ORFormer: Occlusion-Robust Transformer for Accurate Facial Landmark Detection – Supplementary Materials

Jui-Che Chiang[1,3] Hou-Ning Hu[2] *Bo-Syuan Hou[1] *Chia-Yu Tseng[1]
Yu-Lun Liu[1] Min-Hung Chen[3] Yen-Yu Lin[1]

[1]National Yang Ming Chiao Tung University [2]MediaTek Inc. [3]NVIDIA

https://github.com/ben0919/ORFormer

We provide additional implementation details, more ablation studies, the analysis of the computational complexity, and the discussion of the limitations of ORFormer in this supplementary document.

1. Additional Implementation Details

1.1. Model Training

We employ the Adam optimizer [5] along with the cosine annealing warm restart scheduler proposed by Loshchilov *et al*. [9] in all our experiments. The number of iterations for the first restart is set to 5, and the increase factor is set to 2.

The entire training process is carried out on a single NVIDIA GTX 1080 Ti with 11GB of memory. Specifically, for the quantized heatmap generator, we set the learning rate to 0.0007 with a batch size of 128. For deriving the proposed ORFormer, we use a learning rate of 0.0001 with a batch size of 64. For the landmark detection models, we set the learning rate to 0.001 with a batch size of 16.

1.2. Heatmap Definition

ORFormer aims to identify non-visible regions and recover missing features, enabling the generation of high-fidelity heatmaps resilient to challenging scenarios like occlusions, extreme lighting conditions, or extreme head rotations. This capability assists facial landmark detection (FLD) methods in maintaining robustness in such challenging scenarios.

To support FLD methods effectively and efficiently, we employ heatmaps on facial edges (contours) as constraints by following a related approach proposed by Wu *et al*. [15]. Utilizing edge heatmaps alone can reduce computational costs while providing sufficient information for FLD methods.

Heatmap Generation. As illustrated in Fig. 1, for a given face image $I \in \mathbb{R}^{h\times w\times 3}$ and its ground-truth landmark annotations $L = \{l_i\}_{i=0}^{N_L-1}$, we divide L into N_E subsets $L_j \subset L, j = 0, ..., N_E - 1$ to represent the facial edges, such as the cheek and eyebrow. Here, N_L represents the number of landmarks per face, and N_E denotes the number of edges per face. Each facial edge L_j is utilized to interpolate the edge line, thereby forming the binary edge map B_j of the same size as the face image. Subsequently, a distance transform is applied to B_j, computing the nearest distance to the edge line for every pixel, resulting in the formation of the distance map M_j, which is also of the same size as the face image. Finally, we obtain the ground-truth edge heatmap $\hat{H}_j$ used to supervise the quantized heatmap and ORFormer by the following formula:

$$\hat{H}_j(x,y) = \begin{cases} \exp(-\frac{M_j(x,y)^2}{2\sigma^2}), & \text{if } M_j(x,y) < 3\sigma, \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

σ represents the standard deviation of the values in the distance map M_j.

Index Mapping. Our experiments are conducted on three distinct datasets: WFLW [15], COFW [1], and 300W [11]. As illustrated in Fig. 2, the number of landmarks varies across these datasets, leading to differences in the edge heatmap. Consequently, we provide the index mappings between the landmarks and the facial edges in the following.

For the **WFLW** dataset, with N_L equal to 98 and N_E equal to 15, the index mapping is given as follows:

```
Edge 0: [0-32]
Edge 1: [33-37]
Edge 2: [38-41,33]
Edge 3: [42-46]
Edge 4: [46-49,50]
Edge 5: [51-54]
Edge 6: [55-59]
Edge 7: [60-64]
Edge 8: [64-67,60]
Edge 9: [68-72]
```

*means equal contribution

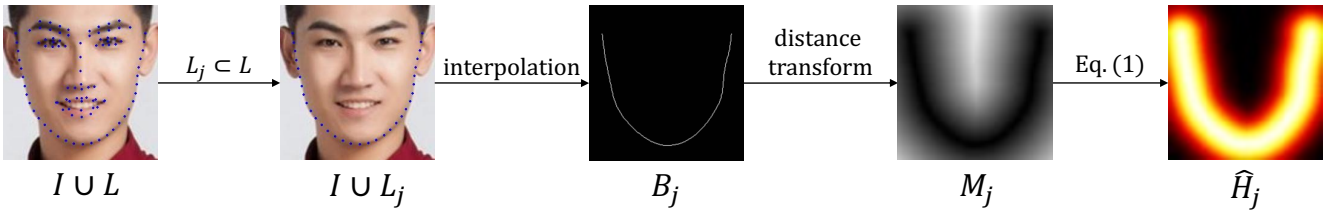

Figure 1. **Generation flow of the ground-truth edge heatmap.**

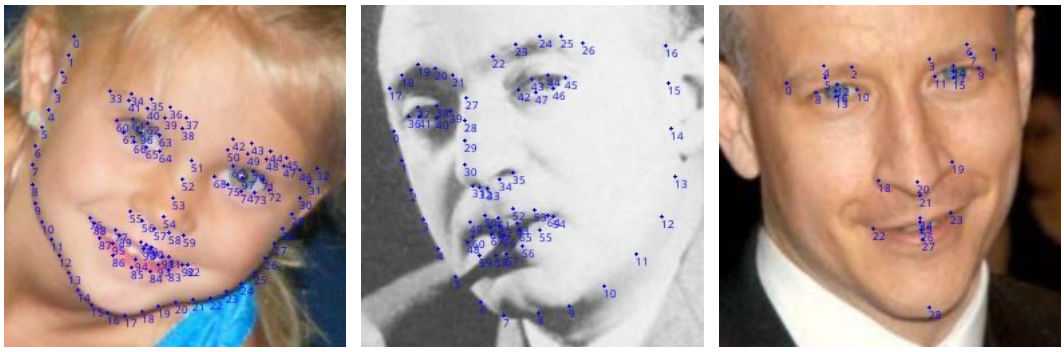

Figure 2. **Visualization of the ground-truth landmarks in different datasets.**

Method (Publication)	Backbone	WFLW-Full			COFW	300W ($NME_{io}\downarrow$)		
		$NME_{io}\downarrow$	$FR_{10\%}\downarrow$	$AUC_{10\%}\uparrow$	$NME_{ip}\downarrow$	Full	Comm.	Chal.
LAB (CVPR18) [15]	Hourglass	5.27	7.56	0.532	-	3.49	2.98	5.19
Wing (CVPR18) [2]	ResNet-50	4.99	6.00	0.550	5.44	-	-	-
HRNet (CVPR19) [12]	HRNet-W18	4.60	4.64	-	-	3.32	2.87	5.15
AWing (ICCV19) [14]	Hourglass	4.36	2.84	0.572	4.94	3.07	2.72	4.52
LUVLi (CVPR20) [6]	DU-Net	4.37	3.12	0.577	-	3.23	2.76	5.16
ADNet (ICCV21) [3]	Hourglass	4.14	2.72	0.602	4.68	2.93	2.53	4.58
PIPNet (IJCV21) [4]	ResNet-101	4.31	-	-	-	3.19	2.78	4.89
HIH (arXiv21) [7]	Hourglass	4.08	2.60	0.605	4.63	3.09	2.65	4.89
SLPT (CVPR22) [16]	HRNet-W18	4.14	2.76	0.595	4.79	3.17	2.75	4.90
RePFormer (arXiv22) [8]	ResNet-101	4.11	-	-	-	3.01	-	-
†STAR (CVPR23) [18]	Hourglass	4.03	2.32	0.611	4.62	2.90	2.52	4.46
LDEQ (CVPR23) [10]	Hourglass	3.92	2.48	0.624	-	-	-	-
ORFormer (Ours)	Hourglass	3.86	1.76	0.622	4.46	2.90	2.53	4.43

Table 1. **Quantitative comparison with state-of-the-art methods on WFLW, COFW, and 300W.** NME is reported for all datasets. For WFLW, FR and AUC with a threshold of 10% are included. The best and second best results are highlighted. The † symbol represents the results we reproduced.

```
Edge 10: [72-75,68]
Edge 11: [76-82]
Edge 12: [82-87,76]
Edge 13: [88-92]
Edge 14: [92-95,88]
```

For the **300W** dataset, with N_L equal to 68 and N_E equal to 13, the index mapping is given as follows:

```
Edge 0: [0-16]
Edge 1: [17-21]
Edge 2: [22-26]
Edge 3: [27-30]
Edge 4: [31-35]
Edge 5: [36-39]
Edge 6: [39-41,36]
Edge 7: [42-45]
Edge 8: [45-47,42]
Edge 9: [48-54]
Edge 10: [54-59,48]
Edge 11: [60-64]
```

Method	Architecture	Loss Functions		WFLW (NME↓)	
		Heatmap	Landmark	Full	Occ.
ADNet [3]	HGNet	AWing	ADL	4.14	5.06
	HGNet+**ORFormer**	AWing	ADL	**4.06** (+1.9%)	**4.94** (+2.4%)
†STAR [18]	HGNet	AWing	STAR	4.03	4.82
	HGNet+**ORFormer**	AWing	STAR	**3.92** (+2.7%)	**4.66** (+3.3%)

Table 2. **Ablation study of enabling ORFormer for landmark detection on WFLW.** NME is reported. The † symbol represents the results we reproduced. The relative performance improvement is calculated based on HGNet.

```
Edge 12: [64-67,60]
```

For the **COFW** dataset, with N_L equal to 29 and N_E equal to 14, the index mapping is given as follows:

```
Edge 0: [0,4,2]
Edge 1: [2,5,0]
Edge 2: [1,6,3]
Edge 3: [3,7,1]
Edge 4: [8,12,10]
Edge 5: [10,13,8]
Edge 6: [9,14,11]
Edge 7: [11,15,9]
Edge 8: [18,21,19]
Edge 9: [20,21]
Edge 10: [22,26,23]
Edge 11: [23,27,22]
Edge 12: [22,24,23]
Edge 13: [23,25,22]
```

2. More Experiments

2.1. Comparisons with State-of-the-Art Methods

Due to limited space in the main paper, we provide the full experimental table here, as shown in Table 1. We also provide more samples for visualization of the output landmark, as shown in Figure 3.

2.2. Ablation Study

2.2.1 Effectiveness of ORFormer

Due to limited space in the main paper, we provide more samples for visualization of the output heatmap of ORFormer, as shown in Figure 4.

2.2.2 Effectiveness of ORFormer's Heatmaps.

To demonstrate the effectiveness of ORFormer for heatmaps generation for facial landmark detection, we compare it to the methods that utilize the same baseline network: ADNet [3] and STAR [18]. The results are presented in Table 2. By incorporating ORFormer's output heatmaps as additional information to the networks, alongside the same loss functions used by ADNet and STAR, our method achieves performance gains, especially in the occlusion subset, showing the effectiveness of ORFormer's heatmap to existing FLD methods.

Method	Distance Function	L2 Loss ↓
ORFormer	concat$\{X-M, M-X\}$	24.86
ORFormer	$\lvert X-M\rvert$	24.43
ORFormer	$(X-M)^2$	**23.87**

Table 3. **Quantitative evaluation on different designs of the distance function of ORFormer's occlusion detection head.** X represents the image patch embedding and M represents the messenger embedding. The label Occ. Head denotes the proposed occlusion detection head. The proposed occlusion detection head is not enabled in the first-row entry. The occlusion-aware cross-attention component is not enabled here. Results highlighted in **bold** represent the best performance. The heatmap regression L2 loss is reported on WFLW.

Method	W's Design	L2 Loss ↓
ORFormer	5×5 conv.	24.26
ORFormer	3×3 conv.	24.05
ORFormer	Fully connected layer	**23.87**

Table 4. **Quantitative evaluation with different filter sizes of W in ORFormer's occlusion detection head**. The occlusion-aware cross-attention component is not enabled here. Results highlighted in **bold** represent the best performance. The heatmap regression L2 loss is reported on WFLW.

2.2.3 ORFormer's Component

Occlusion Detection Head. As mentioned in the paper, we incorporate an occlusion detection head in our proposed ORFormer to identify occluded patches by evaluating the dissimilarity between the image patch embedding X^{l+1} and the messenger embedding M^{l+1}. The patch-specific occlusion map $\alpha^{l+1} = \{\alpha_k^{l+1}\}_{k=0}^{m\times n-1}$ is obtained via

$$\alpha_k^{l+1} = \sigma(W^{l+1} \cdot \texttt{dist}(X_k^{l+1}, M_k^{l+1})), \tag{2}$$

where $\texttt{dist}(\cdot,\cdot)$ calculates the element-wise squared difference between the two embeddings, W^{l+1} represents a fully connected layer that transforms the embedding returned by `dist` into a scalar, and $\sigma(*)$ is the sigmoid function ensuring α_k^{l+1} ranges between 0 and 1.

To explore the difference between designs of distance functions, we compare the heatmap regression quality using L2 loss with various designs of the distance function, as shown in Table 3. We observe that employing the squared

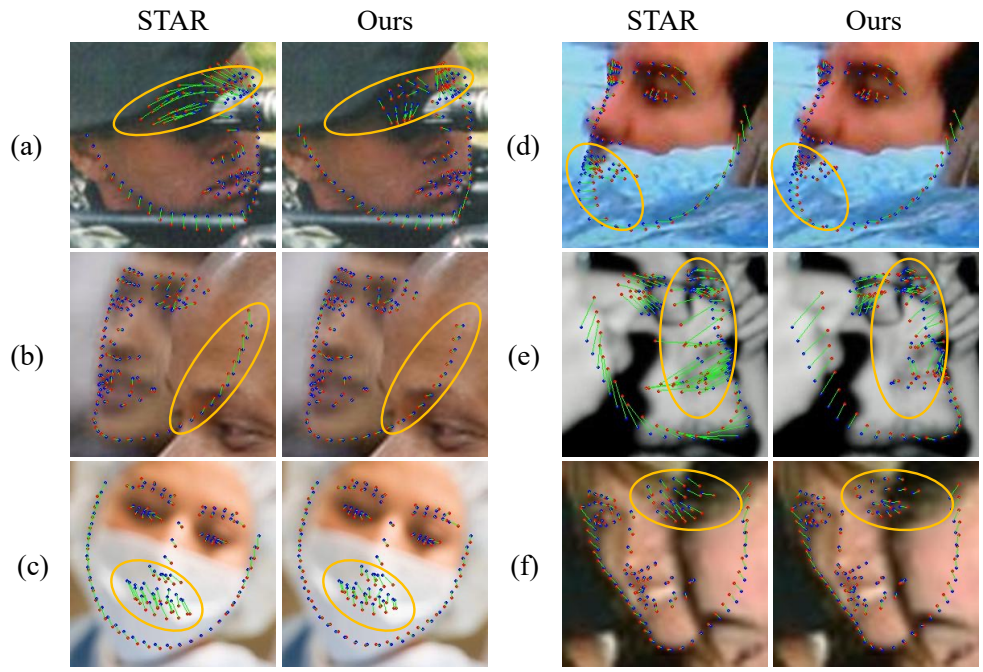

Figure 3. **Qualitative comparison with the reproduced baseline method, STAR, on extreme cases from the test set of WFLW.** The ground-truth landmarks are marked in blue, while the predicted landmarks are in red. The green lines represent the distance between the ground-truth landmarks and the predicted landmarks. Orange ellipses highlight variations between the methods in the challenging areas.

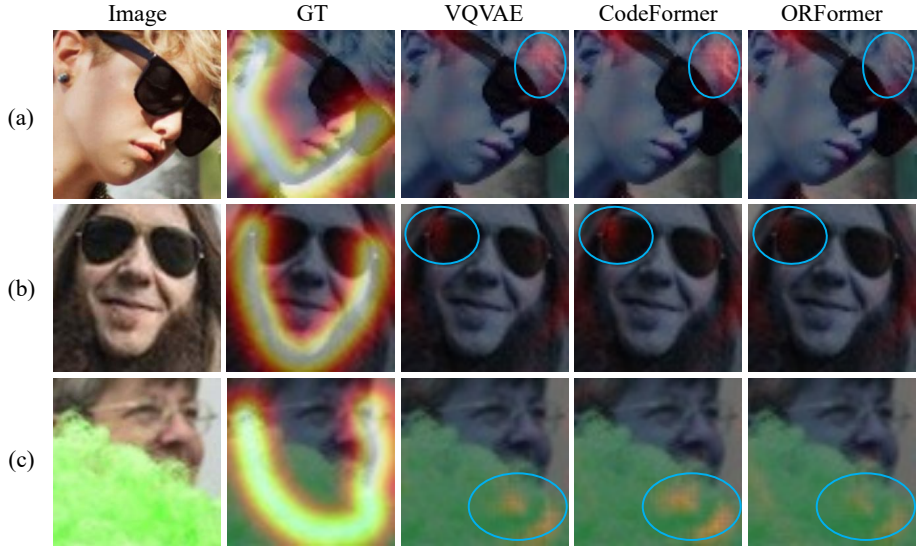

Figure 4. **Qualitative comparison for heatmap generation on WFLW.** GT stands for the ground-truth heatmap. For better visualization, we display the distance heatmap for VQVAE, Codeformer, and ORFormer by computing the pixel-wise L2 distance between their output heatmaps and the GT heatmap. Brighter areas indicate higher errors. The main area of discrepancy is emphasized within an ellipse to highlight variations between the methods.

difference as the distance function in the occlusion detection head yields the best performance. This improvement can be attributed to the squared difference function's capability to impose a larger penalty when there is a large disparity between the image patch embedding X^{l+1} and the messenger embedding M^{l+1}, while still enabling the gradient to propagate continuously.

To explore incorporating more information into OR-

Method	Integration	Pre-trained Weights	Trainable Part	NME_{io} ↓
ORFormer				4.03
ORFormer	✓	✓	Conv.	4.01
ORFormer	✓	✓	All	3.94
ORFormer	✓		All	**3.86**

Table 5. **Quantitative evaluation of different integration methods of ORFormer with Existing FLD Methods** The Conv. label indicates the 1×1 CNN block used to merge the heatmap generated by ORFormer with the feature maps of existing FLD methods. The first row entry represent reproducing STAR [18] without the integration with ORFormer. Results highlighted in **bold** represent the best performance. The landmark detection NME loss is reported on WFLW.

Method	Architecture	Loss Functions		WFLW (NME↓)	
		Heatmap	Landmark	Full	Occ.
ADNet [3]	HGNet	AWing	ADL	4.14	5.06
Ours	HGNet+ORFormer	AWing	ADL	4.06	4.94
†STAR [18]	HGNet	AWing	STAR	4.03	4.82
Ours	HGNet+ORFormer	AWing	STAR	3.92	4.66
Ours	HGNet+ORFormer	L2	NME	**3.86**	**4.57**

Table 6. **Quantitative evaluation of different loss functions of the integration of ORFormer with Existing FLD Methods.** All methods utilize the same backbone. Loss functions highlighted in blue represent the proposed approaches of that work. Results highlighted in bold represent the best performance. The landmark detection NME loss is reported on the WFLW dataset. The † symbol represents the results we reproduced.

Former during occlusion detection, we compare the heatmap regression quality using L2 loss with different filter sizes of W in Eq. 2, as shown in Table 4. For the convolutional layer, we reshape the embedding back to $\mathbb{R}^{m \times n \times d}$ before applying the convolution operation. In contrast, for the fully connected layer, we pass the embedding one by one, equivalent to applying a 1×1 convolutional layer in the shape of $\mathbb{R}^{m \times n \times d}$. Using a larger filter size for the convolutional layer allows the occlusion detection head to consider more information from neighboring embeddings when detecting occlusion. However, we observe that using a fully connected layer performs best. We believe this is because ORFormer operates in the latent space of the quantized heatmap generator, considering one single embedding in this latent space can provide an appropriate receptive field for occlusion detection in human faces.

2.2.4 Integration with FLD Methods

With our ORFormer for occlusion detection and feature recovery, the quantized heatmap generator can produce high-quality heatmaps. We integrate these heatmaps as additional structural guidance into existing FLD methods [3, 18]. Specifically, we concatenate the heatmaps with the feature maps in the early stage and merge them with a single lightweight 1×1 CNN block.

Method	Param. (M)	Self-Att.	Cross-Att.	Occ. Head	Occ.-Aware	L2 Loss ↓	NME↓
VQVAE [13]	1.36					26.72	4.04
CodeFormer [17]	4.32	✓				25.13	4.00
ORFormer (Ours)	4.77	✓	✓			24.35	3.99
	4.78	✓	✓	✓		23.87	3.95
	4.78	✓	✓	✓	✓	20.22	**3.86**

Table 7. **Quantitative evaluation on the proposed components of ORFormer on WFLW.** The heatmap regression L2 loss and he landmark NME loss are reported.

Way of Integration. To explore the best strategy of integrating the heatmap produced by ORFormer into existing FLD methods, we compare the landmark detection accuracy using NME loss with different integration strategies. The results are shown in Table 5. The pre-trained weights are from reproducing STAR [18] with an NME of 4.03. We find that by only fine-tuning the lightweight CNN block, we gain little performance with the help of ORFormer's heatmap. However, if we fine-tune the entire network or train the entire network from scratch without using pre-trained weights, we can achieve larger performance gains.

Loss Function. We also explore alternative choices of the loss function for model integration. As shown in Table 6, by integrating the output heatmaps of ORFormer into existing FLD methods [3, 18] and using the same loss functions, our approach achieves improved performance. Moreover, we obtain the best result using a simple loss function such as L2 loss for heatmap supervision and NME loss for landmark supervision. We believe this is because our heatmap definition differs from ADNet and STAR. While our heatmap is suitable for L2 loss, their heatmap is defined for the use of their proposed specific loss functions.

2.3. Computational Complexity of ORFormer

In Table 7, we show the numbers of trainable parameters of ORFormer. Compared to the conventional ViT, ORFormer enhances ViT to handle occlusions with minimal overhead, with about 10% more trainable parameters.

Even though ORFormer doubles the token count of a regular ViT, the patch token and messenger token compute attention scores separately, affecting the computational complexity linearly and thus minimally impacting the inference time. In Table 8, we integrate our proposed ORFormer into the state-of-the-art baseline, STAR [18], a 4-stack Hourglass network. For fair comparison, we augment the baseline network with one additional stack to align the number of trainable parameters. Our method performs favorably against this augmented baseline with comparable trainable

Method	Architecture	Param. (M)	Mult-Add (G)	Infer. Time (ms)	NME_{io} ↓
†STAR [18]	4-stack HGNet	17	17.4	45	4.03
†STAR [18]	5-stack HGNet	21.5	21.5	63	3.98
Ours	4-stack HGNet+ORFormer	21.8 (+1.4%)	17.9 (−20.6%)	53 (−15.9%)	**3.86**

Table 8. **Ablation study of computation complexity vs NME on WFLW.** HGNet represents the hourglass network. The relative increase/improvement is calculated based on 5-stack HGNet. The † symbol represents the results we reproduced. The inference time is tested on a single NVIDIA GTX 1080 Ti.

parameters, 20.6% fewer mult-add operations, and 15.9% less inference time, showing the advantage of ORFormer.

2.4. Limitations

The first limitation is that ORFormer is particularly effective at handling partially non-visible facial features but struggles with partially deformed facial features. The second limitation is that although ORFormer yields features robust to occlusion, the capability of our method relies on a well-trained quantized heatmap generator, which limits its applicability to tasks related to heatmap generation. In future research, we plan to explore ways to enable ORFormer to handle partially deformed facial features and extend ORFormer to serve as a general feature extractor for various computer vision tasks, where partial occlusions detection and feature recovery are essential, maximizing its impact in the field of computer vision.

References

[1] Xavier P Burgos-Artizzu, Pietro Perona, and Piotr Dollár. Robust face landmark estimation under occlusion. In *ICCV*, 2013. 1

[2] 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 *CVPR*, 2018. 2

[3] Yangyu Huang, Hao Yang, Chong Li, Jongyoo Kim, and Fangyun Wei. Adnet: Leveraging error-bias towards normal direction in face alignment. In *ICCV*, 2021. 2, 3, 5

[4] Haibo Jin, Shengcai Liao, and Ling Shao. Pixel-in-pixel net: Towards efficient facial landmark detection in the wild. *IJCV*, 2021. 2

[5] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1

[6] Abhinav Kumar, Tim K Marks, Wenxuan Mou, Ye Wang, Michael Jones, Anoop Cherian, Toshiaki Koike-Akino, Xiaoming Liu, and Chen Feng. Luvli face alignment: Estimating landmarks' location, uncertainty, and visibility likelihood. In *CVPR*, 2020. 2

[7] Xing Lan, Qinghao Hu, Qiang Chen, Jian Xue, and Jian Cheng. Hih: Towards more accurate face alignment via heatmap in heatmap. *arXiv preprint arXiv:2104.03100*, 2021. 2

[8] Jinpeng Li, Haibo Jin, Shengcai Liao, Ling Shao, and Pheng-Ann Heng. Repformer: Refinement pyramid transformer for robust facial landmark detection. In *IJCAI*, 2022. 2

[9] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 1

[10] Paul Micaelli, Arash Vahdat, Hongxu Yin, Jan Kautz, and Pavlo Molchanov. Recurrence without recurrence: Stable video landmark detection with deep equilibrium models. In *CVPR*, 2023. 2

[11] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *ICCV workshops*, 2013. 1

[12] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019. 2

[13] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. In *NeurIPS*, 2017. 5

[14] Xinyao Wang, Liefeng Bo, and Li Fuxin. Adaptive wing loss for robust face alignment via heatmap regression. In *ICCV*, 2019. 2

[15] 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. 1, 2

[16] Jiahao Xia, Weiwei Qu, Wenjian Huang, Jianguo Zhang, Xi Wang, and Min Xu. Sparse local patch transformer for robust face alignment and landmarks inherent relation learning. In *CVPR*, 2022. 2

[17] Shangchen Zhou, Kelvin Chan, Chongyi Li, and Chen Change Loy. Towards robust blind face restoration with codebook lookup transformer. In *NeurIPS*, 2022. 5

[18] Zhenglin Zhou, Huaxia Li, Hong Liu, Nanyang Wang, Gang Yu, and Rongrong Ji. Star loss: Reducing semantic ambiguity in facial landmark detection. In *CVPR*, 2023. 2, 3, 5, 6