Appendix

This appendix provides additional discussion (Sec. A), full experiments (Sec. B) and PCA qualitative results (Sec. C).

A. Discussion

A.1. Discussion on the Motivation of STAR

ADNet [21] uses a hand-crafted constraint weight, which is a constant so that all landmarks have the same degree of semantic ambiguity. We use λ_1/λ_2 (the anisotropy of distribution) to assess the ambiguity, where a higher value indicates that ambiguity is more severe. As shown in Table. 7, the ambiguities usually change with the samples and the landmarks. As a result, using a hand-crafted setup as ADNet is not the optimal choice. Our STAR is a self-adaptive scheme that dynamically adjusts the degree of semantic ambiguity, bringing obvious improvement.

Face	λ_1/λ_2	Samples	λ_1/λ_2
eye	**1.59**	easy (NME(%) $\leq$ 3.0)	**1.95**
contour	2.60	hard (NME(%) $\geq$ 8.0)	2.20

Table 7. The ambiguity of landmarks in different facial region and samples.

A.2. Discussion on the Influence of Semantic Ambiguity on STAR

We observe that the improvement of STAR in COFW in not obvious compared with STAR in WFLW. And we infer that the unobvious is because the landmarks in COFW distribute on the five sense organs, where the semantic is relatively precise. To verify it, we split the WFLW into two subsets: a COFW-like subset and a subset containing only the face contour. The results in the Table 8 show: 1) Similar to Fig. 6, we use five models to calculate the variance score. The variance on face contour is 1.06, and the result on the other is 0.52. Since higher variance means that semantic ambiguity is more significant, it indicates that ambiguity is more server on the face contour. 2) The NME improvement of STAR on contour is 0.23, and the improvement on the other is 0.12. These results indicate that the improvement of STAR will be more obvious when the ambiguity is serious. In sum, STAR is suitable for the dataset with severe ambiguity and can also improve performance with relatively precise semantics.

WFLW	std	ADNet	STAR
COFW-subset	**0.5172**	3.39	3.27 (+0.12)
Contour	1.0660	6.07	5.84 (**+0.23**)

Table 8. The influence of semantic ambiguity on STAR. STAR has a more obvious improvement in dataset with server ambiguity.

A.3. Discussion on the Effect of STAR on Hard Samples

As shown in Table. 10 and Table. 11, there is a significant improvement in the challenge test sets. We discuss the reason from two aspects. 1) Most hard samples are in challenge test sets. Compared with easy samples, the ambiguities in hard samples are more serious, leaving more room to be improved; 2) STAR has a strong help for hard samples. As discussed in Sec. 5.3, STAR works as a label regularization, which forces the model to pay more attention to structural constraints between landmarks. And this structural information has a significant impact on detecting hard samples, resulting in a more significant improvement in the challenge test.

B. Full Experiments

In this section, we report the full experiments on COFW, 300W and WFLW, including: 1) NME, $\text{FR}_{0.1}$ and $\text{AUC}_{0.1}$ results on WLFW subsets; 2) NME and $\text{FR}_{0.1}$ results on COFW under Inter-Ocular and Inter-Pupil normalization; 3) NME results on 300W under Inter-Ocular and Inter-Pupil normalization.

B.1. Details of Comparison on COFW

The comparison results on COFW under Inter-Ocular normalization and Inter-Pupil normalization are shown in Table 9.

Method	Inter-Ocular		Inter-Pupil	
	NME(%)↓	FR(%)↓	NME(%)↓	FR(%)↓
DAC-CSR [19]	6.03	4.73	-	-
LAB [48]	3.92	0.39	-	-
Coord [44]	3.73	0.39	-	-
SDFL [29]	3.63	0.00	-	-
Heatmap [44]	3.45	0.20	-	-
Human [5]	-	-	5.60	-
TCDCN [56]	-	-	8.05	-
Wing [18]	-	-	5.44	3.75
DCFE [43]	-	-	5.27	7.29
AWing [45]	-	-	4.94	0.99
ADNet [21]	-	-	4.68	0.59
SLPT [50]	3.32	0.00	4.79	1.18
HIH [53]	3.21	0.00	4.63	0.39
STAR (Ours)	**3.21**	**0.00**	**4.62**	0.79

Table 9. NME and $\text{FR}_{0.1}$ comparisons of the STAR under Inter-Ocular normalization and Inter-Pupil normalization on COFW. The threshold for failure rate (FR) is set to 0.1. The best and second best results are marked in colors of red and blue, respectively.

B.2. Details of Comparison on 300W

The comparison results on 300W under Inter-Ocular normalization and Inter-Pupil normalization.

Inter-pupil Normalization

Method	Common Subset	Challenging Subset	Fullset
SDM [54]	5.57	15.40	7.50
CFSS [59]	4.73	9.98	5.76
MDM [41]	4.83	10.14	5.88
RAR [52]	4.12	8.35	4.94
DVLN [49]	3.94	7.62	4.66
DCFE [43]	3.83	7.54	4.55
LAB [48]	3.42	6.98	4.12
Wing [18]	3.27	7.18	4.04
AWing [45]	3.77	6.52	4.31
ADNet [21]	3.51	6.47	4.08
STAR (Ours)	**3.50**	**6.22**	**4.03**

Table 10. Comparing with state-of-the-art methods on 300W under inter-pupil normalisation.

Inter-ocular Normalisation

Method	Common Subset	Challenging Subset	Fullset
PCD-CNN [23]	3.67	7.62	4.44
CPM+SBR [14]	3.28	7.58	4.10
SAN [14]	3.34	6.60	3.98
LAB [48]	2.98	5.19	3.49
DeCaFA [11]	2.93	5.26	3.39
DU-Net [40]	2.90	5.15	3.35
LUVLi [24]	2.76	5.16	3.23
AWing [45]	2.72	4.52	3.07
ADNet [21]	2.53	4.58	2.93
PIPNet [22]	2.78	4.89	3.19
SLPT [50]	2.75	4.90	3.17
HIH [53]	2.65	4.89	3.09
DTLD [25]	2.59	4.50	2.96
STAR (Ours)	**2.52**	**4.32**	**2.87**

Table 11. Comparing with state-of-the-art methods on 300W under inter-ocular normalisation.

B.3. Details of Comparison on WFLW

The comparison results on WFLW test set and its subsets are tabulated in Table 12. STAR yields the competitive performance in NME, $\text{FR}_{0.1}$ and $\text{AUC}_{0.1}$ at SOTA level on all subsets.

C. Qualitative Results

C.1. Further Visualization of the PCA results

Additional visualization of PCA results are shown in Figure 7.

Metric	Method	Testset	Pose	Expression	Illumination	Make-up	Occlusion	Blur
NME(%)↓	ESR [6]	11.13	25.88	11.47	10.49	11.05	13.75	12.20
	SDM [54]	10.29	24.10	11.45	9.32	9.38	13.03	11.28
	CFSS [59]	9.07	21.36	10.09	8.30	8.74	11.76	9.96
	DVLN [49]	6.08	11.54	6.78	5.73	5.98	7.33	6.88
	LAB [48]	5.27	10.24	5.51	5.23	5.15	6.79	6.12
	Wing [18]	5.11	8.75	5.36	4.93	5.41	6.37	5.81
	DeCaFA [11]	4.62	8.11	4.65	4.41	4.63	5.74	5.38
	AWing [45]	4.36	7.38	4.58	4.32	4.27	5.19	4.96
	LUVLi [24]	4.37	7.56	4.77	4.30	4.33	5.29	4.94
	SDFL [29]	4.35	7.42	4.63	4.29	4.22	5.19	5.08
	SDL [27]	4.21	7.36	4.49	4.12	4.05	4.98	4.82
	HIH [53]	4.08	6.87	4.06	4.34	3.85	4.85	4.66
	ADNet [21]	4.14	6.96	4.38	4.09	4.05	5.06	4.79
	PIPNet [22]	4.31	7.51	4.44	4.19	4.02	5.36	5.02
	RePFormer [26]	4.11	7.25	4.22	4.04	3.91	5.11	4.76
	SLPT [50]	4.14	6.96	4.45	4.05	4.00	5.06	4.79
	STAR (Ours)	**4.02**	**6.76**	4.27	**3.97**	**3.83**	**4.80**	**4.58**
$FR_{0.1}$(%)↓	ESR [6]	35.24	90.18	42.04	30.80	38.84	47.28	41.40
	SDM [54]	29.40	84.36	33.44	26.22	27.67	41.85	35.32
	CFSS [59]	20.56	66.26	23.25	17.34	21.84	32.88	23.67
	DVLN [49]	10.84	46.93	11.15	7.31	11.65	16.30	13.71
	LAB [48]	7.56	28.83	6.37	6.73	7.77	13.72	10.74
	Wing [18]	6.00	22.70	4.78	4.30	7.77	12.50	7.76
	DeCaFA [11]	4.84	21.40	3.73	3.22	6.15	9.26	6.61
	AWing [45]	2.84	13.50	2.23	2.58	2.91	5.98	3.75
	LUVLi [24]	3.12	15.95	3.18	2.15	3.40	6.39	3.23
	SDFL [29]	2.72	12.88	1.59	2.58	2.43	5.71	3.62
	SDL [27]	3.04	15.95	2.86	2.72	1.45	5.29	4.01
	HIH [53]	2.60	12.88	1.27	2.43	1.45	5.16	3.10
	ADNet [21]	2.72	12.72	2.15	2.44	1.94	5.79	3.54
	SLPT [50]	2.76	12.27	2.23	1.86	3.40	5.98	3.88
	STAR (Ours)	**2.32**	**11.69**	2.24	**1.58**	**0.98**	**4.76**	3.24
$AUC_{0.1}$ ↑	ESR [6]	0.2774	0.0177	0.1981	0.2953	0.2485	0.1946	0.2204
	SDM [54]	0.3002	0.0226	0.2293	0.3237	0.3125	0.2060	0.2398
	CFSS [59]	0.3659	0.0632	0.3157	0.3854	0.3691	0.2688	0.3037
	DVLN [49]	0.4551	0.1474	0.3889	0.4743	0.4494	0.3794	0.3973
	LAB [48]	0.5323	0.2345	0.4951	0.5433	0.5394	0.4490	0.4630
	Wing [18]	0.5504	0.3100	0.4959	0.5408	0.5582	0.4885	0.4918
	DeCaFA [11]	0.5630	0.2920	0.5460	0.5790	0.5750	0.4850	0.4940
	AWing [45]	0.5719	0.3120	0.5149	0.5777	0.5715	0.5022	0.5120
	LUVLi [24]	0.557	0.310	0.549	0.584	0.588	0.505	0.525
	ADNet [21]	0.6022	0.3441	0.5234	0.5805	0.6007	0.5295	0.5480
	SDFL [29]	0.576	0.315	0.550	0.585	0.583	0.504	0.515
	SDL [27]	0.589	0.315	0.566	0.595	0.604	0.524	0.533
	HIH [53]	0.605	0.358	0.601	0.613	0.618	0.539	0.561
	SLPT [50]	0.595	0.348	0.574	0.601	0.605	0.515	0.535
	STAR (Ours)	**0.6050**	**0.3624**	**0.5839**	**0.6094**	**0.6216**	**0.5379**	**0.5514**

Table 12. Performance Comparison of the STAR and the state-of-the-art methods on WFLW and its subsets. The best and second best results are marked in colors of red and blue, respectively.

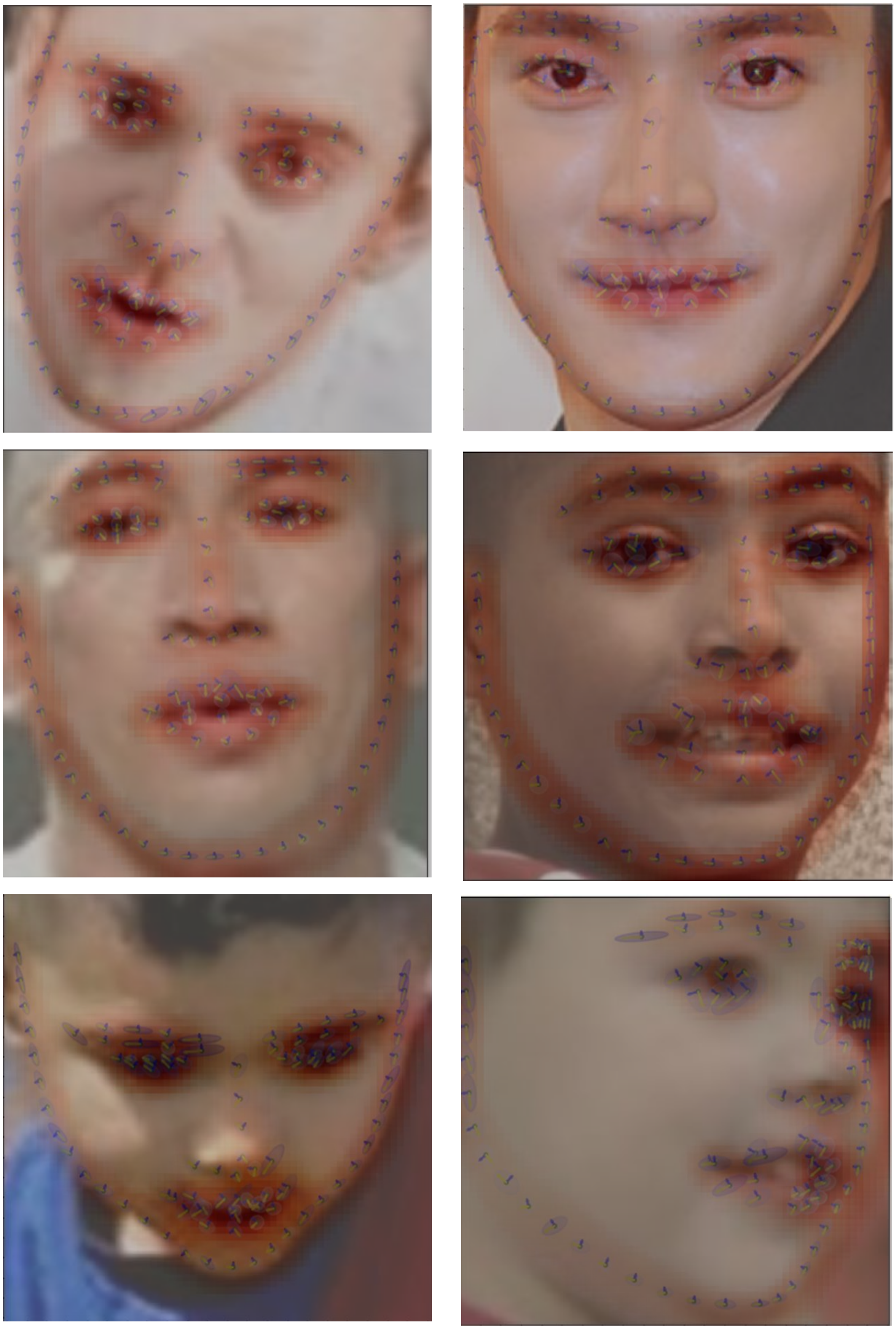

Figure 7. More qualitative results of PCA on WFLW. The yellow and blue arrows indicate the principal component estimated from heatmap via PCA. The shading of the blue ellipse represents the ambiguity strength. (Best view in color and zoom in.)