Generalizable Face Landmarking Guided by Conditional Face Warping

Supplementary Material

1. Introduction

This supplementary material provides the following information: Sec. 2 presents the implementation details in the experiments for the convenience of reproduction. Sec. 3 presents more ablation studies to further validate the effectiveness of our proposed framework. Sec. 4 provides sufficient visualization results as strong evidence for our method. Sec. 5 indicates some failure cases predicted by our model, analyzes the possible limitations then points out the direction of future work.

2. Training Details

We implement our landmark predictor f_θ as SLPT [11] in the above experiments. Each input image is cropped and resized to 256×256, and the training set is augmented with various transformations such as random horizontal flipping, grayscale, occlusion, scaling, rotation, and translation. We select HRNetW180 [9] as the backbone model, with a feature map resolution of 64×64.

To verify the impacts of different model architectures, we compare two different backbones: SBR [2] and HRnet [7]. For the SBR approach [2], we utilize CPM [10] as the detector, and VGG-16 [6] networks to initial four convolutional layers for feature extraction and only three CPM stages are used for heatmap prediction. For the HRNet technique [7], all faces are cropped based on their bounding boxes, centered using calculated formulas, and then resized to 256x256. After that, we perform Data augmentation on images using in-plane rotation, scaling, and random flipping.

3. More Ablation Studies

3.1. Effect of different pose dataset

Previously research employ warp methods such as AutoToon [4] and WarpGAN [5] for facial manipulation, as well as common flow prediction methods like FlowNet [3] and RAFT [8]. These methods either require a one-to-one correspondence between input images or assume minimal deformation between two images. Consequently, for facial images, the positioning of the face also influences the results. However, establishing a one-to-one correspondence between the 300W dataset and the CariFace dataset is challenging and time-consuming. To address this issue, we consider categorizing the datasets into three classes: frontal faces, faces turned right, and faces turned left, each comprising 1000 images. Subsequently, separate training is conducted for each category, and the results can be observed in Table 1.

Table 1. Ablation study on the usage of the dataset.

Settings		300W			
		ALL	Frontal	Left	Right
CariFace	ALL	7.831	7.695	7.879	8.080
	Frontal	8.466	9.077	8.768	7.771

From these results, we can observe variations in the outcomes when training with datasets containing different poses. Additionally, it is evident that training with all available datasets does not necessarily guarantee improved performance. Notably, the best results are achieved when utilizing the frontal 300W dataset in conjunction with all cartoon datasets. This could be attributed to the enhanced flexibility and effectiveness of warping processes when performed from a frontal perspective. That's why we choose this setting for our experiments. This innovative strategy not only improves the overall accuracy and reliability of facial landmark prediction but also simplifies the training process.

4. More Visualization

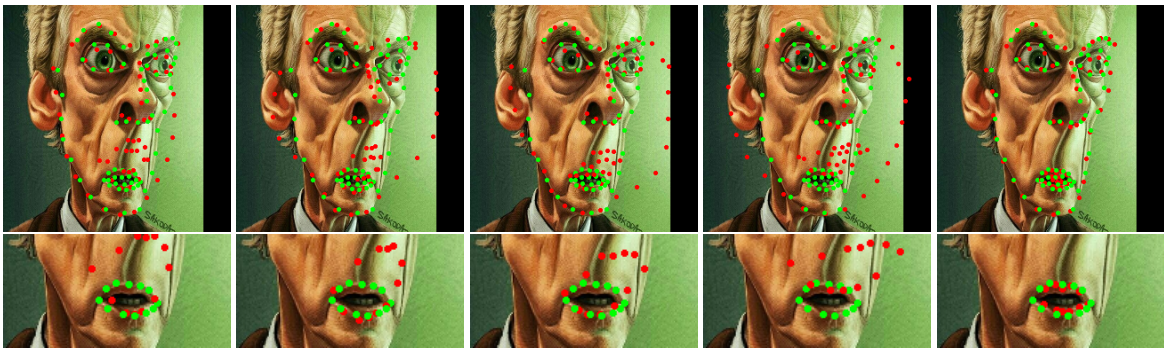
We present more samples to show landmark prediction results of our method under different styles and textures in the CariFace dataset, such as different facial expressions, various head poses, illumination, etc.

Fig. 1 and Fig. 2 demonstrate the effectiveness of our method in accurately predicting facial landmarks across various scenarios, including instances with exaggerated facial features. In Case 1, our method excels at determining the mouth's position when the distance between the nose and the mouth is significantly larger. Case 2 highlights our method's ability to precisely predict the eye edges when they are notably larger than other facial components. Furthermore, Case 3 and 4 showcase our method's capability to accurately estimate facial contours when the face is compressed in both vertical and horizontal directions.

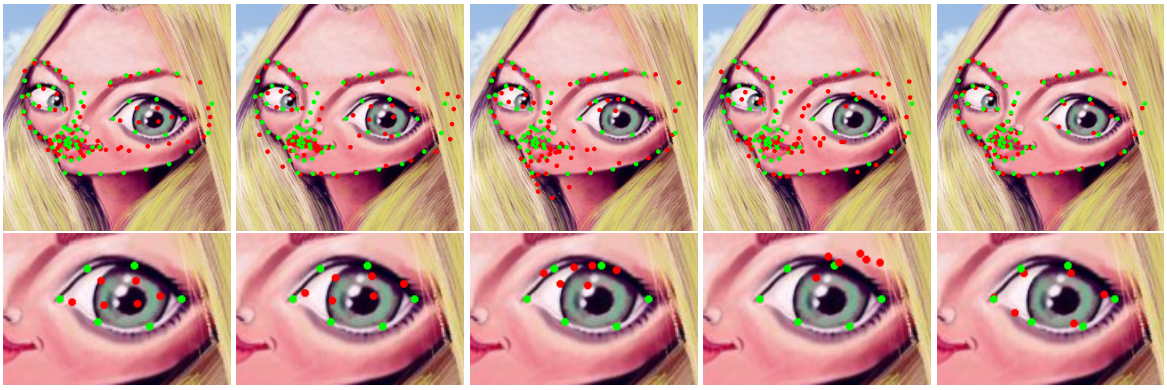
5. Limitations and Discussion

Based on our extensive experiments, our proposed method has achieved impressive results in unsupervised cartoon face landmark detection. Notably, our model exhibits robust performance even when applied to previously unseen domains, surpassing some supervised approaches in certain cases.

Yet, there is still progress for improvements particularly in challenging situations, such as severe occlusion, blur-



(a) Case 1



(b) Case 2

Figure 1. Visual comparisons for various methods in the two DA(300W→CariFace) settings.

ring, and extremely stylized facial contours, as illustrated in Fig. 3. To address these limitations, we have identified two specific issues: 1) our model may struggle to predict facial contours when there is uncertainty in the face boundary. 2) when applied to more challenging scenarios, such as anime dataset [1], our model encounters difficulty in adapting to the domain which is characterized by distinctive features like small noses, mouths, and larger eyes.

For these weaknesses, the most possible solution is to construct a more robust constraint between the warped faces and cartoon faces for better prediction. We leave it in the future work.

References

[1] G Branwen. Danbooru2019 portraits: A large-scale anime head illustration dataset. *Danbooru2019 portraits: A large-scale anime head illustration dataset*, 2019. 2

[2] Xuanyi Dong, Shoou-I Yu, Xinshuo Weng, Shih-En Wei, Yi Yang, and Yaser Sheikh. Supervision-by-registration: An unsupervised approach to improve the precision of facial landmark detectors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 360–368, 2018. 1

[3] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2758–2766, 2015. 1

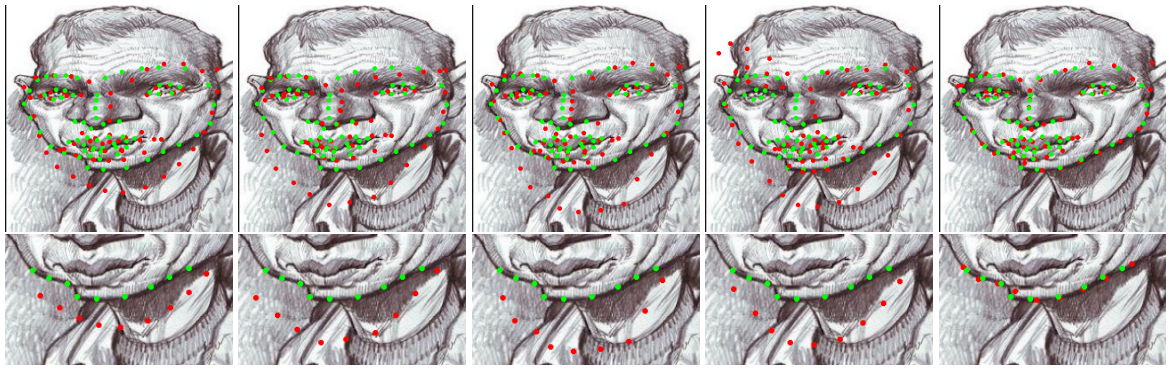
[4] Julia Gong, Yannick Hold-Geoffroy, and Jingwan Lu. Autotoon: Automatic geometric warping for face cartoon generation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 360–369, 2020. 1

[5] Yichun Shi, Debayan Deb, and Anil K Jain. Warpgan: Automatic caricature generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10762–10771, 2019. 1

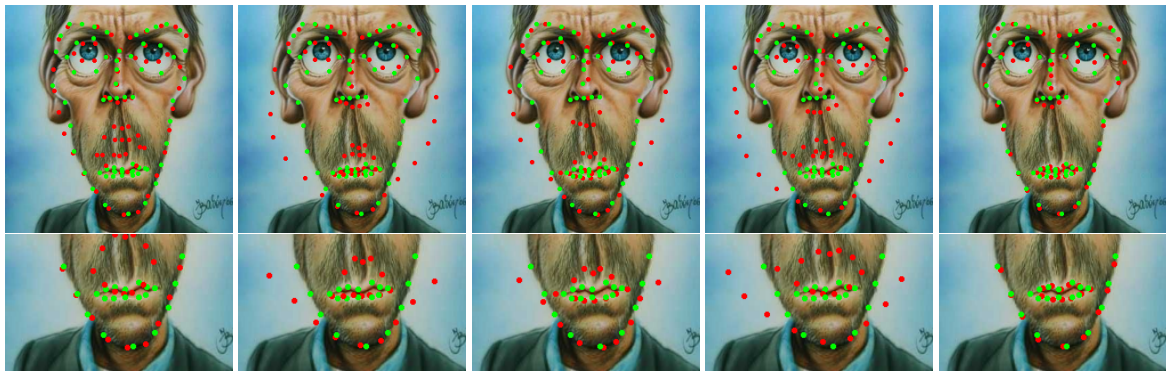
[6] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1

[7] Ke Sun, Yang Zhao, Borui Jiang, Tianheng Cheng, Bin Xiao, Dong Liu, Yadong Mu, Xinggang Wang, Wenyu Liu, and Jingdong Wang. High-resolution representations for labeling pixels and regions. *arXiv preprint arXiv:1904.04514*, 2019. 1

[8] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–*



(a) Case 3



(b) Case 4

Figure 2. Visual comparisons for various methods in the two DA(300W→CariFace) settings.

Figure 3. Visualizations of some typical failures. Red dots represent our predictions.

28, 2020, Proceedings, Part II 16, pages 402–419. Springer, 2020. 1

[9] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3349–3364, 2020. 1

[10] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 4724–4732, 2016. 1

[11] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4052–4061, 2022. 1