39
HR-Net [68]	2.87	5.15	3.32
HG-HSLE [91]	2.85	**5.03**	3.28
AWing [79]	**2.72**	**4.52**	**3.07**
LUVLi (Ours)	**2.76**	5.16	**3.23**

Table 4: NME_{box} and $\text{AUC}^{7}_{\text{box}}$ comparisons on 300-W Test (Split 2), Menpo 2D and COFW-68 datasets.
[Key: **Best**, **Second best**, * = Pretrained on 300W-LP-2D]

	NME_{box} (%) (↓)			$\text{AUC}^{7}_{\text{box}}$ (%) (↑)		
	300-W	Menpo	COFW	300-W	Menpo	COFW
SAN* [23] in [14]	2.86	2.95	3.50	59.7	61.9	51.9
2D-FAN* [7]	2.32	2.16	2.95	66.5	69.0	57.5
KDN [13]	2.49	2.26	-	67.3	68.2	-
Softlabel* [14]	2.32	2.27	2.92	66.6	67.4	57.9
KDN* [14]	**2.21**	**2.01**	**2.73**	**68.3**	**71.1**	60.1
LUVLi (Ours)	2.24	2.18	2.75	**68.3**	70.1	**60.8**
LUVLi* (Ours)	**2.10**	**2.04**	**2.57**	**70.2**	**71.9**	**63.4**

Data Splits and Evaluation Metrics. There are two commonly used train/test splits for 300-W; we evaluate our method on both. *Split* 1*:* The train set contains 3,148 images and full test set has 689 images [72]. *Split* 2*:* The train set includes 3,837 images and test set has 600 images [14]. The model trained on Split 2 is additionally evaluated on the 6,679 near-frontal training images of Menpo 2D and 507 test images of COFW-68 [14]. For Split 1, we use $\text{NME}_{\text{inter-ocular}}$ [68, 72, 79]. For Split 2, we use NME_{box} and AUC_{box} with 7% cutoff [7, 14].

Results: Localization and Cross-Dataset Evaluation. The face alignment results for 300-W Split 1 and Split 2 are summarized in Table 3 and 4, respectively. Table 4 also shows the results of our model (trained on Split 2) on the Menpo and COFW-68 datasets, as in [7, 14]. The results in Table 3 show that our LUVLi landmark localization is competitive with the SOTA methods on Split 1, usually one of the best two. Table 4 shows that LUVLi significantly outperforms the SOTA on Split 2, performing best on 5 out of the 6 cases (3 datasets × 2 metrics). This is particularly impressive on 300-W Split 2, because even though most of the other methods are pre-trained on the 300W-LP-2D dataset (as was our best method, LUVLi*), our method without pre-training still outperforms the SOTA in 2 of 6 cases. Our method performs particularly well in the cross-dataset evaluation on the more challenging COFW-68 dataset, which has multiple externally occluded landmarks.



(a) variance of x (b) variance of y (c) covariance of x, y

Figure 4: Mean squared residual error vs. predicted covariance matrix for all landmarks in 300-W Test (Split 2).

Accuracy of Predicted Uncertainty. To evaluate the accuracy of the predicted uncertainty covariance matrix, $\mathbf{\Sigma}_{Kj} = \begin{bmatrix} \Sigma_{Kj_{xx}} & \Sigma_{Kj_{xy}} \\ \Sigma_{Kj_{xy}} & \Sigma_{Kj_{yy}} \end{bmatrix}$, we compare all three unique terms of this prediction with the statistics of the *residuals* (2D error between the ground-truth location $\mathbf{p}_j$ and the predicted location $\boldsymbol{\mu}_{Kj}$) of all landmarks in the test set. We explain how we do this for $\Sigma_{Kj_{xx}}$ in Figure 4a. First, we bin every landmark of every test image according to the value of the predicted variance in the x-direction $\left(\Sigma_{Kj_{xx}}\right)$. Each bin is represented by one point in the scatter plot. Averaging $\Sigma_{Kj_{xx}}$ across the $N_{\text{bin}} = 734$ landmark points within each bin gives a single predicted $\Sigma_{Kj_{xx}}$ value (horizontal axis). We next compute the residuals in the x-direction of all landmarks in the bin, and calculate the average of the squared residuals to obtain $\Sigma_{xx} = \mathbb{E}(p_{jx} - \mu_{Kj_x})^2$ for the bin. This mean squared residual error, Σ_{xx}, is plotted on the vertical axis. If our predicted uncertainties are accurate, this residual error, Σ_{xx}, should be roughly equal to the predicted uncertainty variance in the x-direction (horizontal axis).

Figure 4 shows that all three terms of our method's predicted covariance matrices are highly predictive of the actual uncertainty: the mean squared residuals (error) are strongly proportional to the predicted covariance values, as evidenced by Pearson correlation coefficients of 0.98 and 0.99. However, decreasing N_{bin} from 734 (plotted in Figure 4) to just 36 makes the correlation coefficients decrease to $0.84, 0.80, 0.72$. Thus, the predicted uncertainties are excellent after averaging but may yet have room to improve.

Uncertainty is Larger for Occluded Landmarks. The COFW-68 [30] test set annotates which landmarks are externally occluded. Similar to [14], we use this to test uncertainty predictions of our model, where the square root of the determinant of the uncertainty covariance is a scalar measure of predicted uncertainty. We report the error, NME_{box}, and average predicted uncertainty, $|\mathbf{\Sigma}_{Kj}|^{1/2}$, in Table 5. We do not use any occlusion annotation from the dataset during training. Like [14], we find that our model's predicted uncertainty is much larger for externally occluded landmarks than for unoccluded landmarks. Furthermore, our method's location estimates are more accurate (smaller NME_{box}) than those of [14] for both occluded and unoccluded landmarks.

Heatmaps vs. Direct Regression for Uncertainty. We tried multiple approaches to estimate the uncertainty distribution from heatmaps, but none of these worked nearly as well as our direct regression using the CEN. We believe this is because in current heatmap-based networks, the resolution of the heatmap (64×64) is too low for accurate uncertainty estimation. This is demonstrated in Figure 5, which shows a histogram over all landmarks in 300-W Test (Split 2) of LUVLi's predicted covariance in the narrowest direction of the covariance ellipse (the smallest eigenvalue of the predicted covariance matrix). The figure shows that in most cases, the uncertainty ellipses are less wide than one heatmap pixel, which explains why heatmap-based methods are not able to accurately capture such small uncertainties.

Table 5: NME_{box} and uncertainty $\left(|\mathbf{\Sigma}_{Kj}|^{1/2}\right)$ on unoccluded and externally occluded landmarks of COFW-68 dataset. [Key: **Best**]

	Unoccluded		Externally Occluded	
	NME_{box}	$\lvert\mathbf{\Sigma}\rvert^{1/2}$	NME_{box}	$\lvert\mathbf{\Sigma}\rvert^{1/2}$
Softlabel [14]	2.30	5.99	5.01	7.32
KDN [14]	2.34	1.63	4.03	11.62
LUVLi (Ours)	**2.15**	9.31	**4.00**	32.49

Table 6: NME and AUC on the AFLW-19 dataset (previous results are quoted from [14, 68]). [Key: **Best**, **Second best**]

	NME_{diag}		NME_{box}	$\text{AUC}^7_{\text{box}}$
	Full	Frontal	Full	Full
CFSS [88]	3.92	2.68	-	-
CCL [89]	2.72	2.17	-	-
DAC-CSR [28]	2.27	1.81	-	-
LLL [61]	1.97	-	-	-
SAN [23]	1.91	1.85	4.04	54.0
DSRN [52]	1.86	-	-	-
LAB (w/o B) [81]	1.85	1.62	-	-
HR-Net [68]	**1.57**	**1.46**	-	-
Wing [27]	-	-	3.56	53.5
KDN [14]	-	-	**2.80**	**60.3**
LUVLi (Ours)	**1.39**	**1.19**	**2.28**	**68.0**

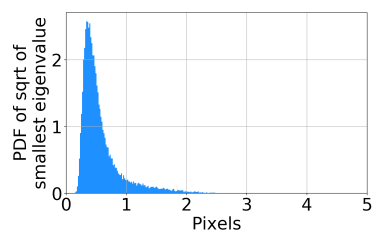


Figure 5: Histogram of the smallest eigenvalue of $\mathbf{\Sigma}_{Kj}$.

5.2. AFLW-19 Face Alignment

On AFLW-19, we train on 20,000 images, and test on two sets: the *AFLW-Full* set (4,386 test images) and the *AFLW-Frontal* set (1,314 test images), as in [68, 81, 89]. Table 6 compares our method's localization performance with other methods that only train on AFLW-19 (without training on any 68-landmark dataset). Our proposed method outperforms not only the other uncertainty-based method KDN [14], but also all previous SOTA methods, by a significant margin on both AFLW-Full and AFLW-Frontal.

Table 7: WFLW-All dataset results for $\text{NME}_{\text{inter-ocular}}$, $\text{AUC}^{10}_{\text{inter-ocular}}$, and $\text{FR}^{10}_{\text{inter-ocular}}$. [Key: **Best**, **Second best**]

	NME(%) (↓)	AUC^{10} (↑)	FR^{10}(%) (↓)
CFSS [88]	9.07	0.366	20.56
DVLN [82]	10.84	0.456	10.84
LAB (w/B) [81]	5.27	0.532	7.56
Wing [27]	5.11	0.554	6.00
DeCaFa (w/DA) [20]	4.62	0.563	4.84
AVS [59]	4.39	**0.591**	4.08
AWing [79]	**4.36**	0.572	**2.84**
LUVLi (Ours)	**4.37**	**0.577**	**3.12**

Table 8: NME_{box} and $\text{AUC}^7_{\text{box}}$ comparisons on MERL-RAV dataset. [Key: **Best**]

Metric (%)	Method	All	Frontal	Half-Profile	Profile
NME_{box}(↓)	DU-Net [72]	1.99	1.89	2.50	1.92
	LUVLi (Ours)	**1.61**	**1.74**	**1.79**	**1.25**
$\text{AUC}^7_{\text{box}}$(↑)	DU-Net [72]	71.80	73.25	64.78	72.79
	LUVLi (Ours)	**77.08**	**75.33**	**74.69**	**82.10**

Table 9: MERL-RAV results on three types of landmarks.

	Self-Occluded	Unoccluded	Externally Occluded
Mean $\widehat{v}_j$	0.13	0.98	0.98
Accuracy (Visible)	0.88	0.99	0.99
NME_{box}	-	1.60	3.53
$\lvert\boldsymbol{\Sigma}\rvert^{0.5}$	-	9.28	34.41
$\lvert\boldsymbol{\Sigma}\rvert^{0.5}_{box} (\times 10^{-4})$	-	1.87	7.00

5.3. WFLW Face Alignment

Landmark localization results for WFLW are shown in Table 7. More detailed results on WFLW are in the supplementary material. Compared to the SOTA methods, LUVLi yields the second best performance on all metrics. Furthermore, while the other methods only predict landmark locations, LUVLi also estimates the prediction uncertainties.

5.4. MERL-RAV Face Alignment

Results of Landmark Localization. Results for all head poses on our MERL-RAV dataset are shown in Table 8.

Results for All Visibility Classes. We analyze LUVLi's performance on all test images for all three types of landmarks in Table 9. The first row is the mean value of the predicted visibility, $\widehat{v}_j$, for each type of landmark. Accuracy (Visible) tests the accuracy of predicting that landmarks are visible when $\widehat{v}_j > 0.5$. The last two rows show the scalar measure of uncertainty, $|\boldsymbol{\Sigma}_{Kj}|^{1/2}$, both unnormalized and normalized by the face box size $\left(|\boldsymbol{\Sigma}|^{0.5}_{box}\right)$ similar to NME_{box}. Similar to results on COFW-68 in Table 5, the model predicts higher uncertainty for locations of externally occluded landmarks than for unoccluded landmarks.

5.5. Ablation Studies

Table 10 compares modifications of our approach on Split 2. Table 10 shows that computing the loss only on the last U-net performs worse than computing loss on all U-nets, perhaps because of the vanishing gradient problem [80]. Moreover, LUVLi's log-likelihood loss without visibility outperforms using MSE loss on the landmark locations (which is equivalent to setting all $\boldsymbol{\Sigma}_{ij} = \mathbf{I}$). We also find that the loss with Laplacian likelihood (13) outperforms the one with Gaussian likelihood (11). Training from scratch is slightly inferior to first training the base DU-Net architecture before fine-tuning the full LUVLi network, consistent with previous observations that the model does not have strongly supervised pixel-wise gradients through the heatmap during training [56]. Regarding the method for estimating the mean, using heatmaps is more effective than direct regression (Direct) from each U-net bottleneck, consistent with previous observations that neural networks have difficulty predicting continuous real values [4, 56]. As described in Section 3.1, in addition to ReLU, we compared two other functions for σ: softmax, and a temperature-scaled softmax (τ-softmax). Results for temperature-scaled softmax and ReLU are essentially tied, but the former is more complicated and requires tuning a temperature parameter, so we chose ReLU for our LUVLi model. Finally, reducing the number of U-nets from 8 to 4 increases test speed by about 2× with minimal decrease in performance.

Table 10: Ablation studies using our method trained on 300W-LP-2D and then fine-tuned on 300-W (Split 2).

Change from LUVLi model:		NME_{box} (%)		$\text{AUC}^7_{\text{box}}$ (%)	
Changed	From → To	300-W	Menpo	300-W	Menpo
Supervision	All HGs → Last HG	2.32	2.16	67.7	70.8
Loss	LUVLi→MSE	2.25	2.10	68.0	71.0
	Lap+vis→Gauss+No-vis	2.15	2.07	69.6	71.6
	Lap+vis → Gauss+vis	2.13	2.05	69.8	71.8
	Lap+vis → Lap+No-vis	2.10	2.05	70.1	71.8
Initialization	LP-2D wts→300-W wts	2.24	2.18	68.3	70.1
	LP-2D wts→ Scratch	2.32	2.26	67.2	69.4
Mean Estimator	Heatmap → Direct	4.32	3.99	41.3	47.5
	ReLU → softmax	2.37	2.19	66.4	69.8
	ReLU → τ-softmax	2.10	2.04	70.1	71.8
No of HG	8 → 4	2.14	2.07	69.5	71.5
—	**LUVLi (our best model)**	**2.10**	**2.04**	**70.2**	**71.9**

6. Conclusions

In this paper, we present LUVLi, a novel end-to-end trainable framework for jointly estimating facial landmark locations, uncertainty, and visibility. This joint estimation not only provides accurate uncertainty predictions, but also yields state-of-the-art estimates of the landmark locations on several datasets. We show that the predicted uncertainty distinguishes between unoccluded and externally occluded landmarks without any supervision for that task. In addition, the model achieves sub-pixel accuracy by taking the spatial mean of the ReLU'ed heatmap, rather than the arg max. We also introduce a new dataset containing manual labels of over 19,000 face images with 68 landmarks, which also labels every landmark with one of three visibility classes. Although our implementation is based on the DU-Net architecture, our framework is general enough to be applied to a variety of architectures for simultaneous estimation of landmark location, uncertainty, and visibility.

References

[1] Akshay Asthana, Tim Marks, Michael Jones, K.H. Tieu, and Rohith M.V. Fully automatic pose-invariant face recognition via 3D pose normalization. In *ICCV*, 2011. 2

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

[3] Yousef Atoum, Joseph Roth, Michael Bliss, Wende Zhang, and Xiaoming Liu. Monocular video-based trailer coupler detection using multiplexer convolutional neural network. In *ICCV*, 2017. 3

[4] Vasileios Belagiannis and Andrew Zisserman. Recurrent human pose estimation. In *FG*, 2017. 2, 8

[5] Lorenzo Bertoni, Sven Kreiss, and Alexandre Alahi. Monoloco: Monocular 3D pedestrian localization and uncertainty estimation. *ICCV*, 2019. 3

[6] Apratim Bhattacharyya, Mario Fritz, and Bernt Schiele. Long-term on-board prediction of people in traffic scenes under uncertainty. In *CVPR*, 2018. 3

[7] Adrian Bulat and Georgios 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. 2, 5, 6, 12

[8] Xavier Burgos-Artizzu, Pietro Perona, and Piotr Dollár. Robust face landmark estimation under occlusion. In *ICCV*, 2013. 5, 6

[9] Xudong Cao, Yichen Wei, Fang Wen, and Jian Sun. Face alignment by explicit shape regression. *IJCV*, 2014. 2

[10] Joao Carreira, Pulkit Agrawal, Katerina Fragkiadaki, and Jitendra Malik. Human pose estimation with iterative error feedback. In *CVPR*, 2016. 2

[11] Rich Caruana. Multitask learning. *Machine learning*, 1997. 4

[12] Olivier Chapelle and Mingrui Wu. Gradient descent optimization of smoothed information retrieval metrics. *Information retrieval*, 2010. 3

[13] Lisha Chen and Qian Ji. Kernel density network for quantifying regression uncertainty in face alignment. In *NeurIPS Workshops*, 2018. 2, 3, 6, 12

[14] Lisha Chen, Hui Su, and Qiang Ji. Face alignment with kernel density deep neural network. In *ICCV*, 2019. 1, 2, 3, 6, 7

[15] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (ELUs). In *ICLR*, 2016. 5

[16] Timothy Cootes, Gareth Edwards, and Christopher Taylor. Active appearance models. *TPAMI*, 2001. 2

[17] Timothy Cootes, Mircea Ionita, Claudia Lindner, and Patrick Sauer. Robust and accurate shape model fitting using random forest regression voting. In *ECCV*, 2012. 2

[18] Timothy Cootes, Christopher Taylor, David Cooper, and Jim Graham. Active shape models-their training and application. *Computer Vision and Image Understanding*, 1995. 2

[19] Timothy Cootes, Gavin Wheeler, Kevin Walker, and Christopher Taylor. View-based active appearance models. *Image and Vision Computing*, 2002. 2

[20] Arnaud Dapogny, Kevin Bailly, and Matthieu Cord. DeCaFA: Deep convolutional cascade for face alignment in the wild. In *ICCV*, 2019. 6, 8, 15, 16

[21] Jiankang Deng, Anastasios Roussos, Grigorios Chrysos, Evangelos Ververas, Irene Kotsia, Jie Shen, and Stefanos Zafeiriou. The Menpo benchmark for multi-pose 2D and 3D facial landmark localisation and tracking. *IJCV*, 2019. 5, 6

[22] Piotr Dollár, Peter Welinder, and Pietro Perona. Cascaded pose regression. In *CVPR*, 2010. 2

[23] Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi Yang. Style aggregated network for facial landmark detection. In *CVPR*, 2018. 6, 7

[24] Xuanyi Dong and Yi Yang. Teacher supervises students how to learn from partially labeled images for facial landmark detection. In *ICCV*, 2019. 6

[25] Xuanyi Dong, Shoou-I Yu, Xinshuo Weng, Shih-En Wei, Yi Yang, and Yaser Sheikh. Supervision-by-registration: An unsupervised approach to improve the precision of facial landmark detectors. In *CVPR*, 2018. 3

[26] Di Feng, Lars Rosenbaum, and Klaus Dietmayer. Towards safe autonomous driving: Capture uncertainty in the deep neural network for lidar 3D vehicle detection. In *ITSC*, 2018. 3

[27] Zhen-Hua Feng, Josef Kittler, Muhammad Awais, Patrik Huber, and Xiao-Jun Wu. Wing loss for robust facial landmark localisation with convolutional neural networks. In *CVPR*, 2018. 7, 8, 16

[28] 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. 7

[29] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *ICML*, 2016. 3

[30] Golnaz Ghiasi and Charless Fowlkes. Occlusion coherence: Detecting and localizing occluded faces. *arXiv preprint arXiv:1506.08347*, 2015. 5, 6, 7

[31] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-PIE. *Image and Vision Computing*, 2010. 5, 14

[32] Nitesh Gundavarapu, Divyansh Srivastava, Rahul Mitra, Abhishek Sharma, and Arjun Jain. Structured aleatoric uncertainty in human pose estimation. In *CVPR Workshops*, 2019. 2

[33] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Weinberger. On calibration of modern neural networks. In *ICML*, 2017. 1

[34] Ali Harakeh, Michael Smart, and Steven Waslander. BayesOD: A bayesian approach for uncertainty estimation in deep object detectors. *arXiv preprint arXiv:1903.03838*, 2019. 3

[35] Yihui He, Chenchen Zhu, Jianren Wang, Marios Savvides, and Xiangyu Zhang. Bounding box regression with uncertainty for accurate object detection. In *CVPR*, 2019. 3

[36] Changbo Hu, Jing Xiao, Iain Matthews, Simon Baker, Jeffrey Cohn, and Takeo Kanade. Fitting a single active appearance model simultaneously to multiple images. In *BMVC*, 2004. 2

[37] Eddy Ilg, Ozgun Cicek, Silvio Galesso, Aaron Klein, Osama Makansi, Frank Hutter, and Thomas Brox. Uncertainty estimates and multi-hypotheses networks for optical flow. In *ECCV*, 2018. 3

[38] Borui Jiang, Ruixuan Luo, Jiayuan Mao, Tete Xiao, and Yuning Jiang. Acquisition of localization confidence for accurate object detection. In *ECCV*, 2018. 3

[39] Licheng Jiao, Fan Zhang, Fang Liu, Shuyuan Yang, Lingling Li, Zhixi Feng, and Rong Qu. A survey of deep learning-based object detection. *arXiv preprint arXiv:1907.09408*, 2019. 4

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

[41] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *NeurIPS*, 2017. 2, 3

[42] Martin Koestinger, Paul Wohlhart, Peter Roth, and Horst Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *ICCV Workshops*, 2011. 5, 6

[43] Samuel Kotz, Tomaz Kozubowski, and Krzysztof Podgórski. Asymmetric multivariate laplace distribution. In *The Laplace distribution and generalizations*. 2001. 4

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

[45] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *NeurIPS*, 2017. 3

[46] Michael Le, Frederik Diehl, Thomas Brunner, and Alois Knol. Uncertainty estimation for deep neural object detectors in safety-critical applications. In *ITSC*, 2018. 1, 3

[47] Dan Levi, Liran Gispan, Niv Giladi, and Ethan Fetaya. Evaluating and calibrating uncertainty prediction in regression tasks. *arXiv preprint arXiv:1905.11659*, 2019. 3

[48] Wenbo Li, Zhicheng Wang, Binyi Yin, Qixiang Peng, Yuming Du, Tianzi Xiao, Gang Yu, Hongtao Lu, Yichen Wei, and Jian Sun. Rethinking on multi-stage networks for human pose estimation. *arXiv preprint arXiv:1901.00148*, 2019. 2

[49] Xiaoming Liu. Discriminative face alignment. *TPAMI*, 2008. 2

[50] Xiaoming Liu. Video-based face model fitting using adaptive active appearance model. *Image and Vision Computing*, 2010. 2

[51] Diogo Luvizon, David Picard, and Hedi Tabia. 2D/3D pose estimation and action recognition using multitask deep learning. In *CVPR*, 2018. 2, 3

[52] Xin Miao, Xiantong Zhen, Xianglong Liu, Cheng Deng, Vassilis Athitsos, and Heng Huang. Direct shape regression networks for end-to-end face alignment. In *CVPR*, 2018. 7

[53] Dimity Miller, Niko Sünderhauf, Haoyang Zhang, David Hall, and Feras Dayoub. Benchmarking sampling-based probabilistic object detectors. In *CVPR Workshops*, 2019. 3

[54] Lukas Neumann, Andrew Zisserman, and Andrea Vedaldi. Relaxed softmax: Efficient confidence auto-calibration for safe pedestrian detection. In *NeurIPS Workshops*, 2018. 3

[55] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016. 2

[56] Aiden Nibali, Zhen He, Stuart Morgan, and Luke Prendergast. Numerical coordinate regression with convolutional neural networks. *arXiv preprint arXiv:1801.07372*, 2018. 2, 8

[57] Ewa Nowara, Tim Marks, Hassan Mansour, and Ashok Veeraraghavany. SparsePPG: towards driver monitoring using camera-based vital signs estimation in near-infrared. In *CVPR Workshops*, 2018. 1

[58] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019. 12

[59] Shengju Qian, Keqiang Sun, Wayne Wu, Chen Qian, and Jiaya Jia. Aggregation via separation: Boosting facial landmark detector with semi-supervised style translation. In *ICCV*, 2019. 5, 6, 8, 16

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

[61] Joseph Robinson, Yuncheng Li, Ning Zhang, Yun Fu, and Sergey Tulyakov. Laplace landmark localization. In *ICCV*, 2019. 7

[62] Sami Romdhani, Shaogang Gong, and Ahaogang Psarrou. A multi-view nonlinear active shape model using kernel PCA. In *BMVC*, 1999. 2

[63] Christos Sagonas, Epameinondas Antonakos, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: Database and results. *Image and Vision Computing*, 2016. 5, 6, 14

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

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

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

[67] Kumar Shridhar, Felix Laumann, and Marcus Liwicki. A comprehensive guide to bayesian convolutional neural network with variational inference. *arXiv preprint arXiv:1901.02731*, 2019. 3

[68] Ke Sun, Yang Zhao, Borui Jiang, Tianheng Cheng, Bin Xiao, Dong Liu, Yadong Mu, Xinggang Wang, Wenyu Liu, and Jingdong Wang. High-resolution representations for labeling pixels and regions. *arXiv preprint arXiv:1904.04514*, 2019. 6, 7, 12, 15, 16

[69] Jaewon Sung and Daijin Kim. Adaptive active appearance model with incremental learning. *Pattern recognition letters*, 2009. 2

[70] Ying Tai, Yicong Liang, Xiaoming Liu, Lei Duan, Jilin Li, Chengjie Wang, Feiyue Huang, and Yu Chen. Towards highly accurate and stable face alignment for high-resolution videos. In *AAAI*, 2019. 2

[71] 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. 2

[72] Zhiqiang Tang, Xi Peng, Kang Li, and Dimitris Metaxas. Towards efficient U-Nets: A coupled and quantized approach. *TPAMI*, 2019. 2, 3, 6, 8, 12

[73] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *CVPR*, 2014. 2

[74] George Trigeorgis, Patrick Snape, Mihalis Nicolaou, Epameinondas Antonakos, and Stefanos Zafeiriou. Mnemonic descent method: A recurrent process applied for end-to-end face alignment. In *CVPR*, 2016. 5, 6

[75] Oncel Tuzel, Tim Marks, and Salil Tambe. Robust face alignment using a mixture of invariant experts. In *ECCV*, 2016. 2

[76] Oncel Tuzel, Fatih Porikli, and Peter Meer. Learning on lie groups for invariant detection and tracking. In *CVPR*, 2008. 2

[77] Georgios Tzimiropoulos. Project-out cascaded regression with an application to face alignment. In *CVPR*, 2015. 2

[78] Georgios Tzimiropoulos and Maja Pantic. Optimization problems for fast AAM fitting in-the-wild. In *ICCV*, 2013. 2

[79] Xinyao Wang, Liefeng Bo, and Li Fuxin. Adaptive wing loss for robust face alignment via heatmap regression. In *ICCV*, 2019. 2, 6, 8, 15, 16

[80] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *CVPR*, 2016. 8

[81] 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. 2, 5, 6, 7, 8, 16

[82] Wenyan Wu and Shuo Yang. Leveraging intra and interdataset variations for robust face alignment. In *CVPR Workshops*, 2017. 8, 16

[83] Xuehan Xiong and Fernando De la Torre. Supervised descent method and its applications to face alignment. In *CVPR*, 2013. 2

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

[85] Kwang Yi, Eduard Trulls, Vincent Lepetit, and Pascal Fua. Lift: Learned invariant feature transform. In *ECCV*, 2016. 3

[86] Stefanos Zafeiriou, George Trigeorgis, Grigorios Chrysos, Jiankang Deng, and Jie Shen. The Menpo facial landmark localisation challenge: A step towards the solution. In *CVPR Workshops*, 2017. 5, 6

[87] 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

[88] Shizhan Zhu, Cheng Li, Chen Loy, and Xiaoou Tang. Face alignment by coarse-to-fine shape searching. In *CVPR*, 2015. 2, 7, 8, 16

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

[90] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Li. Face alignment across large poses: A 3D solution. In *CVPR*, 2016. 5, 6

[91] Xu Zou, Sheng Zhong, Luxin Yan, Xiangyun Zhao, Jiahuan Zhou, and Ying Wu. Learning robust facial landmark detection via hierarchical structured ensemble. In *ICCV*, 2019. 6