Face Alignment with Kernel Density Deep Neural Network

Lisha Chen[1], Hui Su[1,2], Qiang Ji[1]
[1]Rensselaer Polytechnic Institute, [2]IBM Research
chenl21@rpi.edu, huisuibmres@us.ibm.com, qji@ecse.rpi.edu

Abstract

Deep neural networks achieve good performance in many computer vision problems such as face alignment. However, when the testing image is challenging due to low resolution, occlusion or adversarial attacks, the accuracy of a deep neural network suffers greatly. Therefore, it is important to quantify the uncertainty in its predictions. A probabilistic neural network with Gaussian distribution over the target is typically used to quantify uncertainty for regression problems. However, in real-world problems especially computer vision tasks, the Gaussian assumption is too strong. To model more general distributions, such as multimodal or asymmetric distributions, we propose to develop a kernel density deep neural network. Specifically, for face alignment, we adapt state-of-the-art hourglass neural network into a probabilistic neural network framework with landmark probability map as its output. The model is trained by maximizing the conditional log likelihood. To exploit the output probability map, we extend the model to multi-stage so that the logits map from the previous stage can feed into the next stage to progressively improve the landmark detection accuracy. Extensive experiments on benchmark datasets against state-of-the-art unconstrained deep learning method demonstrate that the proposed kernel density network achieves comparable or superior performance in terms of prediction accuracy. It further provides aleatoric uncertainty estimation in predictions.

1. Introduction

Face alignment, or facial landmark localization, is a fundamental step for facial behavior analysis such as face recognition, facial expression estimation, and head pose estimation. Classical work on face alignment mainly adopts a cascade regression framework based on local image features [50, 49, 48, 47], which is sensitive to initialization and achieves limited performance on challenging dataset. With the introduction of deep learning based methods for regression and feature representation learning, the state-of-the-art accuracy in face alignment is achieved [41, 52, 43, 3, 11, 46].

However, the performance of face alignment remains sensitive to face image quality. Challenges such as large head pose, object occlusion or low resolution may lead to poor landmark detection results. Moreover, existing deep learning based methods are susceptible to small image perturbations such as an adversarial attack, which may result in a large difference in the prediction. More importantly, existing deep learning methods cannot predict their output uncertainty. Therefore, it is important to develop a probabilistic deep neural network to quantify the prediction uncertainty and to avoid making over-confident wrong decisions.

To these goals, we propose a Kernel Density Deep Neural Network (KDN). Different from the deterministic approach that gives a point estimation for each input, our model outputs target probability distribution for each input. Moreover, rather than assuming the output follows a Gaussian distribution, the proposed method can capture more general probability distribution, such as multimodal or asymmetric distribution. With the target probability distribution, we can quantify the prediction confidence to distinguish challenging input image caused by large head pose, object occlusion or low resolution and to identify landmarks under occlusion. To further exploit the output probability map, we further extend our model to multi-stage cascade framework so that the probability map produced in the last stage can serve as input to guide the detection in the next stage to progressively improve the landmark detection accuracy.

The contributions of work are summarized as follows:
1) We introduce the Kernel Density Deep Neural Network that produces target probability map, without assuming a specific parametric distribution. The probability map can be used to quantify the uncertainty of the output and to identify the challenging landmarks. And we further extend our model to multiple stages to use the output probability map to progressively improve the landmark detection.
2) We show that the estimated uncertainty in our method can be used to detect occluded landmarks without occlusion supervision.
3) We show the proposed method can be generally extended to other regression tasks such as action unit intensity estimation.

2. Related work

2.1. Probabilistic Neural Network

To quantify uncertainty in neural networks, the probabilistic neural network was proposed to model the conditional target probability distribution given its input and neural networks are used to predict the parameters of the probability distribution. For regression tasks, it is often assumed that the target follows Gaussian distribution [29, 21, 19]. And neural network is used to predict the mean and variance for the Gaussian distribution. In this way, the prediction is given by the mean and the uncertainty is quantified by the variance.

However, for many real-world problems, the target distribution may be more complicated, *e.g*. asymmetric or multimodal, which the Gaussian distribution cannot adequately capture. To deal with this issue, one way is to parameterize a different distribution suitable for the specific problem. For instance, [28] used a Gamma distribution to model the distribution of surgery duration. And [34] used a von Mise distribution to model the distribution of object pose. Another way is to use a mixture distribution which is more flexible in approximating distributions with different shapes. For instance, [1] used a mixture of Gaussian distribution and [34] extended the von Mises distribution to a mixture of von Mises distribution to handle multimodal distribution. These methods typically still have assumptions and are not generally applicable.

We are interested in modeling the distribution of landmark location, which has high probability near the boundary of the facial parts. Therefore the distribution typically does not follow some standard parametric distribution such as Gaussian. And different landmarks may have very different distribution shape.

2.2. Face Alignment

Face alignment is typically treated as a regression problem where given a face image, it aims to localize certain facial key points in the image. Classic methods lie in the categories of Active Shape Model (ASM) [26], Active Appearance Model (AAM) [10, 18, 25, 37], Constrained Local Model (CLM) [20, 38] and Cascade Regression [7, 4, 54, 5, 49]. ASM models the statistical shape of objects, while AAM models both the shape and the appearance features. CLM is similar to AAM that models shape prior using principal component analysis (PCA) that projects both local appearance features and shape features onto the bases. Cascade Regression refines landmark localization stage by stage. These classic methods rely on hand-crafted local image features and are usually sensitive to initializations. They are outperformed by deep learning based methods which use deep feature representation.

Deep learning based method for face alignment was first proposed in [41] and achieved better performance than classical methods. Later on, more works on face alignment using deep learning framework has been explored [53, 52, 43, 12] but they are all based on coordinate regression.

Until recently, fully convolutional neural network (FCN) [23] based methods established new state-of-the-art for face alignment and body pose estimation [42, 27, 45]. And most of these face alignment methods [3, 46] follow the architecture of Stacked Hourglass [27]. The stacked modules refine the network predictions after each stack, similar to the idea of Cascade Regression. Instead of directly predicting the landmark coordinates, it predicts a heatmap with same size as the input image and the landmark location is predicted by the coordinate on the heatmap with largest response. The idea of the heatmap based regression is similar to a fully convolutional neural network which preserves the spatial information of the input image and reduce the parameters brought by fully connected layers.

2.3. Fully Convolutional Network Loss

State-of-the-art face alignment methods adopt FCN structure with a heatmap regression loss. The loss function is typically defined as the mean squared error between the predicted heatmap and the ground truth heatmap. This loss function is originally introduced and widely used in human pose estimation [42, 27, 2, 8, 6]. Besides this typical loss function, there are several other options in literature introduced for solving other problems such as image segmentation.

One choice is to treat the problem as a multi-class classification problem where each pixel location in the heatmap corresponds to one class and use the softmax cross entropy loss over the 2D heatmap. This was used in Mask-RCNN [16] for human body joint estimation. One pixel location with highest probability in the heatmap is selected as the estimation. Solving a regression problem using softmax cross entropy loss also exists in other problems such as face age estimation [31], where we have discrete age labels. These age labels, though discrete, are not independent because labels with close values should be more confusing with each other in classification. Therefore, to some extent, this loss function abandons part of the information provided by the label values. To address this issue, the paper [31] further uses L2 loss of the mean computed from the softmax probability. This idea has been explored in other tasks such as body pose estimation [40], headpose estimation [35]. And [13] further proposes to use L1 loss instead of L2 loss in face age estimation, while [44] proposes to apply wing loss to heatmap regression that is first proposed in [12] for traditional coordinate regression in face alignment. These works, while achieving satisfactory performance in terms of low prediction error, cannot accurately quantify prediction uncertainty. More recently, [15] proposes to estimate covariance matrix besides the mean of a multivariate Gaussian distribution, thus incorporating uncertainty into this framework. All the aforementioned works that add L1, L2 or Gaussian negative log likelihood losses besides the heatmap loss assumes the output distribution as single-modal, implicitly or explicitly. However, we show in this paper that this is often not the case in real-world computer vision problems. Therefore using the mean for inference will lead to erroneous predictions for multimodal cases.

Another choice is to treat the problem as a classification problem for each pixel [32, 17, 33]. There are two types of classification, one is binary classification, where each pixel will be classified as the target or not. And the ground truth binary label is created by assigning 1 to all the pixels in a certain neighborhood around the ground truth target pixel location and 0 to the rest pixels in the heatmap. The other choice is multi-class classification [9], where each pixel is classified as either one of the body or facial part regions or the background. These loss functions are used very often for segmentation tasks. And since it usually uses a softmax or sigmoid cross entropy loss which is also the negative log likelihood of a Categorical distribution, it is able to quantify classification uncertainty. However, they do not achieve as good performance as the previous loss functions as studied in [40] and it is difficult to define the ground truth body or facial part regions given only a single pixel location as the ground truth keypoint location.

Therefore, in this work, we propose a different loss function and a corresponding inference method that achieves state-of-the-art performance and provide good aleatoric uncertainty estimation. Our work differ from previous works in terms of explicitly quantifying uncertainty in fully-convolutional based architecture without adding additional fully connected layers to predicted covariance as [15].

3. The Proposed Method

Our method is built on the probabilistic neural network framework. We assume target $\mathbf{y}$ (landmark coordinates) is a random vector that follows $p(\mathbf{y} \mid \mathbf{x}; \mathbf{\Theta})$, where $\mathbf{x}$ is the input image and $\mathbf{\Theta}$ is the neural network parameter. And $p(\mathbf{y} \mid \mathbf{x}; \mathbf{\Theta})$ is parameterized by the neural network output.

3.1. Kernel Density Network

Instead of assuming the target follows Gaussian distribution as being done by current models, we propose to model the target probability with multi-variate kernel density function [39] in order to capture more general probability distributions, including multimodal and non-symmetric distributions.

Denote m, n as the height and width of the output $\boldsymbol{\pi}(\mathbf{x}; \mathbf{\Theta})$ from the Hourglass module, $\boldsymbol{\mu}_{ij} = [i, j]^T$ as the pixel location in the output map, where $1 \leq i \leq m, 1 \leq j \leq n$. According to the multivariate kernel density distribution [39], the target distribution can be expressed as

$$p(\mathbf{y} \mid \mathbf{x}; \mathbf{\Theta}) = \sum_{i=1}^{m} \sum_{j=1}^{n} K_{\mathbf{\Sigma}}(\mathbf{y} - \boldsymbol{\mu}_{ij}) \boldsymbol{\pi}_{ij}(\mathbf{x}; \mathbf{\Theta}) \quad (1)$$

where $K_{\mathbf{\Sigma}}(\mathbf{y} - \boldsymbol{\mu}_{ij})$ is a Gaussian kernel whose value is the standard 2D Gaussian's probability density at $\mathbf{\Sigma}^{-\frac{1}{2}}(\mathbf{y} - \boldsymbol{\mu}_{ij})$ normalized by $|\mathbf{\Sigma}|^{-\frac{1}{2}}$, *i.e.* $K_{\mathbf{\Sigma}}(\mathbf{y} - \boldsymbol{\mu}_{ij}) = |\mathbf{\Sigma}|^{-\frac{1}{2}} \Phi\big(\mathbf{\Sigma}^{-\frac{1}{2}}(\mathbf{y} - \boldsymbol{\mu}_{ij})\big)$. $\boldsymbol{\pi}(\mathbf{x}; \mathbf{\Theta})$, the output of the neural network, is a weight map of the dimension $m \times n$, where each pixel value $\boldsymbol{\pi}_{ij}(\mathbf{x}; \mathbf{\Theta})$ represents the weight of the Gaussian kernel $K_{\mathbf{\Sigma}}(\mathbf{y} - \boldsymbol{\mu}_{ij})$, $0 \leq \boldsymbol{\pi}_{ij}(\mathbf{x}; \mathbf{\Theta}) \leq 1$ and $\sum_{i=1}^{m} \sum_{j=1}^{n} \boldsymbol{\pi}_{ij}(\mathbf{x}; \mathbf{\Theta}) = 1$.

Thus we form a continuous probability $p(\mathbf{y} \mid \mathbf{x}; \mathbf{\Theta})$ based on the Gaussian kernels $K_{\mathbf{\Sigma}}(\mathbf{y} - \boldsymbol{\mu}_{ij})$ and their corresponding weights $\boldsymbol{\pi}_{ij}(\mathbf{x}; \mathbf{\Theta})$.

It is worth noting that the form of $p(\mathbf{y} \mid \mathbf{x}; \mathbf{\Theta})$ depends on our choice of the kernel function. If we choose a uniform kernel with a range of 1 pixel, it is equivalent to the likelihood for a categorical distribution where each category represents the discrete landmark coordinate. Here we choose a Gaussian kernel to achieve a smoothing effect similar to kernel density estimation.

In this way, we only change the loss function of the neural network for face alignment problem without modifying the heatmap regression based network structure. The goal is to maximize the conditional likelihood without assuming any specific distribution of the target, unlike widely practiced loss function which puts a fixed Gaussian heatmap around the ground truth label as the ground truth heatmap and minimize the L-2 distance between the groundtruth heatmap and the predicted one.

Loss function. The loss function is defined as the negative log conditional likelihood. Given training data $\mathcal{D} = \{\mathbf{x}_k, \mathbf{y}_k \mid k = 1, 2, \ldots, N\}$, we minimize the loss function to get $\mathbf{\Theta}^*$ as shown in Eq.(2).

$$\begin{aligned} \mathbf{\Theta}^* &= \arg\min_{\mathbf{\Theta}} - \sum_{k=1}^{N} \log p(\mathbf{y}_k \mid \mathbf{x}_k; \mathbf{\Theta}) \\ &= \arg\min_{\mathbf{\Theta}} - \sum_{k=1}^{N} \log \sum_{i=1}^{m} \sum_{j=1}^{n} K_{\mathbf{\Sigma}}(\mathbf{y}_k - \mathbf{y}_{ij}) \boldsymbol{\pi}_{ij}(\mathbf{x}_k; \mathbf{\Theta}) \end{aligned} \quad (2)$$

To demonstrate why the proposed loss function based on kernel density benefits the learning process of face alignment, we compute the gradient of the loss w.r.t. the layer before softmax. To simplify the notation, let $w_{kij} = K_{\mathbf{\Sigma}}(\mathbf{y}_k - \mathbf{y}_{ij})$. Denote the layer before softmax for the sample k as f_{kij}, and the layer after softmax as p_{kij}, $p_{kij} = softmax(f_{kij})$. The derivative of the loss contributed by a training sample $\{\mathbf{x}_k, \mathbf{y}_k\}$ can be computed as

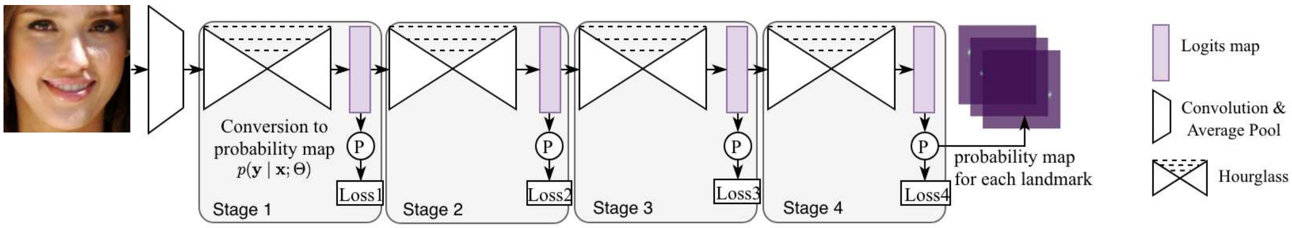


Figure 1: The proposed cascade network structure using Hour Glass module as basic structure, same as [3]. The size of the input image is 256×256. The size of the probability map is 64×64. The loss function for each stage after the Hour Glass module is based on minimizing the negative log conditional likelihood.

$$\frac{\partial Loss_k}{\partial f_{kij}} = \frac{p_{kij}\left(w_{kij} - \sum_{a=1}^{m}\sum_{b=1}^{n} w_{kab}p_{kab}\right)}{\sum_{a=1}^{m}\sum_{b=1}^{n} w_{kab}p_{kab}} \tag{3}$$

where $w_{kij} > 0$ measures the similarity between each pixel location in the heatmap and the ground truth landmark location. The closer the pixel location $[i, j]^T$ is to the ground truth $\mathbf{y}_k$, the higher w_{kij}. $\sum_{a=1}^{m}\sum_{b=1}^{n} w_{kab}p_{kab}$ is the expectation of the similarity over discrete probability distribution $\boldsymbol{\pi}(\mathbf{x}_k; \boldsymbol{\Theta})$.

During training, if $w_{kij} > \sum_{a=1}^{m}\sum_{b=1}^{n} w_{kab}p_{kab}$, *i.e.* the similarity at location $[i, j]^T$ is larger than the average similarity, f_{kij} will increase. Therefore in the beginning, all the pixel locations near the ground truth (similarity greater than the average similarity threshold) will have their probability p_{ij} increased, and pixels far away (similarity smaller than the average similarity threshold) will have their probability decreased. Then the average similarity $\sum_{a=1}^{m}\sum_{b=1}^{n} w_{kab}p_{kab}$ will also increase. With the increasing average similarity, fewer pixels will have their associated probability increased. Then the heatmap will become more concentrated near the ground truth as the training process goes on.

Compared to the softmax cross entropy loss for classification, this loss takes into account the spatial location of each pixel, unlike the softmax loss that treats all the negative classes equally when performing the gradient update. More importantly, in the beginning of the training process, pixels near the ground truth will have their associated probability increased which allows for exploration around the ground truth and prevents overfitting to the ground truth.

Inference. During testing, given a new image $\mathbf{x}^*$ the prediction of landmark locations is obtained by finding the mode of the probability $\mathbf{y}^* = \arg\max_{\mathbf{y}} p(\mathbf{y} \mid \mathbf{x}^*; \boldsymbol{\Theta}^*)$. Specifically, the mode of the continuous probability distribution is obtained by first convolving the softmax map with a Gaussian kernel with the bandwidth used for training to get the discrete probability map and use its local mean around the mode as initialization for gradient ascent. The gradient can be computed by Eq. (4)

$$\begin{aligned}
&\frac{\partial p(\mathbf{y} \mid \mathbf{x}^*; \boldsymbol{\Theta}^*)}{\partial \mathbf{y}} \\
=&\frac{\partial \sum_{i=1}^{m}\sum_{j=1}^{n} K_{\boldsymbol{\Sigma}}(\mathbf{y} - \boldsymbol{\mu}_{ij})\boldsymbol{\pi}_{ij}}{\partial \mathbf{y}} \\
=&-\sum_{i=1}^{m}\sum_{j=1}^{n} K_{\boldsymbol{\Sigma}}(\mathbf{y} - \boldsymbol{\mu}_{ij})\boldsymbol{\pi}_{ij}\Sigma^{-1}(\mathbf{y} - \boldsymbol{\mu}_{ij})
\end{aligned} \tag{4}$$

Covariance of prediction. The proposed target distribution in Eq.(1) is composed of a mixture of Gaussian distributions, thus its covariance matrix can be computed as

$$\begin{aligned}
&\mathrm{Cov}[\mathbf{y} \mid \mathbf{x}; \boldsymbol{\Theta}] \\
=&\Sigma + \sum_{i=1}^{m}\sum_{j=1}^{n}(\mathbf{y}_m - \mathbf{y}_{ij})(\mathbf{y}_m - \mathbf{y}_{ij})^T\boldsymbol{\pi}_{ij}(\mathbf{x}; \boldsymbol{\Theta})
\end{aligned} \tag{5}$$

where $y_m = \sum_{i=1}^{m}\sum_{j=1}^{n} \mathbf{y}_{ij}\boldsymbol{\pi}_{ij}(\mathbf{x}; \boldsymbol{\Theta})$.

The uncertainty of the prediction is quantified by the square root of the determinant of the covariance matrix $|\,\mathrm{Cov}[\mathbf{y} \mid \mathbf{x}; \boldsymbol{\Theta}]|^{\frac{1}{2}}$.

3.2. Cascade Probability Propagation

To take advantage of the probability map, we extend the single stage model to multi-stage so that the probability map from the previous stage can be fed into the next stage to progressively improve the landmark estimation accuracy.

Similar to [30], we want to propagate the estimated probability map to the next stage. For each stage, we will have an estimated probability map $p(\mathbf{y} \mid \mathbf{x}; \boldsymbol{\Theta})$, the raw logits map (before softmax) is concatenated with the down-sampled input image and the feature map with the same size as input to the next stage, as shown in Fig. 1.

The idea is that by propagating the probability map to the next stage, it will guide the network in the next stage implicitly to focus more on the regions in the image with high probability. For example, if the prediction for a certain landmark has high uncertainty, the probability map will has a flat shape, thus encouraging the network at next stage to search in a wider region according to the probability map; if otherwise, the probability map will have a sharp shape and the network at next stage will try to refine the prediction in the nearby neighborhood.

According to our experiments, the probability map in the first stage is more spread-out, *i.e*. the prediction is more uncertain, compared to the predictions in later stages. And the prediction error is larger in the first stage than in later stages. This proves the effectiveness of the Kernel Density Network to model output probability distribution as well as the effectiveness of the cascade framework in improving detection accuracy.

Recently, [22] proposed to use the softlabel loss in multiple stages and reducing the fixed variance of the ground truth heatmap stage by stage for a more fine-grained supervision in later stages. One defect of this method is that since it is sensitive to the fixed variance of the ground truth heatmap, it requires careful tuning of this hyperparameter which can be time-consuming for deep neural networks while our method does not require such tuning but automatically learns a more concentrated probability map stage by stage.

4. Experiments

Datasets. We evaluate our methods on 300W [36], Menpo [51], COFW [4], AFLW [24].
300W has 68 landmark annotation. We first train the method on 300W-LP [55] dataset (61225 faces) which is augmented from original 300W dataset for large yaw pose. And then we fine tune on the original trainset (3837 faces). Testing is performed on 300W testset which contains 600 images.
Menpo contains images from AFLW and FDDB with landmark re-annotation following the 68 landmark annotation scheme. It has two subsets, frontal which has 68 landmark annotation for near frontal faces (6679 samples) and profile which has 39 landmark annotation for profile faces (2300 samples). We use the frontal set for cross dataset evaluation.
COFW has 1345 training samples and 507 testing samples, whose facial images are all partially occluded. The original dataset is annotated with 29 landmarks. We also use the COFW-68 test set [14] which has 68 landmarks re-annotation for cross dataset evaluation.
AFLW contains 24386 faces with large head pose up to $120°$ for yaw and $90°$ for pitch and roll. We follow [53] to conduct our experiments on AFLW-full dataset with 19 landmarks annotation where 20000 and 4386 samples are used for training and testing respectively.

Evaluation metrics. We evaluate our algorithm using standard normalised mean error (NME) and Cumulative Errors Distribution (CED) curve. In addition, the area-under-the-curve (AUC), the failure rate (FR) for a maximum error of 0.07 and the negative log likelihood (NLL) at the ground truth location are reported.
Normalized Mean Error (NME) Same as in [3], the NME is defined as the average point-to-point Euclidean distance between the ground truth ($\mathbf{y}_{gt}$) and predicted ($\mathbf{y}_{pred}$) landmark locations normalized by the ground truth facial bounding box size $d = \sqrt{w_{bbox} * h_{bbox}}$ where w_{bbox} and h_{bbox} are the width and height of the bounding box. NME $= \frac{1}{N}\sum_{k=1}^{N} \frac{||\mathbf{y}_{pred}^{(k)} - \mathbf{y}_{gt}^{(k)}||_2}{d}$.
Area under the Curve (AUC) Based on the NME in the test dataset, we can draw a Cumulative Error Distribution (CED) Curve with NME as the horizontal axis and percentage of test images as the vertical axis. Then the AUC is computed as the area under that curve for each test dataset.

Implementation details. To make a fair comparison with the SoA method using softlabel loss [3], we use the same training and testing procedure.
Training procedure: The initial learning rate is 10^{-4} for 15 epochs using a minibatch of 10, then dropped to 10^{-5} and 10^{-6} after every 15 epochs and keep training until convergence. Adam optimizer is used. We apply random augmentations such as random cropping, rotation, flipping, scale noise, color jittering, occlusion, *etc*.
Testing procedure: We follow the standard testing procedure. The face is cropped using the ground truth bounding box defined in 300W by the extreme locations of the 68 ground truth landmark points. The cropped face is rescaled to 256×256 before passed to the network. We did not use any other transformation/normalization of the face for fair comparisons.
Overall complexity: The total number of parameters is $23,820,176 \approx 24M$ in the network with 4 HourGlass modules. With 1 Nvidia RTX 2080 Ti GPU, 1 Xeon CPU, TensorFlow 1.14.0, it takes about 26min to train 1 epoch on 300W-LP dataset and 1.5min on 300W-train dataset. The inference speed is around 10 fps. Our inference is based on the mode of the predicted continuous distribution, obtained by gradient ascent (details in Section 3.1).

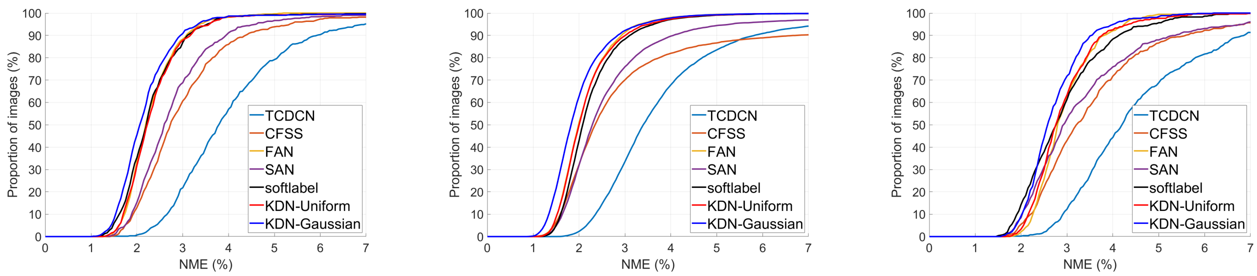
4.1. Comparison with existing approaches

We perform test on 300W test dataset. The result of softlabel, KDN-Uniform and KDN-Gaussian are implemented by ourselves using the same structure but different loss functions. To make a fair comparison, they are trained using the same random seed. Result of the softlabel based on our implementation is slightly worse than [3]. The results are shown in Table 1. The CED curves for the 300W test dataset and Menpo Challenge dataset are shown in Fig. 2a and 2b respectively.

We could see from Table 1 that compared to the softlabel loss, our loss function achieves comparable or better performance in terms of NME, AUC and FR and compared to the pseudo NLL computed from softlabel method by normalizing the final heatmap as the probability map, our method gives significantly better NLL. We also compare the results of using a uniform kernel instead of a Gaussian. Using a uniform kernel is equivalent to treating the problem as classi-

Table 1: Prediction results on 300W-test, Menpo-frontal and COFW-68 test (%)

Dataset	300W-test				Menpo-frontal				COFW-68 test			
Metric / Method	NME	AUC	FR	NLL	NME	AUC	FR	NLL	NME	AUC	FR	NLL
TCDCN [52]	4.15	42.1	4.83	-	4.04	46.2	5.84	-	4.71	35.8	8.68	-
CFSS [53]	3.09	56.7	1.83	-	3.91	57.4	9.75	-	3.79	49.0	4.34	-
FAN [3]	2.32	66.5	**0.00**	-	2.16	69.0	0.21	-	2.95	57.5	**0.00**	-
SAN [11]	2.86	59.7	1.00	-	2.95	61.9	3.11	-	3.50	51.9	3.94	-
softlabel	2.32	66.6	**0.33**	4.67	2.27	67.4	0.24	4.53	2.92	57.9	**0.00**	5.27
KDN-Uniform	2.38	65.9	0.50	**2.78**	2.19	68.7	**0.19**	2.92	2.92	58.0	0.20	4.13
KDN-Gaussian (proposed)	**2.21**	**68.3**	0.50	2.93	**2.01**	**71.1**	**0.19**	**2.87**	**2.73**	**60.1**	**0.00**	**3.21**



(a) 300W testset (b) Menpo-frontal (c) COFW-68 testset

Figure 2: CED curves of different methods on 300W-test, Menpo-frontal, COFW-68 test

fication and the target distribution as categorical with the categories representing the different pixel locations. This will result in a very sharp probability map, *i.e*. an over-confident prediction. And the loss function does not take into account the spatial relationship between different categories.

Fig. 2a shows that our method performs better in some challenging images than the softlabel method. And using a uniform kernel generates slightly larger error compared to using a Gaussian kernel. This is partly because using a uniform kernel introduces quantization error during both training and testing. The quantization error also exists in most recent work adopting the heatmap regression framework, which obtains the landmark coordinate prediction by taking the coordinates of the max value from the output heatmap. But since the heatmap is 4 times smaller than both the width and height of the original input image. This will lead to the downsampling error which makes it difficult to distinguish between the locations of two very close but different landmarks. This can be a big problem for dense landmark schemes. Previous works usually either do not address this issue or address this issue by a heuristic post-processing method such as the implementation provided in [3]. Different from these works, our method constructs a continous mixture of 2D Gaussian distribution from the predicted heatmap. Therefore during testing, we are able to find the mode of the continous distribution even if it lies between two pixels.

4.1.1 Cross-dataset Evaluation

Besides 300W testset, we evaluate the proposed method on Menpo dataset, COFW-68 testset for cross dataset evaluation. The results are shown in Table 1. The method is trained on 300W-LP and fine-tuned on 300W Challenge train set for 68 landmarks. Though the proposed method has similar or marginal improvement on 300W testset and Menpo-frontal dataset, we can see that for cross dataset evaluation on more challenging dataset such as COFW with heavy occlusion, the proposed method shows better performance, especially in terms of NLL.

4.1.2 Probability map visualization

Fig. 3 demonstrates that the proposed method can distinguish between occluded uncertain landmarks and non-occluded landmarks based on predicted heatmap. For occluded landmarks, the predicted heatmap usually has a flatter shape than the non-occluded ones. While the traditional softlabel regression methods can hardly demonstrate the predictive uncertainty in occluded landmarks. Kernel Density Network with a uniform kernel is also able to distinguish occluded landmarks, but it has a sharper shape compared to Gaussian kernel. Similar as in Kernel Density Estimation, Gaussian kernel to some extent smooths the estimated distribution compared to a uniform kernel.

Therefore, our predicted heatmap may be used to detect occlusion without occlusion annotation as supervision, unlike work in [48, 47].

Fig. 4 demonstrates that the proposed method can capture distribution with a more flexible shape. For landmarks lie on the facial contour, the predicted heatmap usually has a shape along the local edge of the face. While the traditional softlabel regression method still predicts a circular shape that represents a standard 2D Gaussian.

Figure 3: Sample heatmaps generated from two methods for **occluded** landmarks (best viewed in color and magnified). The 1st row is the proposed kernel density method, the 2nd row is the softlabel method. The displayed landmarks are subsets of the 68 points, *i.e*. the first 3 columns show point 1,5,9,13,17; the last column show point 31,46,37,49,55.

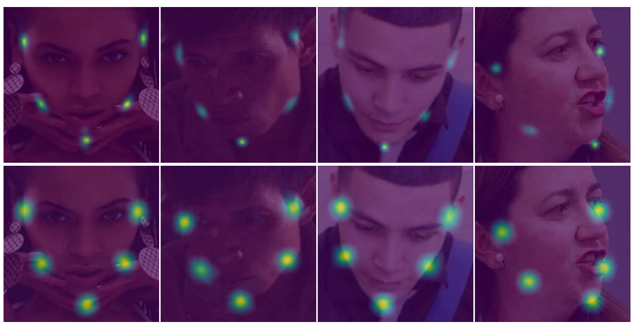


Figure 4: Sample heatmaps generated from two methods with **flexible distribution shape** (best viewed in color and magnified). The 1st row is the proposed kernel density method, the 2nd row is the softlabel method. The displayed landmarks are subsets of the 68 points, *i.e*. point 1,5,9,13,17.

4.1.3 Occlusion Dataset

We evaluate the occlusion detection quantitatively on COFW and AFLW-full. For COFW, we report the results on original testset with 29 points annotation and on COFW-68 testset [14] with model trained on 300W train set. Note that for occlusion detection, we are only using the square root of the determinant of covariance computed from the probability map but not any occlusion annotation from dataset or from manual augmentation during training. To compute pseudo variance for softlabel method, we first normalize the heatmap to make the non-negative values sum to one, then treat the normalized heatmap as a probability map to compute variance. KDN-Uniform and KDN-Gaussian generally achieve better precision/recall than softlabel. Since there are other causes of uncertainty besides occlusion, occluded landmarks should have higher uncertainty but not vice versa.

Table 2: Occlusion dataset prediction result (%)

Dataset	COFW-29				AFLW-full			
Metric / Method	NME	AUC	FR	NLL	NME	AUC	FR	NLL
SAN [11]	-	-	-	-	4.04	54.0	11.88	-
Wing [12]	-	-	-	-	3.56	53.5	7.52	-
Softlabel	2.51	64.3	0.97	6.13	2.87	59.3	4.99	5.75
KDN-Uniform	2.52	64.4	**0.79**	4.32	2.91	58.1	5.24	4.21
KDN-Gaussian	**2.28**	**67.8**	**0.79**	**3.19**	**2.80**	**60.3**	**4.67**	**3.56**

Table 3: Occlusion detection result (precision/recall %)

Method	COFW-68	COFW-29	AFLW-full
softlabel	56/40	61/40	61/40
KDN-Uniform	**70/40**	**76/40**	72/40
KDN-Gaussian	**70/40**	75/40	**73/40**

Table 4: Occluded *vs*. non-occluded points performance

Dataset	COFW-68 testset			
Occlusion state	non-occluded		occluded	
Method	NME (%)	uncertainty	NME (%)	uncertainty
softlabel	**2.30**	5.99	5.01	7.32
KDN-Uniform	2.46	1.25	4.45	7.89
KDN-Gaussian	2.34	1.63	**4.03**	11.62

4.1.4 Challenging conditions

We evaluate different methods on challenging conditions caused by either low resolution or high noise. We manually add different scales of noise to clean 300W testset and plot the prediction error in NME in Fig. 5a, where we can see that for each method, the prediction error generally increases with noise scale but the proposed method performs best under noisy conditions. In Fig. 5b we show the NME versus the resolution of the input image in pixels.

4.2. Ablation Study

If not specified, ablation study is performed on 300W test set with models trained on 300W-LP and fine-tuned on 300W trainset.

4.2.1 Kernel Density Network

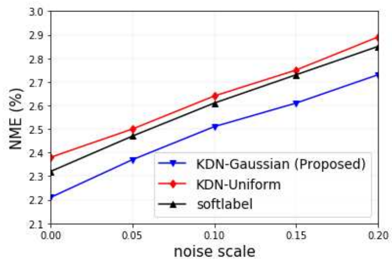
To analyze the effect of the proposed Kernel Density Network, we evaluate the performance of a single stage network in terms of prediction accuracy and uncertainty quantification. Table 5 shows the comparison of results generated from different loss function with a single stage. The proposed loss function is better than the result from softlabel loss.

Table 5: Single stage's prediction accuracy on 300W testset

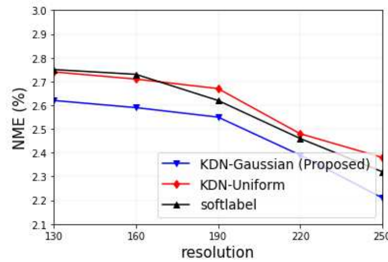
Method	NME	AUC	FR	NLL
Softlabel	2.58	62.5	1.00	4.79
KDN-Uniform	2.57	63.1	1.00	**2.95**
KDN-Gaussian	**2.52**	**63.9**	**0.50**	3.01

4.2.2 Multi-stage Cascade

The multi-stage cascade network is trained end-to-end. To analyze the effect of multiple stages, we evaluate the performance of each stage. The NME and average uncertainty at each stage is shown in Fig. 6. From the table we can see that the next stage refines the previous stage's prediction



(a) NME under different scales of noise.



(b) NME under different resolutions.

Figure 5: Sensitivity (NME) under different challenging conditions.

progressively. After each stage, the prediction error reduces and the predicted uncertainty also reduces.

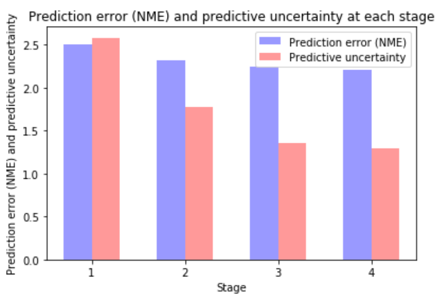


Figure 6: Uncertainty and prediction error at each stage.

4.3. Extension to Other Tasks

Theoretically the proposed method can be widely applied to any regression tasks whose target values are bounded. To demonstrate the generalizability to other tasks, we evaluate the methods on facial action unit intensity estimation.

4.3.1 Facial action unit intensity estimation

We use BP4D dataset and use the metric mean absolute error (MAE) and intra-class correlation (ICC). We divide the dataset into training and testing by different subjects, *i.e*. training set consists of subjects with odd index and testing set consists of subjects with even index. Results are shown in Table 6. The performance of KDN-Gaussian is not always the best in terms of accuracy, but it gives consistent improvement over KDN-Uniform.

Table 6: Action unit intensity estimation on BP4D dataset

Method	MAE	ICC
Deterministic	0.847	0.628
Gaussian	**0.748**	**0.664**
KDN-Uniform	0.795	0.559
KDN-Gaussian	0.757	0.588

5. Conclusion

This paper introduced a Kernel Density Deep Neural Network to quantify aleatoric uncertainty in face alignment, and for a more general distribution thus our method is applicable to other regression tasks. Since previous works using fixed variance Gaussian blob heatmap for supervision (softlabel) such as [3] do not quantify different uncertainties of different landmarks, which makes it difficult to apply to real-world problems and tasks that depend on face alignment. To our best knowledge, this is the first work to explicitly address the uncertainty quantification in fully-convolutional neural network based regression problems with a more flexible distribution than Gaussian. We show that uncertainty can be used to detect occlusion without occlusion supervision. Besides, our model provides a principled way of inference using the mode of the predicted continuous distribution to reduce quantization error compared to previous post-processing method such as interpolation [11] or heuristic method [3]. Moreover, in a multi-stage framework, the average predicted uncertainty is reduced stage by stage automatically without manually tuning the variance of the Gaussian blob heatmap in each stage.

We hope this work can benefit the landmark localization community as well as other deep ordinary regression tasks and provide a different perspective in designing the loss function to consider label distribution and aleatoric uncertainty.

Acknowledgment This work is partially supported by Cognitive Immersive Systems Laboratory (CISL), a collaboration between IBM and RPI, and also a center in IBM's AI Horizon Network.

References

[1] Christopher M. Bishop. Mixture density networks. Technical report, 1994.

[2] Adrian Bulat and Georgios Tzimiropoulos. Human pose estimation via convolutional part heatmap regression. In *ECCV*, 2016.

[3] 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 *International Conference on Computer Vision*, 2017.

[4] Xavier P. Burgos-Artizzu, Pietro Perona, and Piotr Dollár. Robust face landmark estimation under occlusion. In *Proceedings of the 2013 IEEE International Conference on Computer Vision*, ICCV '13, pages 1513–1520, Washington, DC, USA, 2013. IEEE Computer Society.

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

[6] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1302–1310, 2017.

[7] Dong Chen, Shaoqing Ren, Yichen Wei, Xudong Cao, and Jian Sun. Joint cascade face detection and alignment. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 109–122, Cham, 2014. Springer International Publishing.

[8] Yu Chen, Chunhua Shen, Xiu-Shen Wei, Lingqiao Liu, and Jian Yang. Adversarial posenet: A structure-aware convolutional network for human pose estimation. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1221–1230, 2017.

[9] Xiao Chu, Wanli Ouyang, hongsheng Li, and Xiaogang Wang. Crf-cnn: Modeling structured information in human pose estimation. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 316–324. Curran Associates, Inc., 2016.

[10] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. In Hans Burkhardt and Bernd Neumann, editors, *Computer Vision — ECCV'98*, pages 484–498, Berlin, Heidelberg, 1998. Springer Berlin Heidelberg.

[11] Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi Yang. Style aggregated network for facial landmark detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 379–388, 2018.

[12] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[13] Bin-Bin Gao, Hong-Yu Zhou, Jianxin Wu, and Xin Geng. Age estimation using expectation of label distribution learning. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 712–718. International Joint Conferences on Artificial Intelligence Organization, 7 2018.

[14] Golnaz Ghiasi and Charless C. Fowlkes. Occlusion coherence: Detecting and localizing occluded faces. *CoRR*, abs/1506.08347, 2015.

[15] Nitesh B. Gundavarapu, Divyansh Srivastava, Rahul Mitra, Abhishek Sharma, and Arjun Jain. Structured aleatoric uncertainty in human pose estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

[16] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2017.

[17] Eldar Insafutdinov, Leonid Pishchulin, Bjoern Andres, Mykhaylo Andriluka, and Bernt Schiele. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In *ECCV*, 2016.

[18] F. Kahraman, G. Muhitin, S. Darkner, and R. Larsen. An active illumination and appearance model for face alignment. *Turkish Journal of Electrical Engineering and Computer Science*, 18(4):677–692, 2010.

[19] Alex Kendall and Yarin Gal. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? (Nips), 2017.

[20] Neeraj Kumar, Peter N. Belhumeur, and Shree K. Nayar. Facetracer: A search engine for large collections of images with faces. In *The 10th European Conference on Computer Vision (ECCV)*, October 2008.

[21] Quoc Le, Alex Smola, and Stéphane Canu. Heteroscedastic Gaussian Process Regression. *Proceedings of the 22nd international conference on Machine learning ICML 05*, 227:489–496, 2005.

[22] 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. *CoRR*, abs/1901.00148, 2019.

[23] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 3431–3440, 2015.

[24] Peter M. Roth Martin Koestinger, Paul Wohlhart and Horst Bischof. Annotated Facial Landmarks in the Wild: A Large-scale, Real-world Database for Facial Landmark Localization. In *Proc. First IEEE International Workshop on Benchmarking Facial Image Analysis Technologies*, 2011.

[25] Iain Matthews and Simon Baker. Active appearance models revisited. *International Journal of Computer Vision*, 60(2):135–164, Nov 2004.

[26] Stephen Milborrow and Fred Nicolls. Locating facial features with an extended active shape model. In *Proceedings of the 10th European Conference on Computer Vision: Part IV*, ECCV '08, pages 504–513, Berlin, Heidelberg, 2008. Springer-Verlag.

[27] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII*, pages 483–499, 2016.

[28] Nathan H. Ng, Rodney A. Gabriel, Julian McAuley, Charles Elkan, and Zachary C. Lipton. Predicting surgery duration with neural heteroscedastic regression. In *MLHC*, volume 68 of *Proceedings of Machine Learning Research*, pages 100–111. PMLR, 2017.

[29] D.A. Nix and A.S. Weigend. Estimating the mean and variance of the target probability distribution. In *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94)*, pages 55–60 vol.1, 1994.

[30] Onur Ozdemir, Benjamin Woodward, and Andrew A. Berlin. Propagating uncertainty in multi-stage bayesian convolutional neural networks with application to pulmonary nodule detection. *CoRR*, abs/1712.00497, 2017.

[31] Hongyu Pan, Hu Han, Shiguang Shan, and Xilin Chen. Mean-Variance Loss for Deep Age Estimation from a Face. *CVPR 2018*, pages 5285–5294.

[32] George Papandreou, Tyler Zhu, Nori Kanazawa, Alexander Toshev, Jonathan Tompson, Christoph Bregler, and Kevin P. Murphy. Towards accurate multi-person pose estimation in the wild. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3711–3719, 2017.

[33] Leonid Pishchulin, Eldar Insafutdinov, Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, Peter V. Gehler, and Bernt Schiele. Deepcut: Joint subset partition and labeling for multi person pose estimation. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4929–4937, 2016.

[34] Sergey Prokudin, Peter V. Gehler, and Sebastian Nowozin. Deep directional statistics: Pose estimation with uncertainty quantification. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part IX*, pages 542–559, 2018.

[35] Nataniel Ruiz, Eunji Chong, and James M. Rehg. Fine-grained head pose estimation without keypoints. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.

[36] Christos Sagonas, Epameinondas Antonakos, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge. *Image Vision Comput.*, 47(C):3–18, Mar. 2016.

[37] J. Saragih and R. Goecke. A nonlinear discriminative approach to aam fitting. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8, 2007. Exported from https://app.dimensions.ai on 2018/11/15.

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

[39] Jeffrey S. Simonoff. *Smoothing Methods in Statistics*. Springer, 1996.

[40] Xiao Sun, Bin Xiao, Shuang Liang, and Yichen Wei. Integral human pose regression. *arXiv preprint arXiv:1711.08229*, 2017.

[41] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deep convolutional network cascade for facial point detection. In *Computer Vision - CVPR IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE Computer Society Conference on Computer Vision and Pattern Recognition. . 10.1109/CVPR.2013.446., Proceedings*, pages 3476–3483, 2013.

[42] Jonathan Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'14, pages 1799–1807, Cambridge, MA, USA, 2014. MIT Press.

[43] George Trigeorgis, Patrick Snape, Mihalis A. Nicolaou, Epameinondas Antonakos, and Stefanos Zafeiriou. Mnemonic descent method: A recurrent process applied for end-to-end face alignment. In *CVPR*, pages 4177–4187. IEEE Computer Society, 2016.

[44] Xinyao Wang, Liefeng Bo, and Fuxin Li. Adaptive wing loss for robust face alignment via heatmap regression. *CoRR*, abs/1904.07399, 2019.

[45] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 4724–4732, 2016.

[46] 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.

[47] Yue Wu, Chao Gou, and Qiang Ji. Simultaneous facial landmark detection, pose and deformation estimation under facial occlusion. In *CVPR*, pages 5719–5728. IEEE Computer Society, 2017.

[48] Yue Wu and Qiang Ji. Robust facial landmark detection under significant head poses and occlusion. In *ICCV*, pages 3658–3666. IEEE Computer Society, 2015.

[49] Xuehan Xiong and Fernando De la Torre. Global supervised descent method. In *CVPR*, pages 2664–2673. IEEE Computer Society, 2015.

[50] Xuehan Xiong and Fernando De la Torre Frade. Supervised descent method and its applications to face alignment. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, May 2013.

[51] S. Zafeiriou, G. Trigeorgis, G. Chrysos, J. Deng, and J. Shen. The menpo facial landmark localisation challenge: A step towards the solution. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2116–2125, July 2017.

[52] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Facial landmark detection by deep multi-task learning. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 94–108, Cham, 2014. Springer International Publishing.

[53] Shizhan Zhu, Cheng Li, Chen Change Loy, and Xiaoou Tang. Face alignment by coarse-to-fine shape searching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4998–5006, 2015.

[54] Shizhan Zhu, Cheng Li, Chen Change Loy, and Xiaoou Tang. Unconstrained face alignment via cascaded compositional learning. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3409–3417, 2016.

[55] Xiangyu Zhu, Zhen Lei, Stan Z Li, et al. Face alignment in full pose range: A 3d total solution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.