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 $\lambda_i = 1$ 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 $(-30°, -30°)$ and color jittering $(0.6, 1.4)$ as in [72], while those from 300-W Split 2, AFLW-19, WFLW-98 and MERL-RAV datasets are augmented randomly using scaling $(0.8 - 1.2)$, rotation $(-50°, 50°)$, 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 2.5×10^{-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 2×10^{-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 $\boldsymbol{\mu}_{ij}$ of each of the heatmap is defined as

$$\boldsymbol{\mu}_{ij} = \begin{bmatrix} \mu_{ijx} \\ \mu_{ijy} \end{bmatrix} = \frac{\sum_{x,y} \sigma\big(\boldsymbol{H}_{ij}(x,y)\big) \begin{bmatrix} x \\ y \end{bmatrix}}{\sum_{x,y} \sigma\big(\boldsymbol{H}_{ij}(x,y)\big)} \tag{16}$$

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

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 $\boldsymbol{\mu}_{Kj}$ and a covariance matrix $\boldsymbol{\Sigma}_{Kj}$. We can view this as our network predicting that a human labeler of that image will effectively select the landmark location $\mathbf{p}_j$ for

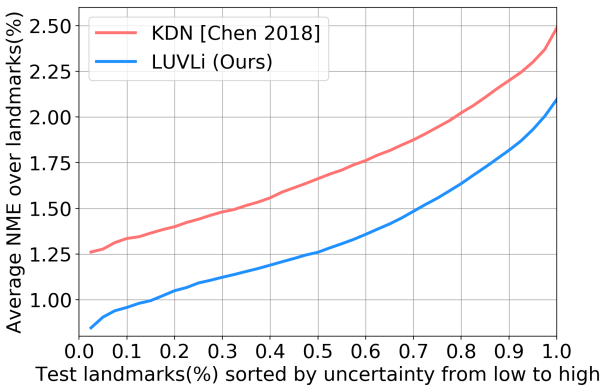

Figure 6: Average NME vs sorted uncertainty, averaged across landmarks in an image.

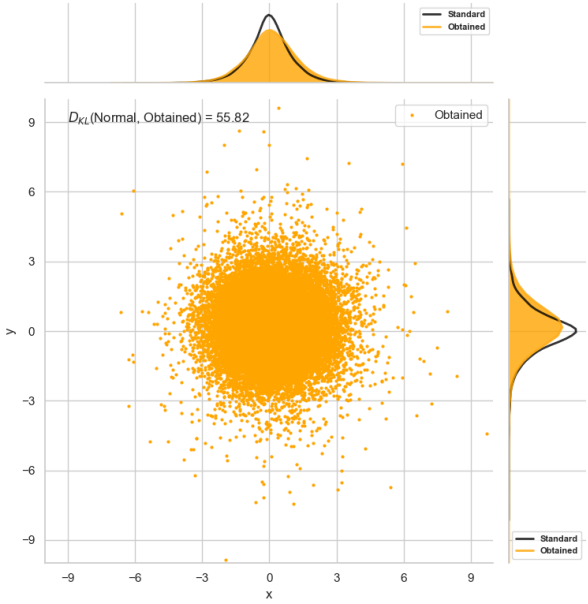

Figure 7: Scatter plot of transformed ground-truth locations, $\mathbf{p}'_j = \mathbf{\Sigma}_{Kj}^{-0.5}(\mathbf{p}_j - \boldsymbol{\mu}_{Kj})$, for 300-W Test (Split 2). The histograms (orange) of their x and y coordinates are very close to the the marginal pdf (black curves) of the Standard Laplacian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

that image from the Laplacian distribution from (12) with mean $\boldsymbol{\mu}_{Kj}$ and covariance $\mathbf{\Sigma}_{Kj}$:

$$\mathbf{p}_j \sim P(\mathbf{z}|\boldsymbol{\mu}_{Kj}, \mathbf{\Sigma}_{Kj}) = \frac{e^{-\sqrt{3(\mathbf{z}-\boldsymbol{\mu}_{Kj})^T \mathbf{\Sigma}_{Kj}^{-1}(\mathbf{z}-\boldsymbol{\mu}_{Kj})}}}{\frac{2\pi}{3}\sqrt{|\mathbf{\Sigma}_{Kj}|}}. \tag{17}$$

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, $\mathbf{p}_j$, using the predicted mean and covariance for that landmark as follows:

$$\mathbf{p}'_j = \mathbf{\Sigma}_{Kj}^{-0.5}(\mathbf{p}_j - \boldsymbol{\mu}_{Kj}). \tag{18}$$

If our method's predictions are correct, then from (17), $\mathbf{p}_j \sim P(\mathbf{z}|\boldsymbol{\mu}_{Kj}, \mathbf{\Sigma}_{Kj})$. Hence, $\mathbf{p}'_j$ is drawn from the transformed distribution $P(\mathbf{z}')$, where $\mathbf{z}' = \mathbf{\Sigma}_{Kj}^{-0.5}(\mathbf{z} - \boldsymbol{\mu}_{Kj})$:

$$\mathbf{p}'_j \sim P(\mathbf{z}'|\mathbf{0}, \mathbf{I}) = \frac{e^{-\sqrt{3\mathbf{z}'^T\mathbf{z}'}}}{2\pi/3}. \tag{19}$$

After this simple transformation (transforming the labeled ground-truth location $\mathbf{p}_j$ of each landmark using its predicted mean and covariance), we have transformed our network's prediction about $\mathbf{p}_j$ into a prediction about $\mathbf{p}'_j$ 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 $\mathbf{p}'_j$ 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, $\mathbf{p}'_j$, 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 $\mathbf{p}'_j$) 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 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.

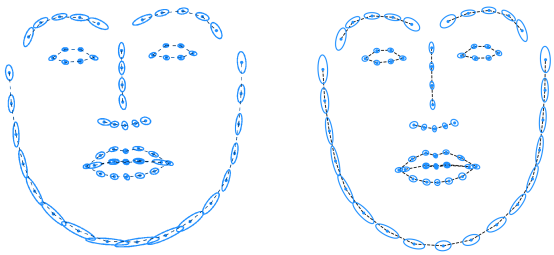

Figure 8: Variation across three human labelers [63] (*left*) versus uncertainties computed by our proposed method on frontal images of Multi-PIE dataset (*right*).

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.

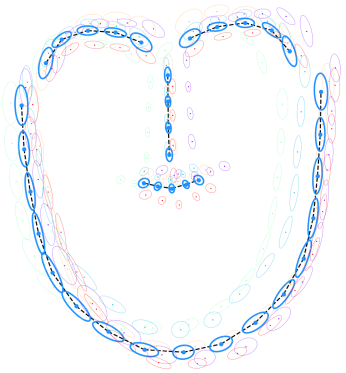

Figure 9: Our model's uncertainty predictions for some individual frontal face images from the Multi-PIE dataset (various colors), overlaid with the mean uncertainty predictions across all frontal Multi-PIE faces (blue, copied from Figure 8).

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 NME_{box} and a larger value of $\text{AUC}^{7}_{\text{box}}$.

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 $\mathbf{p}_j$ from the last hourglass using (13) for the Lapla-

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

Likelihood	$\text{NME}_{\text{box}}(\%)$ (↓)	$\text{AUC}^{7}_{\text{box}}(\%)$ (↑)	NLL (↓)
Laplacian	2.10	70.1	0.51
Gaussian	2.13	69.8	0.66

Table 12: $\text{NME}^{\text{vis}}_{\text{box}}$ comparisons on MERL-RAV dataset. [Key: **Best**]

Metric (%)	Method	All	Frontal	Half-Profile	Profile
$\text{NME}^{\text{vis}}_{\text{box}}$(↓)	DU-Net [72]	2.27	1.91	2.77	3.10
	LUVLi (Ours)	**1.84**	**1.75**	**1.99**	**2.03**
NME_{box}(↓)	DU-Net [72]	1.99	1.89	2.50	1.92
	LUVLi (Ours)	**1.61**	**1.74**	**1.79**	**1.25**

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.

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 13, we report results on the entire test set (All), which we also reported in Table 7. In Table 13, 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. $\text{NME}_{\text{inter-ocular}}$, $\text{AUC}^{10}_{\text{inter-ocular}}$, and $\text{FR}^{10}_{\text{inter-ocular}}$ 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 13 compares our method's landmark localization results with those of other state-of-the-art methods on the WFLW dataset. Our method performs performs in the top two methods on all the 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. MERL-RAV Face Alignment

We next define a modified version of the evaluation metric NME that may be more appropriate for face images with extreme head pose. Whereas NME as defined in (15) divides by the total number of landmarks N_p, the modified NME instead divides by the number of visible landmarks. This metric, which we call NME^{vis}, computes the mean across only the visible (unoccluded and externally occluded) landmarks:

$$\text{NME}^{\text{vis}}(\%) = \frac{1}{\sum_j v_j} \sum_{j=1}^{N_p} v_j \frac{\|\mathbf{p}_j - \boldsymbol{\mu}_{Kj}\|_2}{d} \times 100, \quad (20)$$

If all of the facial landmarks are visible, then this reduces to our previous definition of NME (15).

We define $\text{NME}^{\text{vis}}_{\text{box}}$ as the special case of NME^{vis} in which the normalization d is set to the geometric mean of the width and height of the ground-truth bounding box $\left(\sqrt{w_{\text{bbox}} \cdot h_{\text{bbox}}}\right)$, as in NME_{box} [7, 14, 86]. Results for all head poses on MERL-RAV dataset using the metric $\text{NME}^{\text{vis}}_{\text{box}}$ are shown in Table 12. We also repeat the NME_{box} numbers from Table 8. Clearly, the $\text{NME}^{\text{vis}}_{\text{box}}$ and NME_{box} numbers are very close for the frontal subsets but are different for half-profile and profile subsets. This is because half-profile and (especially) profile face images have fewer visible landmarks (more self-occluded landmarks), which causes the denominator in (20) to be smaller for these images.

A2.4. Additional Qualitative Results

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

A2.5. Video Demo of LUVLi

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 a person 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.

Table 13: $\text{NME}_{\text{inter-ocular}}$ and $\text{AUC}^{10}_{\text{inter-ocular}}$ comparison between our proposed method and the state-of-the-art landmark localization methods on the WFLW dataset.
[Key: **Best**, **Second best**; (w/DA) = uses more data; (w/B) = uses boundary; (↓) = smaller is better; (↑) = larger is better]

Metric	Method	All	Head Pose	Expression	Illumination	Make-up	Occlusion	Blur
$\text{NME}_{\text{inter-ocular}}(\%)$ (↓)	CFSS [88]	9.07	21.36	10.09	8.30	8.74	11.76	9.96
	DVLN [82]	10.84	46.93	11.15	7.31	11.65	16.30	13.71
	LAB (w/B) [81]	5.27	10.24	5.51	5.23	5.15	6.79	6.32
	Wing [27]	5.11	8.75	5.36	4.93	5.41	6.37	5.81
	DeCaFA (w/DA) [20]	4.62	8.11	**4.65**	4.41	4.63	5.74	5.38
	HR-Net [68]	4.60	7.94	4.85	4.55	**4.29**	5.44	5.42
	AVS [59]	4.39	8.42	4.68	**4.24**	4.37	5.60	**4.86**
	AWing [79]	**4.36**	**7.38**	**4.58**	4.32	**4.27**	**5.19**	4.96
	LUVLi (Ours)	**4.37**	**7.56**	4.77	**4.30**	4.33	**5.29**	**4.94**
$\text{AUC}^{10}_{\text{inter-ocular}}(\uparrow)$	CFSS [88]	0.366	0.063	0.316	0.385	0.369	0.269	0.303
	DVLN [82]	0.456	0.147	0.389	0.474	0.449	0.379	0.397
	LAB (w/B) [81]	0.532	0.235	0.495	0.543	0.539	0.449	0.463
	Wing [27]	0.554	0.310	0.496	0.541	0.558	0.489	0.492
	DeCaFA (w/DA) [20]	0.563	0.292	0.546	0.579	0.575	0.485	0.494
	AVS [59]	**0.591**	**0.311**	**0.549**	**0.609**	**0.581**	**0.517**	**0.551**
	AWing [79]	0.572	**0.312**	**0.515**	0.578	0.572	0.502	0.512
	LUVLi (Ours)	**0.577**	0.310	**0.549**	**0.584**	**0.588**	**0.505**	**0.525**
$\text{FR}^{10}_{\text{inter-ocular}}(\%)$ (↓)	CFSS [88]	20.56	66.26	23.25	17.34	21.84	32.88	23.67
	DVLN [82]	10.84	46.93	11.15	7.31	11.65	16.30	13.71
	LAB (w/B) [81]	7.56	28.83	6.37	6.73	7.77	13.72	10.74
	Wing [27]	6.00	22.70	4.78	4.30	7.77	12.50	7.76
	DeCaFA(w/DA) [20]	4.84	21.40	3.73	3.22	6.15	9.26	6.61
	AVS [59]	4.08	18.10	4.46	2.72	4.37	7.74	4.40
	AWing [79]	**2.84**	**13.50**	**2.23**	**2.58**	**2.91**	**5.98**	**3.75**
	LUVLi (Ours)	**3.12**	**15.95**	**3.18**	**2.15**	**3.40**	**6.39**	**3.23**

A2.6. 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 to the right of the image.

Acknowledgements

We would like to thank Lisha Chen from RPI for providing results from their method and Zhiqiang Tang and Shijie Geng from Rutgers University for providing their pre-trained models on 300-W (Split 1). 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). 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 are also grateful to Maitrey Mehta from the University of Utah who volunteered for the demo. We also acknowledge anonymous reviewers for their feedback that helped in shaping the final manuscript.

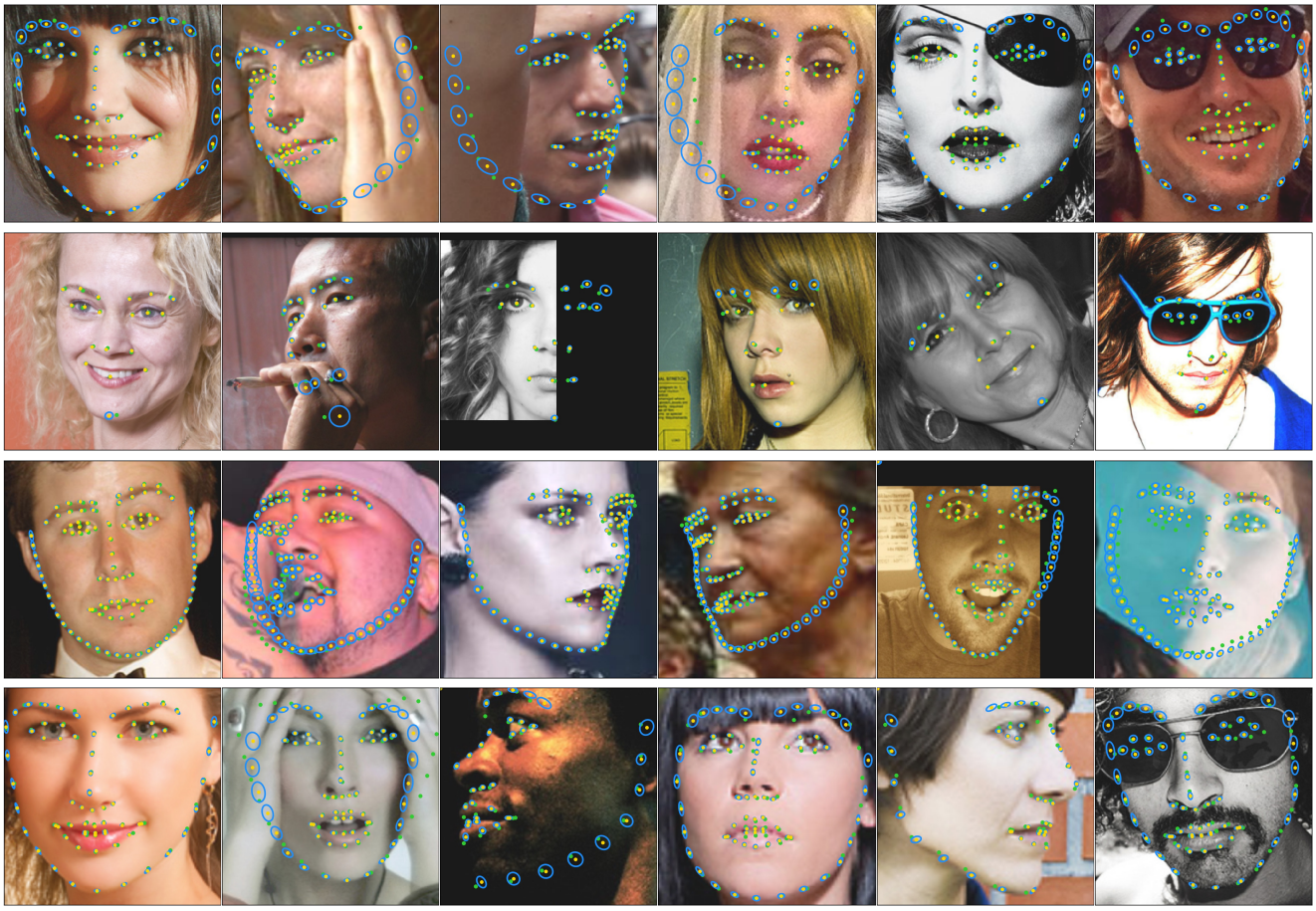

Figure 10: Results of our LUVLi face alignment on example face images from four face datasets. *Top row:* 300-W. *Second row:* AFLW-19. *Third row:* WFLW. *Bottom row:* MERL-RAV. Ground-truth (green) and predicted (yellow) landmark locations are shown. The estimated uncertainty of the predicted location 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.