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

Kumar, Abhinav; Marks, Tim; Mou, Wenxuan; Wang, Ye; Cherian, Anoop; Jones, Michael J.; Liu, Xiaoming; Koike-Akino, Toshiaki; Feng, Chen

TR2020-067 June 09, 2020

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.

IEEE Conference on Computer Vision and Pattern Recognition (CVPR)
LUVLi Face Alignment: Estimating Landmarks’ Location, Uncertainty, and Visibility Likelihood

Abhinav Kumar*,1, Tim K. Marks*,2, Wenxuan Mou*,3, Ye Wang2, Michael Jones2, Toshi Koike-Akino2, Anoop Cherian2, Xiaoming Liu4, Chen Feng5
abhinav3663@gmail.com, tmarks@merl.com, wenxuanmou@gmail.com,
ywang, mjones, koike, cherian@merl.com, liuxm@cse.msu.edu, cfeng@nyu.edu
1University of Utah, 2Mitsubishi Electric Research Labs (MERL), 3University of Manchester, 4Michigan State University, 5New 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.

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
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 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, 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 (DUNet) [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 $\ell_2$ 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 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 \( \mu_{ij} \), estimated covariance matrix \( \Sigma_{ij} \), and estimated visibility \( \tilde{v}_{ij} \) are tied together by the LUVLi loss function \( L_{ij} \), 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 parameterized: 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.

3.1. Mean Estimator

Let \( H_{jj}(x, y) \) denote the value at pixel location \((x, y)\) of the \( j \)th landmark’s heatmap from the \( i \)th U-net. The landmark’s location estimate \( \mu_{ij} = [\mu_{ijx}, \mu_{ijy}]^T \) is given by first post-processing the pixels of the heatmap \( H_{ij} \) with a function \( \sigma \), then taking the weighted spatial mean of the result (See (16) in the supplementary material). We considered three different functions for \( \sigma \): 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 \( \sigma \) 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 vari-
ables. 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 \( p_j \). By convention, if the landmark is not visible \( (v_j = 0) \), then \( p_j = \emptyset \), a special symbol indicating non-existence. Together, these variables are distributed according to an unknown distribution \( p(v_j, p_j) \). The marginal Bernoulli distribution \( p(v_j) \) captures the probability of visibility, \( p(p_j|v_j=1) \) denotes the distribution of the landmark location when it is visible, and \( p(p_j|v_j=0) = 1_{\emptyset}(p_j) \), where \( 1_{\emptyset} \) 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 \( z \) of each landmark \( j \) via

\[
q(v, z) = q_v(v)q_z(z|v),
\]

where \( q_v(v = 1) = \hat{v}_{ij} \), \( q_v(v = 0) = 1 - \hat{v}_{ij} \),

\[
\hat{v}_{ij} \text{ is the predicted probability of visibility, and}
\]

\[
q_z(z|v = 1) = \mathcal{P}(z|\mu_{ij}, \Sigma_{ij}), \quad q_z(z|v = 0) = \emptyset,
\]

where \( \mathcal{P} \) denotes the likelihood of the landmark being at location \( z \) given the estimated mean \( \mu_{ij} \) and covariance \( \Sigma_{ij} \).

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

\[
L_{ij} = -\ln q(v_j, p_j) = -\ln q_v(v_j) - \ln q_z(z|v_j) = -(1 - v_j) \ln (1 - \hat{v}_{ij}) - v_j \ln (\hat{v}_{ij}) - v_j \ln (\mathcal{P}(p_j|\mu_{ij}, \Sigma_{ij})),
\]

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 \( p_j \) 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

\[
\mathbb{E}[-\ln q(v_j, p_j)] = \mathbb{E}\left[ \ln \frac{p(v_j, p_j)}{q(v_j, p_j)} - \ln p(v_j, p_j) \right] = \text{DKL}(p(v_j, p_j)||q(v_j, p_j)) + \mathbb{E}[-\ln p(v_j, p_j)],
\]

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

Further, since

\[
\mathbb{E}[-\ln q(v_j, p_j)] = \mathbb{E}_{v_j \sim \ln(p_v)}[-\ln q(v_j)] + p_v \mathbb{E}_{p_j \sim \ln(p(v_j|v_j=1))}[-\ln \mathcal{P}(p_j|\mu_{ij}, \Sigma_{ij})],
\]

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

\[
\text{DKL}(p(v_j)||q(v_j)) + p_v \text{DKL}(p(p_j|v_j=1)||\mathcal{P}(z|\mu_{ij}, \Sigma_{ij}))
\]

### 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:

\[
\mathcal{P}(z|\mu_{ij}, \Sigma_{ij}) = \frac{\exp(-\frac{1}{2}(z-\mu_{ij})^T \Sigma_{ij}^{-1}(z-\mu_{ij}))}{2\pi \sqrt{\det(\Sigma_{ij})}}.
\]

Substituting (10) into (5), we have

\[
L_{ij} = -(1 - v_j) \ln (1 - \hat{v}_{ij}) - v_j \ln (\hat{v}_{ij}) + v_j \frac{1}{2} \ln |\Sigma_{ij}| + v_j \sqrt{3} \frac{(p_j - \mu_{ij})^T \Sigma_{ij}^{-1}(p_j - \mu_{ij})}{2T_2}
\]

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:

\[
\mathcal{P}(z|\mu_{ij}, \Sigma_{ij}) = \frac{e^{-\frac{1}{2}((z-\mu_{ij})^T \Sigma_{ij}^{-1}(z-\mu_{ij}))}}{2\pi \sqrt{\det(\Sigma_{ij})}}.
\]

Substituting (12) in (5), we have

\[
L_{ij} = -(1 - v_j) \ln (1 - \hat{v}_{ij}) - v_j \ln (\hat{v}_{ij}) + v_j \frac{1}{2} \ln |\Sigma_{ij}| + v_j \sqrt{3} \frac{(p_j - \mu_{ij})^T \Sigma_{ij}^{-1}(p_j - \mu_{ij})}{2T_2}
\]

In (13), \( T_2 \) is the 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 \( \Sigma_{ij} \) is the identity matrix and if all landmarks are assumed to be visible, then (11) simply reduces to the squared \( \ell_2 \) distance, and (13) reduces to the \( \ell_2 \) 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 \( 2 \times 2 \) covariance matrix \( \Sigma_{ij} \), which is symmetric positive definite. The three degrees of freedom of \( \Sigma_{ij} \) are captured by its Cholesky decomposition: a lower-triangular matrix \( L_{ij} \) such that \( \Sigma_{ij} = L_{ij} L_{ij}^T \). To estimate the elements of \( L_{ij} \), 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-
We call the new dataset AFLW\cite{Cao_2015_CVPR} dataset. In addition to landmark locations, ev-

eaching linear layer whose input is the bottleneck of the U-net (128 × 4 × 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 \( \mathbf{L}_{ij} \) of a covariance matrix must have positive diagonal elements, we pass the corresponding entries of the output through an ELU activation function\cite{clevert2015fast}, 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 \( v_j \), we add another fully connected linear layer whose input is the bottleneck of the U-net (128 × 4 × 4 = 2,048 dimensions) and output is an \( N_p \)-dimensional vector. This is passed through a sigmoid activation so the predicted visibility \( \hat{v}_{ij} \) 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 \( \mathbf{L}_{ij} \) 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:

\[
\mathcal{L} = \sum_{i=1}^{K} \lambda_i \mathbf{L}_i, \quad \text{where} \quad \mathbf{L}_i = \frac{1}{N_p} \sum_{j=1}^{N_p} \mathbf{L}_{ij} . \quad (14)
\]

At test time, each landmark’s mean and Cholesky coefficients are derived from the \( K \)th (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 uncer-
nainty, we now introduce a new dataset with entirely new, manual labels of over 19,000 face images from the AFLW\cite{Cao_2015_CVPR} 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 \((p_j = \emptyset)\). 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.

Existing Datasets. The most commonly used publicly available datasets for evaluation of 2D face alignment are summarized in Table 1. The 300-W dataset\cite{Cao_2015_CVPR} uses a 68-landmark system that was originally used for Multi-PIE\cite{munkhoeva2010multi}. Menpo 2D\cite{2017arXiv170202907K,swersky2015efficient} 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\cite{2017arXiv170202907K,torabi2017300} 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\cite{zhu2016face} and AFLW-68\cite{Cao_2015_CVPR} 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 ±120° yaw and ±90° pitch. Focusing on yaw, we group the images into five pose classes: frontal,
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-RA V 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-RA V dataset.
Training and testing protocols are described in the supple-
mental 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:

$$NME(\%) = \frac{1}{N_p} \sum_{j=1}^{N_p} \| \mathbf{p}_j - \mu_{K_j} \|_2 \times 100,$$  \hspace{1cm} (15)

where $v_j$, $p_j$, and $\mu_{K_j}$ respectively denote the visibility,
ground-truth and predicted location of landmark $j$ from the
$K$th (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
($\sqrt{w_{bbox} \cdot h_{bbox}}$), while NME$_{inter-ocular}$ [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].

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
NME$_{inter-ocular}$ [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 method (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 out-
performs 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.

![Figure 3: Unoccluded, externally occluded, and self-occluded landmarks.](image)

<table>
<thead>
<tr>
<th>Pose</th>
<th>Side</th>
<th>#Train</th>
<th>#Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frontal</td>
<td></td>
<td>8,778</td>
<td>2,195</td>
</tr>
<tr>
<td>Half-</td>
<td>Left half</td>
<td>1,800</td>
<td>295</td>
</tr>
<tr>
<td>Profile</td>
<td>Right half</td>
<td>1,221</td>
<td>306</td>
</tr>
<tr>
<td>Profile</td>
<td>Left</td>
<td>2,080</td>
<td>521</td>
</tr>
<tr>
<td>Profile</td>
<td>Right</td>
<td>2,190</td>
<td>548</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Pose</th>
<th>Side</th>
<th>#Train</th>
<th>#Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frontal</td>
<td></td>
<td>8,778</td>
<td>2,195</td>
</tr>
<tr>
<td>Half-</td>
<td>Left half</td>
<td>1,800</td>
<td>295</td>
</tr>
<tr>
<td>Profile</td>
<td>Right half</td>
<td>1,221</td>
<td>306</td>
</tr>
<tr>
<td>Profile</td>
<td>Left</td>
<td>2,080</td>
<td>521</td>
</tr>
<tr>
<td>Profile</td>
<td>Right</td>
<td>2,190</td>
<td>548</td>
</tr>
</tbody>
</table>

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

| Table 3: NME$_{inter-ocular}$ on 300-W Common, Challenge, and Full datasets (Split 1). [Key: Best, Second best] |
|-----------------|-----------------|-----------------|-----------------|
|                  | Common          | Challenge       | Full            |
|                 | NME$_{inter-ocular}$ (%) (↓) |                 |                 |
| 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 [65]     | 2.87            | 5.15            | 3.32            |
| HG-Hsle [91]    | 2.85            | 5.03            | 3.28            |
| A Wing [79]     | 2.72            | 4.52            | 3.07            |
| LUVLi (Ours)    | **2.76**        | 5.16            | **3.23**        |

Table 4: NME$_{box}$ and AUC$_{box}$ comparisons on 300-W Test (Split 2), Menpo 2D and COFW-68 datasets. [Key: Best, Second best, * = Pretrained on 300W-LP-2D]
Accuracy of Predicted Uncertainty. To evaluate the accuracy of the predicted uncertainty covariance matrix, $\Sigma_{K_j} = \begin{bmatrix} \Sigma_{K_{j_{xx}}} & \Sigma_{K_{j_{xy}}} \\ \Sigma_{K_{j_{yx}}} & \Sigma_{K_{j_{yy}}} \end{bmatrix}$, we compare all three unique terms of this prediction with the statistics of the residuals (2D error between the ground-truth location $p_j$ and the predicted location $\mu_{K_j}$) of all landmarks in the test set. We explain how we do this for $\Sigma_{K_{j_{xx}}}$ in Figure 4a. First, we bin every landmark of every test image according to the value of the predicted variance in the $x$-direction ($\Sigma_{K_{j_{xx}}}$). Each bin is represented by one point in the scatter plot. The $\Sigma_{K_{j_{xx}}}$ of landmarks within the bin is averaged to obtain a single estimate for the 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 $\Sigma_{xx} = \mathbb{E}[(p_{j_{x}} - \mu_{K_{j_{x}}})^2]$ for the bin. This mean squared residual error, $\Sigma_{xx}$, is plotted on the vertical axis. If our predicted uncertainties are accurate, this residual error, $\Sigma_{xx}$, should be roughly equal to the predicted uncertainty variance in the $x$-direction (plotted on the 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 the Pearson correlation coefficients of 0.98 and 0.99. Even better, the predicted and residual covariance values are roughly equal (fairly close to the line $y = x$), especially for smaller residuals.

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, $\text{NME}_{\text{box}}$, and average predicted uncertainty, $|\Sigma_{K_j}|^{1/2}$, 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 $\text{NME}_{\text{box}}$) 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 ($64 \times 64$) 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.

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 NME_{inter-ocular}, AUC^{10}_{inter-ocular}, and FR\(_{10}\). [Key: Best, Second best]

<table>
<thead>
<tr>
<th>Metric (%)</th>
<th>Method</th>
<th>All</th>
<th>Frontal</th>
<th>Half-Profile</th>
<th>Profile</th>
</tr>
</thead>
<tbody>
<tr>
<td>NME_{box}</td>
<td>CFSS [88]</td>
<td>9.07</td>
<td>0.366</td>
<td>20.56</td>
<td></td>
</tr>
<tr>
<td></td>
<td>DVNL [82]</td>
<td>10.84</td>
<td>0.456</td>
<td>10.84</td>
<td></td>
</tr>
<tr>
<td></td>
<td>LAB (w/B) [81]</td>
<td>5.27</td>
<td>0.532</td>
<td>7.56</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Wing [27]</td>
<td>5.11</td>
<td>0.554</td>
<td>6.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>DeCaFa (w/Da) [20]</td>
<td>4.62</td>
<td>0.563</td>
<td>4.84</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AVS [59]</td>
<td>4.39</td>
<td>0.591</td>
<td>4.98</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AWing [79]</td>
<td>4.46</td>
<td>0.572</td>
<td>2.84</td>
<td></td>
</tr>
<tr>
<td></td>
<td>LUVLi (Ours)</td>
<td>4.47</td>
<td>0.577</td>
<td>3.12</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>NME(_{box}^\tau)</th>
<th>Metric (%)</th>
<th>Method</th>
<th>All</th>
<th>Frontal</th>
<th>Half-Profile</th>
<th>Profile</th>
</tr>
</thead>
<tbody>
<tr>
<td>CFSS [88]</td>
<td>1.99</td>
<td>1.74</td>
<td>1.79</td>
<td>1.25</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DVNL [82]</td>
<td>1.61</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LAB (w/B) [81]</td>
<td>71.80</td>
<td>73.25</td>
<td>64.78</td>
<td>72.79</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wing [27]</td>
<td>77.08</td>
<td>75.33</td>
<td>74.69</td>
<td>82.10</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 8: NME_{box} and AUC_{box}^{\tau} comparisons on MERL-RAV dataset. [Key: Best]

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

<table>
<thead>
<tr>
<th>Metric (%)</th>
<th>Method</th>
<th>All</th>
<th>Frontal</th>
<th>Half-Profile</th>
<th>Profile</th>
</tr>
</thead>
<tbody>
<tr>
<td>NME_{box}</td>
<td>DU-Net [72]</td>
<td>1.99</td>
<td>1.89</td>
<td>2.50</td>
<td>1.92</td>
</tr>
<tr>
<td></td>
<td>LUVLi (Ours)</td>
<td>1.61</td>
<td>1.74</td>
<td>1.79</td>
<td>1.25</td>
</tr>
<tr>
<td>AUC_{box}^{\tau}</td>
<td>DU-Net [72]</td>
<td>71.80</td>
<td>73.25</td>
<td>64.78</td>
<td>72.79</td>
</tr>
<tr>
<td></td>
<td>LUVLi (Ours)</td>
<td>77.08</td>
<td>75.33</td>
<td>74.69</td>
<td>82.10</td>
</tr>
</tbody>
</table>

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. Accuracy (Visible) tests the accuracy of predicting that landmarks are visible when \(\hat{v}_j > 0.5\). The last two rows show the scalar measure of uncertainty, \(|\Sigma|^{1/2}\), both unnormalized and normalized by the face box size (\(|\Sigma|^{1/2}_{box}\)) 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 \(\Sigma_{ij} = I\)). 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 \(\sigma\): softmax and a temperature-scaled softmax (\(\tau\)-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.

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

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

Boru Jiang, Ruixuan Luo, Jiayuan Mao, Tete Xiao, and Yuning Jiang. Acquisition of localization confidence for accurate object detection. In ECCV, 2018. 3


Vahid Kazemi and Josephine Sullivan. One millisecond face alignment. In CVPR Workshops, 2014. 2


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


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


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

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


Samir Romdhani, Shaoang Gong, and Ahaogang Psarrou. A multi-view nonlinear active shape model using kernel PCA. In BMVC, 1999. 2


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

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

Patrick Sauer, Timothy Cootes, and Christopher Taylor. Accurate regression procedures for active appearance models. In BMVC, 2011. 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
[75] Oncel Tuzel, Tim Marks, and Salil Tambe. Robust face alignment using a mixture of invariant experts. In ECCV, 2016. 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
[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 inter-dataset 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
[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
[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
LUVLi Face Alignment: Estimating Landmarks’ Location, Uncertainty, and Visibility Likelihood

Supplementary Material

A1. Implementation Details

Images are cropped using the detector bounding boxes provided by the dataset 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 [7].

Training. We modified the PyTorch [58] code for DU-Net [72], keeping the number of U-nets $K = 8$ as in [72]. Unless otherwise stated, we use the 2D Laplacian likelihood (12) as our landmark location likelihood, and therefore we use (13) as our final loss function. All U-nets have equal weights $\lambda = 1$ in (14). For all datasets, visibility $v_j = 1$ is assigned to unoccluded landmarks (those that are not labeled as occluded) and to landmarks that are labeled as externally occluded. Visibility $v_j = 0$ is assigned to landmarks that are labeled as self-occluded and landmarks whose labels are missing.

Training images for 300-W Split 1 are augmented randomly using scaling $(0.75 - 1.25)$, rotation $(-30^\circ, -30^\circ)$ and color jittering $(0.6, 1.4)$ and random occlusion, as in [7].

The RMSprop optimizer is used as in [7, 72], with batch size 24. Training from scratch takes 100 epochs and starts with learning rate $2.5 \times 10^{-4}$, which is divided by 5, 2, and 2 at epochs 30, 60, and 90 respectively [72]. When we initialize from pretrained weights, we finetune for 50 epochs using the LUVLi loss: 20 with learning rate $10^{-4}$, followed by 30 with learning rate $2 \times 10^{-5}$. We consider the model saved in the last epoch as our final model.

Testing. Whereas heatmap based methods [7, 68, 72] adjust their pixel output with a quarter-pixel offset in the direction from the highest response to the second highest response, we use the spatial mean as the landmark location without carrying out any adjustment nor shifting the heatmap even by a quarter of a pixel. We do not need to implement a sub-pixel shift, because our spatial mean over the ReLUed heatmaps already performs sub-pixel location prediction.

Spatial Mean The spatial mean $\mu_{ij}$ of each of the heatmap is defined as

$$\mu_{ij} = \begin{bmatrix} \mu_{ijx} \\ \mu_{ijy} \end{bmatrix} = \frac{\sum_{x,y} \sigma(H_{ij}(x,y))}{\sum_{x,y} \sigma(H_{ij}(x,y))}$$

where $\sigma(H_{ij}(x,y))$ denotes the output of post-processing the heatmap pixel with a function $\sigma$.

A2. Additional Experiments and Results

We now provide additional results evaluating our system’s performance in terms of both localization and uncertainty estimation.

A2.1. System Trained on 300-W

A2.1.1 Training

For Split 1, we initialized using the pre-trained DU-Net model available from the authors of [72], then fine-tuned on the 300-W training set (Split 1) using our proposed architecture and LUVLi loss. For Split 2, for the experiments in which we pre-trained on 300W-LP-2D, we pre-trained from scratch on 300W-LP-2D using heatmaps (using the original DU-Net architecture and loss). We then fine-tuned on the 300-W training set (Split 2) using our proposed architecture and LUVLi loss.

A2.1.2 Comparison with KDN [13]

To compare directly with Chen et al. [13], in Figure 6 we plot normalized mean error (NME) vs. predicted uncertainty (rank, from smallest to largest), as in Figure 1 of [13]. (We obtained the predicted uncertainty and NME data of [13] from the authors.) The figure shows that for our method as well as for [13], there is a strong trend that higher predicted uncertainties correspond to larger location errors. However, the errors of our method are significantly smaller than the errors produced by [13].

A2.1.3 Verifying Predicted Uncertainty Distributions

For every image, for each landmark $j$, our network predicts a mean $\mu_{ij}$ and a covariance matrix $\Sigma_{ij}$. We can view this as our network predicting that a human labeler of that image will effectively select the landmark location $p_j$ for
Figure 6: Average NME vs sorted uncertainty, averaged across landmarks in an image.

Figure 7: Scatter plot of transformed ground-truth locations, $p_j' = \Sigma_{Kj}^{-0.5}(p_j - \mu_{Kj})$, for 300-W Test (Split 2). The histograms (orange) of their x and y coordinates are very close to the marginal pdf (black curves) of the Standard Laplacian distribution $P(z'|0, I)$.

that image from the Laplacian distribution from (12) with mean $\mu_{Kj}$ and covariance $\Sigma_{Kj}$:

$$p_j \sim P(z|\mu_{Kj}, \Sigma_{Kj}) = \frac{e^{-\frac{1}{\sqrt{3}}(z - \mu_{Kj})^T \Sigma_{Kj}^{-1}(z - \mu_{Kj})}}{2\pi \sqrt{|\Sigma_{Kj}|}}.$$  

(17)

If we had multiple labels (e.g., ground-truth landmark locations from multiple human labelers) for a single landmark in one image, then it would be straightforward to evaluate how well our method’s predicted probability distribution matches the distribution of labeled landmark locations. Unfortunately, face alignment datasets only have a single ground-truth location for each landmark in each image. This makes it difficult, but not impossible, to evaluate how well the human labels for images in the test set fit our method’s predicted uncertainty distributions. We propose the following method for verifying the predicted probability distributions.

Suppose we transform the ground-truth location of a landmark, $p_j$, using the predicted mean and covariance for that landmark as follows:

$$p_j' = \Sigma_{Kj}^{-0.5}(p_j - \mu_{Kj}).$$  

(18)

If our method’s predictions are correct, then from (17), $p_j \sim P(z|\mu_{Kj}, \Sigma_{Kj})$. Hence, $p_j'$ is drawn from the transformed distribution $P'(z')$, where $z' = \Sigma_{Kj}^{-0.5}(z - \mu_{Kj})$:

$$p_j' \sim P(z'|0, I) = \frac{e^{-\frac{1}{\sqrt{3}}z'^Tz'}}{2\pi^{3/2}}.$$  

(19)

After this simple transformation (transforming the labeled ground-truth location $p_j$ of each landmark using its predicted mean and covariance), we have transformed our network’s prediction about $p_j$ into a prediction about $p_j'$ that is much easier to evaluate, because the distribution in (19) is simply a standard 2D Laplacian distribution—it no longer depends on the predicted mean and covariance.

Thus, our method predicts that after the transformation (18), every ground-truth landmark location $p_j'$ is drawn from the same standard 2D Laplacian distribution (19). Now that we have an entire population of transformed labels that our model predicts are all drawn from the same distribution, it is easy to verify whether the labels fit our model’s predictions. Figure 7 shows a scatter plot of the transformed locations, $p_j'$, for all landmarks in all test images of 300-W (Split 2). We plot the histogram of the marginalized landmark locations ($x$- or $y$-coordinate of $p_j'$) in orange above and to the right of the plot, and overlay the marginal pdf of the standard Laplacian (19) in black. The excellent match between the transformed landmark locations and the standard Laplacian distribution indicates that our model’s predicted uncertainty distributions are quite accurate. Since Kullback-Leibler (KL) divergence is invariant to affine transformations like the one in (18), we can evaluate the KL-divergence (printed at the top of the scatterplot) between the standard 2D Laplacian distribution (19) and the distribution of the transformed landmark locations (using their 2D histograms) as a numerical measure of how well the predictions of our model fit the distribution of labeled locations.
A2.1.4 Relationship to Variation Among Human Labelers on Multi-PIE

We test our Split 2 model on 812 frontal face images of all subjects from the Multi-PIE dataset [31], then compute the mean of the uncertainty ellipses predicted by our model across all 812 images. To compute the mean, we first normalize the location of each landmark using the inter-ocular distance, as in [63], and also normalize the covariance matrix by the square of the inter-ocular distance. We then take the average of the normalized locations across all faces to obtain the mean landmark location. The covariance matrices are averaged across all faces using the log-mean-exponential technique. The mean location and covariance matrix of each landmark (averaged across all faces) is then used to plot the results which are shown on the right in Figure 8.

We compare our model predictions with Figure 5 of [63], shown on the left of Figure 8. To create that figure, [63] tasked three different human labelers with annotating the same frontal face images from the Multi-PIE database of 80 different subjects in frontal pose with neutral expression. For each landmark, they plotted the covariance of the label locations across the three labelers using an ellipse. Note the similarity between our model’s predicted uncertainties (on the right of Figure 8 and the covariance across human labelers (on the left of Figure 8), especially around the eyes, nose, and mouth. Around the outside edge of the face, note that our model predicts that label locations will vary primarily in the direction parallel to the edge, which is precisely the pattern observed across human labelers.

A2.1.5 Sample Uncertainty Ellipses on Multi-PIE

To illustrate how the predicted uncertainties output by our method vary across different subjects from Multi-PIE, in Figure 9 we overlay our model’s mean uncertainty predictions (in blue, copied from right side of Figure 8) with our model’s predicted uncertainties of some of the individual Multi-PIE face images (in various colors). To simplify the figure, we plot all landmarks except for the eyes, nose, and mouth.

Figure 8: Variation across three human labelers [63] (left) versus uncertainties computed by our proposed method on frontal images of Multi-PIE dataset (right).

Figure 9: Our model’s uncertainty predictions for some individual frontal face images from the Multi-PIE dataset (various colors), overlaid with the mean uncertainty predictions across all frontal Multi-PIE faces (blue, copied from Figure 8).

A2.1.6 Laplacian vs. Gaussian Likelihood

We have described two versions of our model: one whose loss function (13) uses a 2D Laplacian probability distribution (12), and another whose loss function (11) uses a 2D Gaussian probability distribution (10). We now discuss the question of which of these two models performs better.

The numerical comparisons are shown in Table 11. The numbers in the first two columns of the table were also presented in the ablation studies table, Table 10.

Comparing the Predicted Locations. If we consider only the errors of the predicted landmark locations, the first two columns of Table 11 show that the Laplacian model is slightly better: The Laplacian model has a smaller value of \( \text{NME}_{\text{box}} \) and a larger value of \( \text{AUC}_{\text{box}} \).

Comparing the Predicted Uncertainties. To compare the two models’ predicted uncertainties as well as their predicted locations, we consider the probability distributions over landmark locations that are predicted by each model. We want to know which model’s predicted probability distributions better explain the ground-truth locations of the landmarks in the test images. In other words, we want to know which model assigns a higher likelihood to the ground-truth landmark locations (i.e., which model yields a lower negative log-likelihood on the test data). We compute the negative log-likelihood of the ground-truth locations \( p_j \) from the last hourglass using (13) for the Lapla-
cian model and \((11)\) for the Gaussian model. The results, in the last column of Table 11, show that the Laplacian model gives a lower negative log-likelihood. In other words, the ground-truth landmark locations have a higher likelihood under our Laplacian model. We conclude that the learned Laplacian model explains the human labels better than the learned Gaussian model.

Table 11: Comparison of our model with Laplacian likelihood vs. with Gaussian likelihood, on 300-W Test (Split 2). [Key: \((\uparrow)\) = higher is better;\((\downarrow)\) = lower is better ]

<table>
<thead>
<tr>
<th>Likelihood</th>
<th>NME box (%) ((\downarrow))</th>
<th>AUC box (%) ((\uparrow))</th>
<th>NLL (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Laplacian</td>
<td>2.10</td>
<td>70.1</td>
<td>0.51</td>
</tr>
<tr>
<td>Gaussian</td>
<td>2.13</td>
<td>69.8</td>
<td>0.66</td>
</tr>
</tbody>
</table>

A2.2. WFLW Face Alignment

Data Splits and Implementation Details. The training set consists of 7,500 images, while the test set consists of 2,500 images. In Table 12, we report results on the entire test set (All), which we also reported in Table 7. In Table 12, we additionally report results on several subsets of the test set: large head pose (326 images), facial expression (314 images), illumination (698 images), make-up (206 images), occlusion (736 images), and blur (773 images). The images are cropped using the detector bounding boxes provided by [68] and resized to 256 × 256.

We first train the images with the heatmaps on proxy ground-truth heatmaps, then finetune using our proposed LUVLi loss. NME\textsubscript{inter-ocular}, AUC\textsubscript{inter-ocular}, and FR\textsubscript{inter-ocular} are used as evaluation metrics, as in [20, 68, 79]. We report AUC and FR with cutoff 10% as in [20, 68, 79].

Results of Facial Landmark Localization Table 12 compares our method’s landmark localization results with those of other state-of-the-art methods on the WFLW dataset. Our method performs best (tied with AWing [79]) on the NME metric, and performs in the top two methods on the other two metrics. Importantly, all of the other methods only predict landmark locations—they do not predict the uncertainty of their estimated landmark locations. Not only does our method place in the top two on all three landmark localization metrics, but our method also accurately predicts its own uncertainty of landmark localization.

A2.3. System Trained on MERL-RAV

Demo Video. We include a short demo video of our LUVLi model that was trained on our new MERL-RAV dataset. The video demonstrates our method’s ability to predict landmarks’ visibility (i.e., whether they are self-occluded) as well as their locations and uncertainty. We take a simple face video of someone turning his head from frontal to profile pose and run our method on each frame independently. Overlaid on each frame of video, we plot each estimated landmark location in yellow, and plot the predicted uncertainty as a blue ellipse. To indicate the predicted visibility of each landmark, we modulate the transparency of the landmark (of the yellow dot and blue ellipse). Landmarks whose predicted visibility is close to 1 are shown as fully opaque, while landmarks whose predicted visibility is close to zero are fully transparent (are not shown). Landmarks with intermediate predicted visibilities are shown as partially transparent.

In the video, notice that as the face approaches the profile pose, points on the far edge of the face begin to disappear, because the method correctly predicts that they are not visible (are self-occluded) when the face is in profile pose.

A2.4. More Qualitative Results

In Figure 10, we show example results on images from four datasets on which we tested.

A2.5. Examples from our MERL-RAV Dataset

Figure 11 shows several sample images from our MERL-RAV dataset. The ground-truth labels are overlaid on the images. On each image, unoccluded landmarks are shown in green, externally occluded landmarks are shown in red, and self-occluded landmarks are indicated by black circles in the face schematic on the right.

Acknowledgements

We would like to thank Lisha Chen from RPI, Zhiqiang Tang and Shijie Geng from Rutgers University, and Wei Tang from North Western University for their help during this project. We also had very useful discussions with Peng Gao from Chinese University of Hong Kong on the loss functions and Moitreya Chatterjee from University of Illinois Urbana-Champaign. We would also like to thank Adrian Bulat and Georgios Tzimiropoulos from the University of Nottingham for detailed discussions on getting bounding boxes for 300-W (Split 2).
Table 12: NME$_{\text{inter-ocular}}$ and AUC$_{\text{inter-ocular}}^{10}$ comparison between our proposed method and the state-of-the-art landmark prediction methods on the WFLW dataset.

[Key: Best, Second best; (w/DA) = uses more data; (w/B) = uses boundary; (↓) = smaller is better; (↑) = larger is better]

<table>
<thead>
<tr>
<th>Metric</th>
<th>Method</th>
<th>All</th>
<th>Head Pose</th>
<th>Expression</th>
<th>Illumination</th>
<th>Make-up</th>
<th>Occlusion</th>
<th>Blur</th>
</tr>
</thead>
<tbody>
<tr>
<td>NME$_{\text{inter-ocular}}$(%) (↓)</td>
<td>CFSS [88]</td>
<td>9.07</td>
<td>21.36</td>
<td>10.09</td>
<td>8.30</td>
<td>8.74</td>
<td>11.76</td>
<td>9.96</td>
</tr>
<tr>
<td></td>
<td>DVLN [82]</td>
<td>10.84</td>
<td>46.93</td>
<td>11.15</td>
<td>7.31</td>
<td>11.65</td>
<td>16.90</td>
<td>13.71</td>
</tr>
<tr>
<td></td>
<td>LAB (w/B) [81]</td>
<td>5.27</td>
<td>10.24</td>
<td>5.51</td>
<td>5.23</td>
<td>5.15</td>
<td>6.79</td>
<td>6.32</td>
</tr>
<tr>
<td></td>
<td>Wing [27]</td>
<td>5.11</td>
<td>8.75</td>
<td>5.36</td>
<td>4.93</td>
<td>5.41</td>
<td>6.37</td>
<td>5.81</td>
</tr>
<tr>
<td></td>
<td>DeCaFA (w/DA) [20]</td>
<td>4.62</td>
<td>8.11</td>
<td>4.65</td>
<td>4.41</td>
<td>4.63</td>
<td>5.74</td>
<td>5.38</td>
</tr>
<tr>
<td></td>
<td>HR-Net [68]</td>
<td>4.60</td>
<td>7.94</td>
<td>4.85</td>
<td>4.55</td>
<td>4.29</td>
<td>5.44</td>
<td>5.42</td>
</tr>
<tr>
<td></td>
<td>AVS [59]</td>
<td>4.39</td>
<td>8.42</td>
<td>4.68</td>
<td>4.24</td>
<td>4.37</td>
<td>5.60</td>
<td>4.86</td>
</tr>
<tr>
<td></td>
<td>AWing [79]</td>
<td>4.36</td>
<td>7.38</td>
<td>4.58</td>
<td>4.32</td>
<td>4.27</td>
<td>5.19</td>
<td>4.96</td>
</tr>
<tr>
<td></td>
<td>LUVLi (Ours)</td>
<td>4.37</td>
<td>7.56</td>
<td>4.77</td>
<td>4.30</td>
<td>4.33</td>
<td>5.29</td>
<td>4.94</td>
</tr>
<tr>
<td>AUC$_{\text{inter-ocular}}^{10}$(↑)</td>
<td>CFSS [88]</td>
<td>0.366</td>
<td>0.063</td>
<td>0.316</td>
<td>0.385</td>
<td>0.369</td>
<td>0.269</td>
<td>0.303</td>
</tr>
<tr>
<td></td>
<td>DVLN [82]</td>
<td>0.456</td>
<td>0.147</td>
<td>0.389</td>
<td>0.474</td>
<td>0.449</td>
<td>0.379</td>
<td>0.397</td>
</tr>
<tr>
<td></td>
<td>LAB (w/B) [81]</td>
<td>0.532</td>
<td>0.235</td>
<td>0.495</td>
<td>0.543</td>
<td>0.539</td>
<td>0.449</td>
<td>0.463</td>
</tr>
<tr>
<td></td>
<td>Wing [27]</td>
<td>0.554</td>
<td>0.310</td>
<td>0.496</td>
<td>0.541</td>
<td>0.558</td>
<td>0.489</td>
<td>0.492</td>
</tr>
<tr>
<td></td>
<td>DeCaFA (w/DA) [20]</td>
<td>0.563</td>
<td>0.292</td>
<td>0.546</td>
<td>0.579</td>
<td>0.575</td>
<td>0.485</td>
<td>0.494</td>
</tr>
<tr>
<td></td>
<td>AVS [59]</td>
<td>0.591</td>
<td>0.311</td>
<td>0.549</td>
<td>0.609</td>
<td>0.581</td>
<td>0.517</td>
<td>0.551</td>
</tr>
<tr>
<td></td>
<td>AWing [79]</td>
<td>0.572</td>
<td>0.312</td>
<td>0.515</td>
<td>0.578</td>
<td>0.572</td>
<td>0.502</td>
<td>0.512</td>
</tr>
<tr>
<td></td>
<td>LUVLi (Ours)</td>
<td>0.577</td>
<td>0.310</td>
<td>0.549</td>
<td>0.584</td>
<td>0.588</td>
<td>0.505</td>
<td>0.529</td>
</tr>
<tr>
<td>FR$_{\text{inter-ocular}}$(%) (↓)</td>
<td>CFSS [88]</td>
<td>20.56</td>
<td>66.26</td>
<td>23.25</td>
<td>17.34</td>
<td>21.84</td>
<td>32.88</td>
<td>23.67</td>
</tr>
<tr>
<td></td>
<td>DVLN [82]</td>
<td>10.84</td>
<td>46.93</td>
<td>11.15</td>
<td>7.31</td>
<td>11.65</td>
<td>16.30</td>
<td>13.71</td>
</tr>
<tr>
<td></td>
<td>LAB (w/B) [81]</td>
<td>7.56</td>
<td>28.83</td>
<td>6.37</td>
<td>6.73</td>
<td>7.77</td>
<td>13.72</td>
<td>10.74</td>
</tr>
<tr>
<td></td>
<td>Wing [27]</td>
<td>6.00</td>
<td>22.70</td>
<td>4.78</td>
<td>4.30</td>
<td>7.77</td>
<td>12.50</td>
<td>7.76</td>
</tr>
<tr>
<td></td>
<td>DeCaFA (w/DA) [20]</td>
<td>4.84</td>
<td>21.40</td>
<td>3.73</td>
<td>3.22</td>
<td>6.15</td>
<td>9.26</td>
<td>6.61</td>
</tr>
<tr>
<td></td>
<td>AVS [59]</td>
<td>4.08</td>
<td>18.10</td>
<td>4.46</td>
<td>2.72</td>
<td>4.37</td>
<td>7.74</td>
<td>4.40</td>
</tr>
<tr>
<td></td>
<td>AWing [79]</td>
<td>2.84</td>
<td>13.50</td>
<td>2.23</td>
<td>2.58</td>
<td>2.91</td>
<td>5.98</td>
<td>3.75</td>
</tr>
<tr>
<td></td>
<td>LUVLi (Ours)</td>
<td>3.12</td>
<td>15.95</td>
<td>3.18</td>
<td>2.15</td>
<td>3.40</td>
<td>6.39</td>
<td>3.23</td>
</tr>
</tbody>
</table>
Figure 10: Results of our LUVLi face alignment on example face images from four face datasets. Top row: 300-W. Second row: AFLW-19. Third row: WFLW. Bottom row: MERL-RAV. 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). In the MERL-RAV images (bottom row), the predicted visibility of each landmark controls its transparency. In particular, the predicted locations of landmarks with predicted visibility close to zero (such the points on the far side of the profile face in the third image of the bottom row) are 100% transparent (not shown).
Figure 11: Sample images from our MERL-RAV dataset with **unoccluded** landmarks shown in green, **externally occluded** landmarks shown in red, and self-occluded landmarks indicated by black circles in the face schematic on the right of each image.