Appendix

The appendix is organized as follows:

- In Sec. A1, we provide additional analysis and implementation details of our method.
- In Sec. A2, we provide additional results on WFLW [84], COFW [7], and 300W [66].

A1. Additional Details

A1.1. Additional discussion on Soft-argmax

As mentioned in Sec. 4.1, our work focuses on solving the problem of mismatch with Soft-argmax. The mismatch occurs because the loss is not convex w.r.t. the heatmap's elements; as illustrated in the example of Eq. (6) and Eq. (7). Our proposed method is based on the structured learning framework of loss-augmented inference [69, 74, 77], where the loss is in the form of a log-sum-of-exponentials, which is convex w.r.t. the heatmap's elements. In other words, there will be no local minimum with respect to the heatmap, and hence, no mismatch. In Fig. A1, we visualize more examples to show the convergence efficiency comparison between STAR and ours.

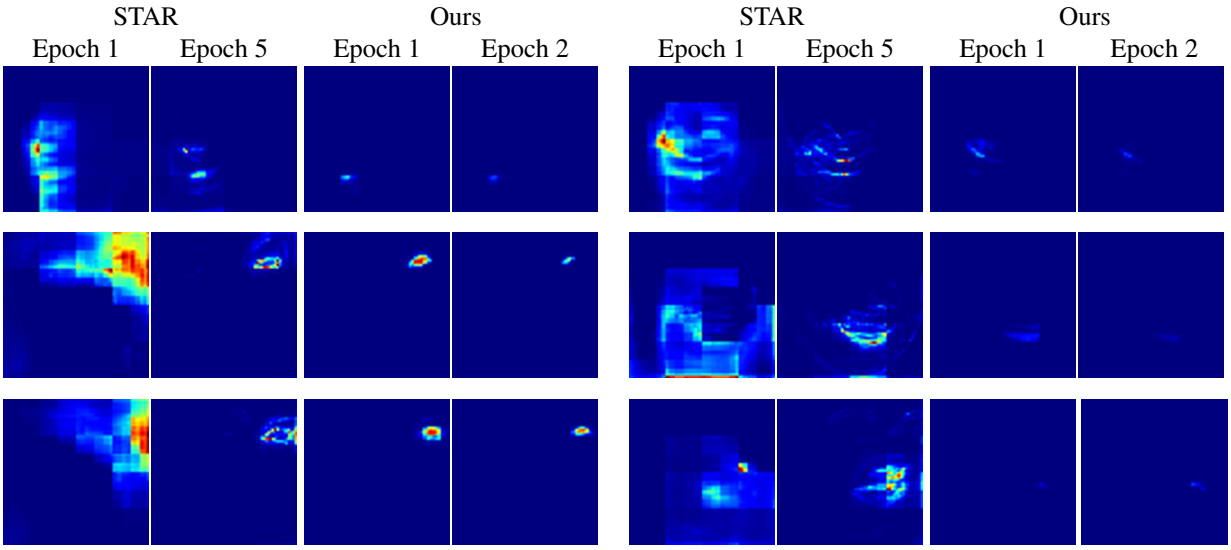


Figure A1. **More comparisons of convergence efficiency between STAR's** `Soft-argmax` **and ours** following the settings in Fig. 4.

A1.2. Additional justification for label smoothing

In Fig. 1, we demonstrate the pipeline of our image-aware label smoothing, which creates a "soft ground-truth". To verify that this softness resembles semantic ambiguity, in Fig. A2, we ask 5 users to annotate facial images following similar process of WFLW [84]. As expected, the annotations consist of variations. Importantly, we observe a strong similarity between our label smoothing and the distribution of human annotations. This further confirms that our assumption that the variation is along the image edges is correct.

A1.3. Implementation Details

Training details. The training is conducted on 4 NVIDIA A6000 GPUs with 48GB memory. The training batch size is set to 128. We used the Adam optimizer [39] with a learning rate of 0.001. The model is trained for 200 epochs and takes roughly 10 hours to finish. For the loss function, we follow Huang et al. [35], Zhou et al. [98] and incorporate AAM [35]. AMM is trained on an auxiliary task of predicting edge heatmap ($\mathcal{L}_{\texttt{Awing}}$) derived from the boundary lines. The total loss function is as follows,

$$\mathcal{L}_{\texttt{total}} = \mathcal{L}_{\texttt{Awing}} + \lambda \mathcal{L}_{\texttt{ours}}, \tag{A14}$$

where $\mathcal{L}_{\texttt{Awing}}$ is the AwingLoss [81] for training the AAM [35] and λ controls the weight of our loss $\mathcal{L}_{\texttt{ours}}$ from Eq. (13).
Label smoothing. Here we provide the implementation details of our label smoothing. Please also refer to Fig. 1. Given N landmarks $\boldsymbol{y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_N]$ of an input facial image of size 256px, we the follow hand-designed procedure [81, 84]

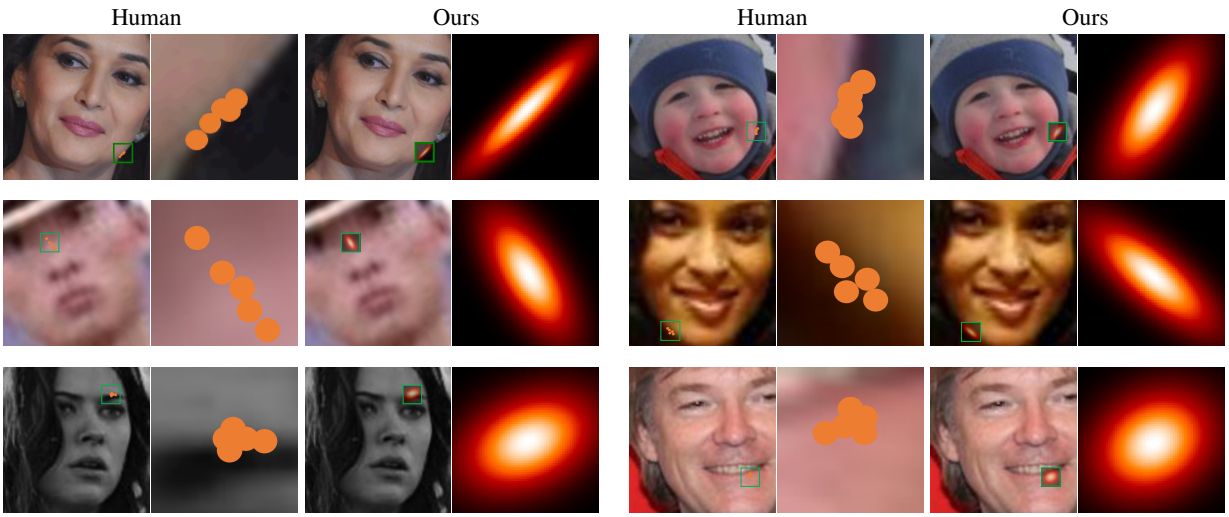
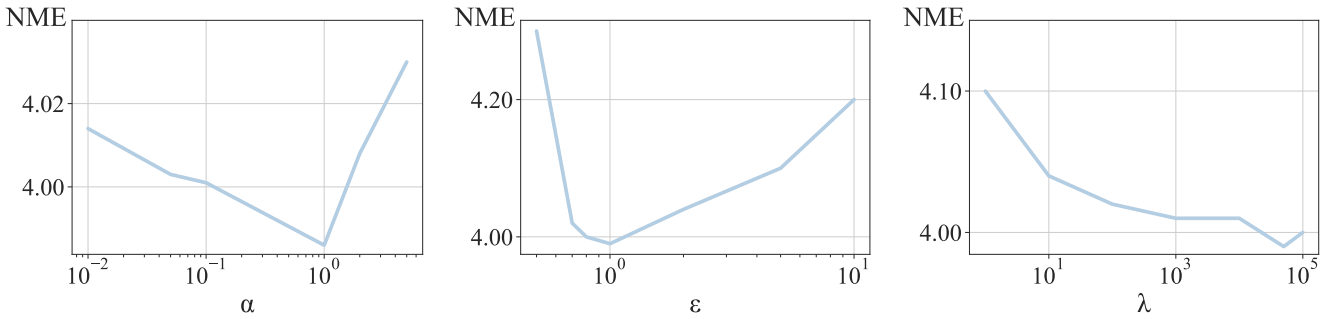


Figure A2. **Comparison between real-life annotation ambiguity and our label smoothing.** The "Human" columns are annotated by 5 users. The "Ours" columns are generated from our label smoothing.

and construct an edge heatmap $\boldsymbol{E}$ of size 64px. We then apply the `Pytorch` Gaussian blur with a kernel size of 9 and the sharpness filter with a factor of 5 onto $\boldsymbol{E}$. The refined edge heatmap is denoted as $\boldsymbol{E}'$. For each landmark $\boldsymbol{y}_n$, we consider the patch of size $2k+1$ around $\boldsymbol{y}_n$ on $\boldsymbol{E}'$, denoted as $\boldsymbol{E}'_n$. On the other hand, we construct a Gaussian heatmap $\boldsymbol{N}_n$ centered around $\boldsymbol{y}_n$ with a kernel size of 5. After both $\boldsymbol{E}'_n$ and $\boldsymbol{N}_n$ are normalized so that their maximum value is 1, we construct a joint heatmap $\boldsymbol{M}_n = 0.01\boldsymbol{E}'_n + \boldsymbol{N}_n$. Finally, we build the directional Gaussian heatmap $\boldsymbol{G}$ for image-aware label smoothing based on the covariance of $\boldsymbol{M}_n$, i.e. $\boldsymbol{G}_n = \mathcal{N}(\boldsymbol{y}_n, \gamma\Sigma_{\boldsymbol{M}_n})$. The γ is set to 0.001 for COFW but 0.01 for WFLW and 300W.

For each image, we sample 10 $\boldsymbol{y}'_n$ following the distribution $\boldsymbol{G}$ from our label smoothing. The sampling strategy is lightweight and does not have an observable impact on the overall inference speed.

Hyperparameters. We identify the following hyper-parameters which affect the performance of our approach: α (Eq. 12), ϵ (Eq. 11), and λ (Eq. A14). We experiment with different choices, *i.e.*, a grid search, on these hyperparameters to observe their impact on the performance. The results are reported in Fig. A3. It is shown that our method is sensitive to the choices of ϵ and λ. Overall, we choose $\alpha = 1, \epsilon = 1$ and $\lambda = 5000$.



(a) Different choices of α. (b) Different choices of ϵ. (c) Different choices of λ.

Figure A3. **The NME(↓) of different choices of hyper-parameters on WFLW.**

Architecture. Our architecture strictly follows the stacked hourglass network used in STAR and ADNet. Specifically, it has a parameter size of 13.48M and the FLOPs are 17.54G. The throughput is 50 images of 256 pixels by 256 pixels per second.

A2. Additional Results

In the main paper, we do not use the same evaluation settings across the three datasets presented in Tab. 1 to Tab. 4. This is because not all works report their results on all datasets (WFLW, COFW, and 300W) or all evaluation metrics (NME, FR, and AUC) and settings (inter-ocular or inter-pupil), especially the older methods pre-2020. Additionally, some works do not have open-sourced code or are not reproducible. Note, that our experiment setup follows the state-of-the-art STAR [98], i.e., we compared our method to the same baseline they do. For completeness, we report the comparison of additional evaluation metrics in the following paragraphs.

Additional ablation studies. In Tab. A1, we evaluate the robustness of our label smoothing under challenging conditions. We observe a consistent gain on most subsets.

Label smoothing	WFLW-L	WFLW-E	WFLW-I	WFLW-M	WFLW-O	WFLW-B
✗	6.68	**4.26**	3.94	3.93	4.77	**4.56**
✓	**6.58**	**4.26**	**3.90**	**3.89**	**4.74**	4.57

Table A1. **Ablation of label smoothing on the subsets of WFLW.** We ablate our label smoothing and report the inter-ocular NME↓ on the six subsets of WFLW [8].

Additional results on WFLW. In Tab. A2 and Tab. A3, we report the comparison of FR(↓) and AUC(↑) to SOTA facial landmark detection methods on the six subsets of WFLW. We achieve competitive results on all subsets. Notably, our method excels in the "largepose" subset. We provide qualitative results of the six subsets in Fig. A4, Fig. A5, Fig. A6, Fig. A7, Fig. A8, and Fig. A9.

Method	Type	WFLW-L	WFLW-E	WFLW-I	WFLW-M	WFLW-O	WFLW-B
Wing [28]	C	22.70	4.78	4.30	7.77	12.50	7.76
LAB [84]	H	28.83	6.37	6.73	7.77	13.72	10.74
DeCaFA [18]	H	21.40	3.73	3.22	6.15	9.26	6.61
AWing [81]	H	13.50	2.23	2.58	2.91	5.98	3.75
ADNet [35]	H	12.72	2.15	2.44	1.94	5.79	3.54
HIH [43]	H	12.88	**1.27**	2.43	1.45	5.16	3.10
STAR [98]	H	10.77	2.24	1.58	**0.98**	4.76	2.98
Ours	H	**9.23**	1.92	**0.86**	1.46	**3.95**	**2.33**

Table A2. **Comparison to SOTA facial landmark detection methods on the subsets of WFLW.** We report the FR↓ on the six subsets of WFLW [8], *i.e.*, large pose (WFLW-L), expression (WFLW-E), illumination (WFLWI), make-up (WFLW-M), occlusion (WFLW-O), and blur (WFLW-B). Type "C" and "H" stand for coordinate-regression and heatmap-regression methods.

Method	Type	WFLW-L	WFLW-E	WFLW-I	WFLW-M	WFLW-O	WFLW-B
Wing [28]	C	0.310	0.496	0.541	0.558	0.489	0.491
LAB [84]	H	0.235	0.495	0.543	0.539	0.449	0.463
DeCaFA [18]	H	0.292	0.546	0.579	0.575	0.485	0.494
AWing [81]	H	0.312	0.515	0.578	0.572	0.502	0.512
ADNet [35]	H	0.344	0.523	0.581	0.601	0.530	0.548
HIH [43]	H	0.358	**0.601**	**0.613**	0.618	**0.539**	**0.561**
STAR [98]	H	0.362	0.584	0.609	**0.622**	0.538	0.551
Ours	H	**0.370**	0.587	0.612	0.619	**0.539**	0.552

Table A3. **Comparison to SOTA facial landmark detection methods on the subsets of WFLW.** We report the AUC↑ on the six subsets of WFLW [8], *i.e.*, large pose (WFLW-L), expression (WFLW-E), illumination (WFLWI), make-up (WFLW-M), occlusion (WFLW-O), and blur (WFLW-B). Type "C" and "H" stand for coordinate-regression and heatmap-regression methods.

Additional results on COFW. In Tab. A4, we report more quantitative comparisons on COFW. We also provide qualitative results of COFW in Fig. A10.

Method	Inter-Pupil		Inter-Ocular	
	NME ↓	FR ↓	NME ↓	FR ↓
TCDCN [93]	8.05	-	-	-
Wu and Ji [85]	5.93	-	-	-
DAC-CSR [27]	-	-	6.03	4.73
Wing [28]	5.44	3.75	-	-
DCFE [78]	5.27	7.29	-	-
LAB [84]	-	-	3.92	0.39
SDFL [52]	-	-	3.63	**0.00**
SLPT [86]	4.79	1.18	3.32	**0.00**
Awing [81]	4.94	0.99	-	-
ADNet [35]	4.68	0.59	-	-
HIH [43]	4.63	**0.39**	-	-
STAR [98]	4.62	0.79	3.21	**0.00**
Ours	**4.58**	0.79	**3.15**	**0.00**

Table A4. **Additional comparison to SOTA on COFW.**

Method	Full	Comm.	Chal.
SDM [88]	7.50	5.57	15.40
CFSS [99]	5.76	4.73	9.98
MDM [76]	5.88	4.83	10.14
RAR [87]	4.94	4.12	8.36
DVLN [83]	4.66	3.94	7.62
HG-HSLE [100]	4.59	3.94	7.24
DCFE [78]	4.55	3.83	7.54
LAB [84]	4.12	3.42	6.98
Wing [28]	4.04	**3.27**	7.18
Awing [81]	4.31	3.77	6.52
ADNet [35]	4.08	3.51	6.47
STAR [98]	**4.03**	3.55	6.22
Ours	**4.03**	3.51	**6.09**

Table A5. **Inter-pupil NME↓ comparisons on 300W and common/challenging subsets.**

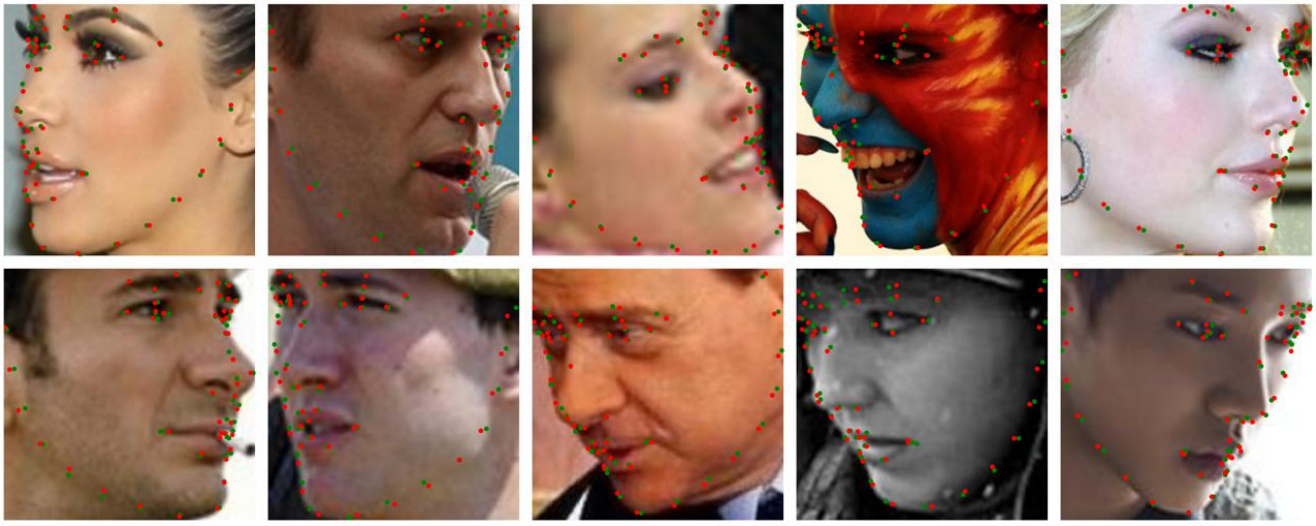


Figure A4. **Qualitative results on the "largepose" subset of WFLW.** The ground truth landmarks are marked in green while our predictions are marked in red.

Additional results on 300W. In Tab. A5, we report more quantitative comparisons on COFW. We provide qualitative results of the two subsets of 300W in Fig. A11 and Fig. A12.

Figure A5. **Qualitative results on the "expression" subset of WFLW.** The ground truth landmarks are marked in green while our predictions are marked in red.

Figure A6. **Qualitative results on the "illumination" subset of WFLW.** The ground truth landmarks are marked in green while our predictions are marked in red.

Figure A7. **Qualitative results on the "makeup" subset of WFLW.** The ground truth landmarks are marked in green while our predictions are marked in red.

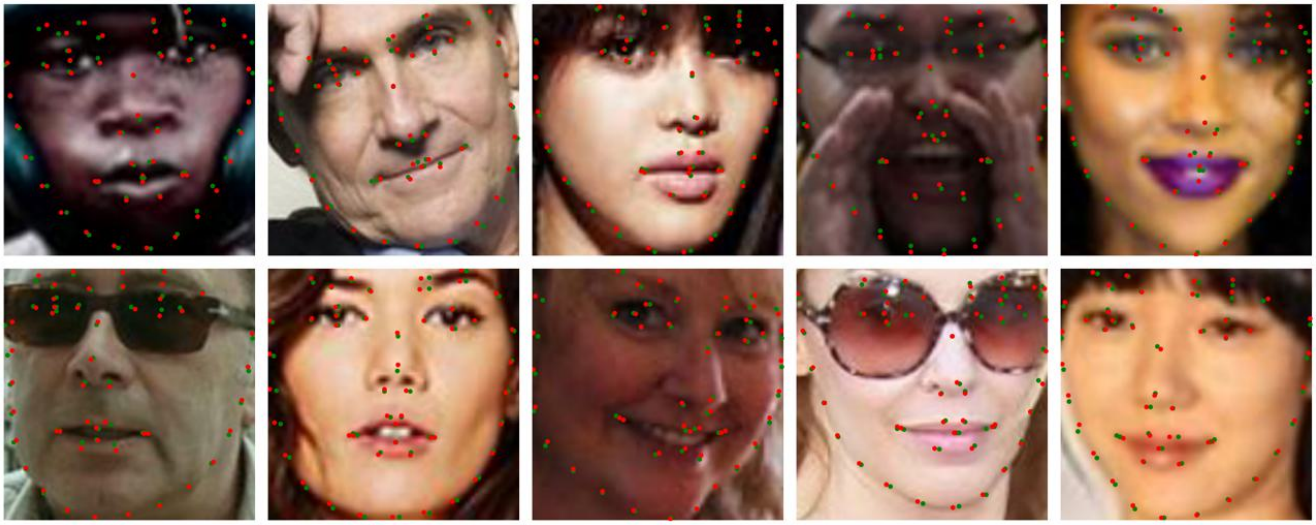


Figure A8. **Qualitative results on the "occlusion" subset of WFLW.** The ground truth landmarks are marked in green while our predictions are marked in red.

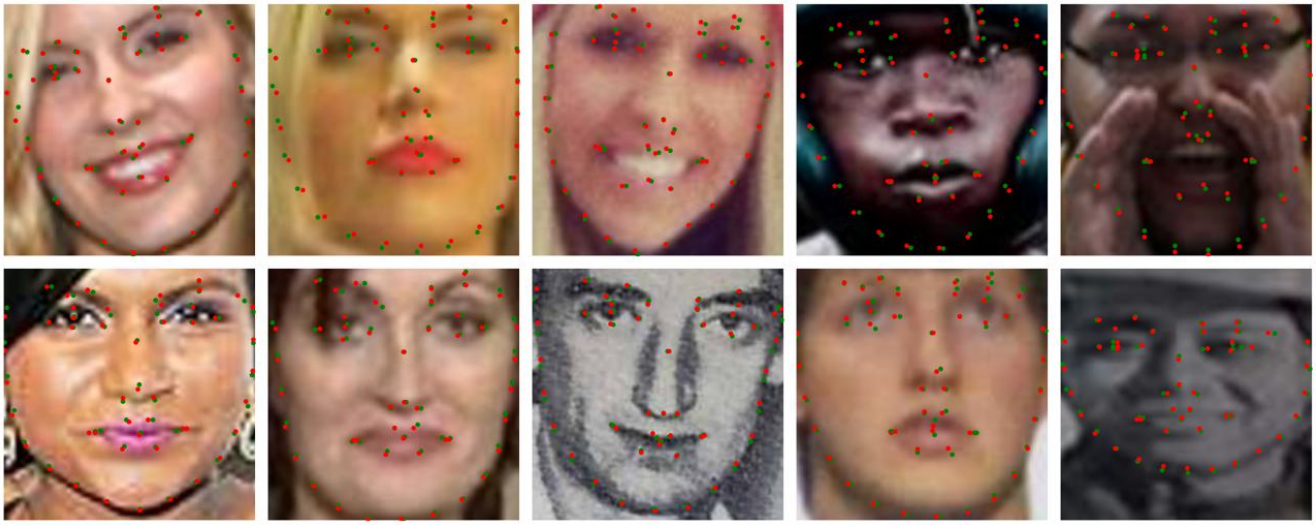


Figure A9. **Qualitative results on the "blur" subset of WFLW.** The ground truth landmarks are marked in green while our predictions are marked in red.

Figure A10. **Qualitative results on COFW.** The ground truth landmarks are marked in green while our predictions are marked in red.

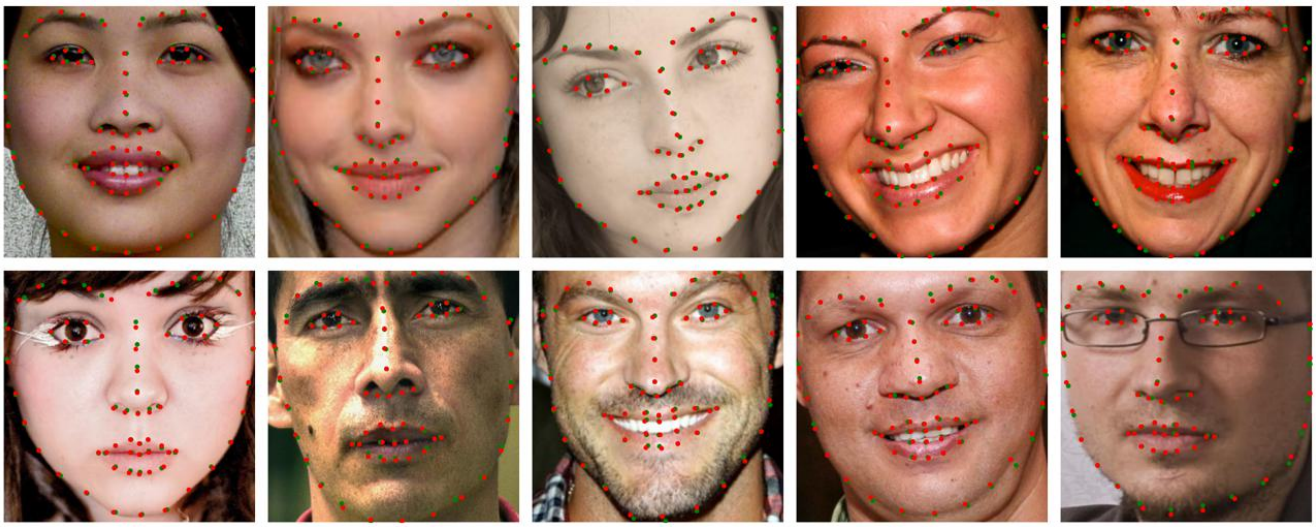


Figure A11. **Qualitative results on the "common" subset of COFW.** The ground truth landmarks are marked in green while our predictions are marked in red.

Figure A12. **Qualitative results on the "challenge" subset of 300W.** The ground truth landmarks are marked in green while our predictions are marked in red.