Common issues: As per reviewers request, we compare with most recent SoA methods using official code with the same metric. Results are shown in Table 1. We extend our work (CNN part) to use 4 stacks of HourGlass modules with intermediate supervision, this gives improved performance as shown in Table 1, which clearly shows our method outperforms the best compared state-of-the-art (SoA) methods on most datasets given the similar number of parameters as [5]. And our method works particularly well on challenging datasets such as Menpo-profile, COFW-68 and 300VW-category3, 300W-train challenge set, significantly outperforming the compared SoA methods.

Table 1: Comparison with other recent SoA methods (%)

Dataset	300W-test-all			Menpo-frontal			Menpo-profile			COFW-68 test		
Metric Method	NME	AUC	FR	NME	AUC	FR	NME	AUC	FR	NME	AUC	FR
FAN (reported in [5])	-	66.9	-	-	67.5	-	-	-	-	-	-	-
SAN [1]	2.86	59.7	1.00	2.95	61.9	3.11	11.71	20.7	48.39	3.50	51.9	3.94
Our method	2.25	67.8	0.17	2.16	69.0	0.18	4.71	49.0	24.30	2.65	61.8	0.00
Dataset	300VW-category1			300VW-category2			300VW-category3					
FAN (reported in [5])	-	72.1	-	-	71.2	-	-	64.1	-			
SAN [1]	2.58	64.5	1.10	2.57	63.2	0.42	4.06	52.9	7.19			
Our method	2.08	70.9	0.29	2.07	70.1	0.04	2.54	67.4	2.01			

In Table 2, we compare with some most recent best results reported, in the 300W protocol that trains on LFPW-train, HELEN-train, AFW and tests on LFPW-test, HELEN-test, ibug and use NME normalized with inter-ocular distance as the metric.

Table 2: Comparison with SoA methods on 300W dataset using 300W protocol (NME normalized with inter-ocular distance %)

Subset Method	Com.	Chal.	Full
SA [3]	3.45	6.38	4.02
Wing [2]	**3.27**	7.18	4.04
SAN [1]	3.34	6.60	3.98
Our method	3.33	**6.29**	**3.91**

1. Reviewer 1 (R1)

1.1 *Original reported performance score.* We also listed the performance score (AUC) reported in the original paper [5] in Table 1. The discrepancy may be caused by different versions of official code.

1.2 *Original benchmark protocol.* To compare with other SoA methods, we reported performance under the original benchmark protocol (300W protocol) widely used by other works in Table 2.

1.3 *How the 3D model was acquired.* Following CE-CLM [50], the 3D model is inferred from 2D annotations using structure from motion. We also tried to use models learned from 3D datasets (e.g. BP4D and Facewarehouse) but found the annotation scheme discrepancy led to inaccurate results.

2. Reviewer 2 (R2)

2.1 *Novelty.* Although there are works on CNN-CRF, especially for image segmentation and body landmark detection, there are very few works applying CNN-CRF to facial landmark detection. Compared to body landmarks, facial landmarks have many more points and require accurate localization on the facial contour, thus existing CNN+CRN methods on body landmarks are impractical or not accurate enough to be directly applied to facial landmark detection. Theoretically, our model differs from existing CNN-CRF methods in explicitly employing a fully connected (rather a tree) CRF model and a pose-dependent instead of a fixed pairwise energy function to capture structural relationship variations caused by head pose and deformation. And we perform exact conditional learning and inference compared to widely used approximate methods like mean-field and therefore we have more accurate estimation of the full covariance matrix, which quantifies the structured aleatoric uncertainty. Both R1 and R3 acknowledge the novelties of our model.

2.2 *Citation and comparison with SoA.* We provide comparison with SoA methods in Table 1 and Table 2. For other related work, we will discuss and compare them in the revised paper.

3. Reviewer 3 (R3)

3.1 *Unclear experimental methodology.* The baseline is FAN [5] since it performs best among the compared methods. And it uses the same methodology that pretrains the model using 300W-LP. This is applied to all experiments in our original paper where we train one model and test it on all datasets. For a fair comparison, we conduct additional experiment to compare with most recent SoA methods following the same 300W protocol and metrics in Table 2.

3.2 *Missing link to similar work on Continuous CRF.* Different from the two existing continuous CRF works listed by the reviewer, our work models the unary potential weight as dependent on the input, which captures structured heteroscedastic aleatoric uncertainty. Besides, compared to Continuous Conditional Neural Fields (CCNF) applied to face landmark detection, our structured output is directly defined as the location of the facial landmark while in the CCNF it is the probability of the certain landmark being aligned at each pixel location of the image. We will cite and discuss the continuous CRF papers the reviewer mentioned in the revised paper.

3.3 *Clarification on Gaussian NLL.* Gaussian NLL refers to Gaussian negative log likelihood, computed from the mean and covariance from the softmax probability map and embedded in the unary potential. It is the loss when $C_{ij} = \mathbf{0}$.

References

[1] X. Dong et al. Style aggregated network for facial landmark detection. In *CVPR*, 2018.

[2] Z. Feng et al. Wing loss for robust facial landmark localisation with convolutional neural networks. In *CVPR*, 2018.

[3] Z. Liu et al. Semantic alignment: Finding semantically consistent ground-truth for facial landmark detection. In *CVPR*, 2019.