Cascaded Dual Vision Transformer for Accurate Facial Landmark Detection

Technical Appendix

In this supplementary material, we provide more details and results omitted from the main paper for brevity. Specifically, in Sec. A, we introduce the GPU memory requirement; in Sec. B, we compare our method by training and testing networks with similar computational capacity; in Sec. C, we investigate the impact of input image resolution; and in Sec. D, we present visual comparisons on the COFW and 300W datasets.

A. GPU Memory Requirement

In Tab. S1, we report the memory required for each GPU during training, as well as the number of parameters for different numbers of prediction blocks.

#Pred. Blocks	2	4	6	8	10	12
Memory (GB)	4.4	6.3	8.2	10.3	12.1	14.2
#Param. (M)	24.4	48.4	72.4	96.4	120.4	144.4

Table S1. Memory required for each GPU during training, and number of parameters for different numbers of prediction blocks.

B. Comparison on Similar Compute Capacity

Our proposed Long Skip Connection avoid losing useful information due to intermediate supervision and make deeper network architectures feasible. However, improved performance is not solely attributed to in creased computational capacity. When we use 4 prediction blocks and reduce the dimension of the feature maps to (160, 32, 32), the number of parameters in our network is comparable to other baselines. As reported in Tab. S2, the NME score still surpasses the previous SOTA method LDEQ [2] by 0.08, indicating the effectiveness of our proposed architecture.

C. Comparison on Different Image Resolutions

We investigate the influence of different input image resolutions, as shown in Fig. S1. Specifically, D-VIT improves the performance by 0.09, 0.08 and 0.07 at resolutions of 64px, 128px, and 256px, respectively, indicating that our proposed method is not sensitive to the input image size.

Method	#Param. (M)	NME(↓)	FR_{10}(↓)	AUC_{10}(↑)
HIH [1]	22.7	4.08	2.60	60.5
SPIGA [3]	60.3	4.06	2.08	60.6
STAR [4]	13.4	4.02	2.32	60.5
LDEQ [2]	21.8	3.92	2.48	62.4
Ours_nstack4	21.0	3.84	2.44	63.3

Table S2. Comparisons on WFLW dataset. We reduce the number of network parameters to 21M, denoted as "Ours_nstack4". Our proposed method still shows effectiveness.

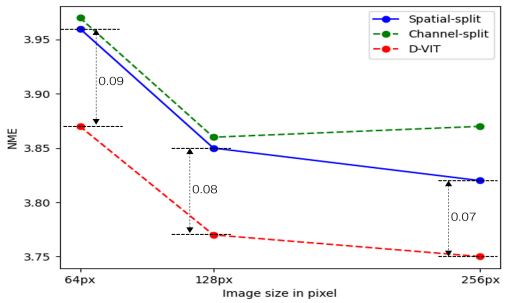


Figure S1. NME against different input image sizes on WFLW dataset.

D. Visual Results on COFW and 300W

In this section, we present the qualitative comparison results on COFW and 300W. Fig. S3 and Fig. S2 show the results of different prediction blocks. Fig. S4 and Fig. S5 show the comparisons of different skip connection strategies. With the help of our proposed D-ViT and LSC, the detection accuracy for landmarks is improved.

Spatial-split Channel-split D-ViT

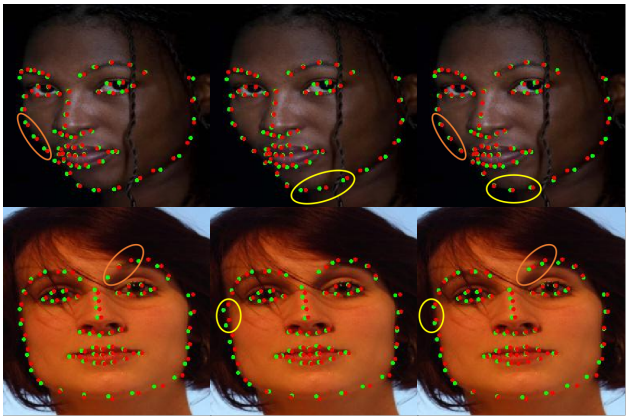
Figure S2. Visual comparison of different prediction blocks on 300W. Green and red points represent the predicted and ground-truth landmarks, respectively.

ResCBSP DenC LSC

Figure S4. Qualitative results of different skip connection strategies on COFW by using 8 prediction blocks. Green and red points represent the predicted and ground-truth landmarks, respectively.

Spatial-split Channel-split D-ViT

Figure S3. Visual results on the COFW dataset which contains heavy occlusions. Geometric relations among landmarks play a crucial role in accurately predicting landmarks on occluded parts (indicated by orange and yellow circles). Our D-ViT captures both semantic image features and the underlying geometric features among landmarks, enabling our model to make more accurate predictions even in the presence of occlusions.



ResCBSP DenC LSC

Figure S5. Qualitative results of different skip connection strategies on 300W by using 8 prediction blocks. Green and red points represent the predicted and ground-truth landmarks, respectively.

References

[1] Xing Lan, Qinghao Hu, and Jian Cheng. HIH: towards more accurate face alignment via heatmap in heatmap. *CoRR*, abs/2104.03100, 2021. 1

[2] Paul Micaelli, Arash Vahdat, Hongxu Yin, Jan Kautz, and Pavlo Molchanov. Recurrence without recurrence: Stable video landmark detection with deep equilibrium models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22814–22825, June 2023. 1

[3] Andrés Prados-Torreblanca, José M Buenaposada, and Luis Baumela. Shape preserving facial landmarks with graph attention networks. In *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022*. BMVA Press, 2022. 1

[4] Zhenglin Zhou, Huaxia Li, Hong Liu, Nanyang Wang, Gang Yu, and Rongrong Ji. Star loss: Reducing semantic ambiguity in facial landmark detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15475–15484, June 2023. 1