Supplementary Material

Zhonglin Sun
zhonglin.sun@qmul.ac.uk
Georgios Tzimiropoulos
g.tzimiropoulos@qmul.ac.uk

School of Electronic Engineering and Computer Science
Queen Mary university of London
London, UK

1 Introduction

This is the supplementary material for the paper **Part-based Face Recognition with Vision Transformers**. We first exhibit the detailed choice of data augmentation we used to enhance fViT in **Section** 2.1.1. Then we list the model details adopted for our fViT in **Section** 2.1.2. Effect of data augmentation, the overlapping rate of landmarks and the comparison of choice of landmark CNN are also included as the additional ablation study in **Section** 2.2. Finally, we describe learned landmark network is effective for the side task of Application to unsupervised landmark discovery in **Section** 2.3.

2 Additions to section 4: Experiments

2.1 Implementation details

2.1.1 Training details

For training the Transformer, we opted to use a large amount of data augmentation compared to the original face recognition setting used in ResNets. Specifically, we used stochastic depth regularization with probability 0.1 [6], resize & crop in the range $[0.9, 1.0]$, RandAugment [1] with magnitude of 2, and without the solarize and invert operations, Mixup [16] with alpha=0.5 and probability of 0.2, Cutout with value 0.1, and weight decay 1e-1 for the ViT backbone and 5e-2 for the Landmark CNN. We adopted AdamW [8] and the cosine learning rate decay followed by warm-up of 5 epochs, while we trained in total for 34 epochs. All networks are trained from scratch.

Model	Hidden size	Parameters	FLOPS
part fViT-B	768	66M	12.64G
fViT-B	768	63M	12.58G
Resnet-100	-	65M	12.10G
part fViT-S	512	46M	8.96G
fViT-S	512	43M	8.90G
Resnet-50	-	43.59M	6.33G

Table 1: Network sizes and FLOPS for our fViT and Part fViT and Resnet

© 2022. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

2.1.2 Model details

To fairly compare with Resnet [7] which is used as the backbone in most recent methods, we constructed our fViT in order to have a similar model size and FLOPS with Resnet-100. Our base configuration for fViT, called fViT-B, has 12 layers, 11 attention heads and $d = 768$. We also built a fViT-S. Our models and Resnet-100 are compared in Table 1. As can be observed, our fViT-B has similar model size and FLOPS with Resnet-100. Our landmark network is a MobilenetV3 [4] unless otherwise specified. All models are implemented in PyTorch [10].

2.2 Additional Ablation Study

2.2.1 Effect of different data augmentations

Here we present effectiveness of the choice of different augmentations suggested in [14] starting from random filp. Results can be found in 2, we can observe that with more data augmentation are added, more accurate results will be gained.

Exp	Flip	Randaug	Res&Crop	Stostich	Mixup	Cutout	Warm-up	LFW	CFP-FP	AgeDB-30	IJB-C
1	√							99.63	95.72	97.1	95.29
2	√	√						99.68	96.84	97.55	95.87
3	√	√	√					99.70	97.23	97.26	95.98
4	√	√	√	√				99.73	97.40	97.30	96.05
5	√	√	√	√	√			99.76	98.19	97.60	96.13
6	√	√	√	√	√	√		99.78	98.37	97.67	96.23
7	√	√	√	√	√	√	√	99.80	98.78	97.85	96.37

Table 2: Impact of data augmentation

2.2.2 Degree of overlapping

We also examiate the degree of overlapping patches trained by our network, we calculate the mean and variance overlap rate of the closest patches, listed in Table 3.The overlap rate for the large pose datasets CFP-FP& IJB-C is higher than that for other datasets.

	LFW	CFP-FP	AgeDB-30	IJB-C
R=16	0.5007±0.0002	0.5250±0.0016	0.4980±0.0002	0.5099±0.0007
R=49	0.3993±0.0002	0.4665±0.0064	0.3997±0.0001	0.4279±0.0003
R=196	0.2681±0.0001	0.2950±0.0010	0.2684±0.00008	0.2789±0.0005

Table 3: The overlap rate of the neighboring patches obtained by our part fVIT-B with R=16, 49 and 196

2.2.3 Effect of different landmark CNNs

We conducted an experiment to evaluate the impact of using different CNNs for landmark network. Specifically, we also chose Resnet-50[3]. The model used is the part fViT-B, with $R = 196$ landmarks. Table 4 shows the obtained results. We conclude that a larger landmark CNN does not further boost the final accuracy.

Landmark Network	LFW	CFP-FP	AgeDB	IJB-C
fViT	99.85	99.01	98.13	97.21
part fViT (MobilenetV3)	99.83	**99.21**	98.29	**97.29**
part fViT (ResNet50)	99.81	99.14	**98.35**	97.11

Table 4: Impact of landmark CNNs on face recognition accuracy.

2.3 Application to unsupervised landmark discovery

We opted for a quantitative evaluation of the facial landmarks discovered by our landmark CNN using the evaluation protocol and codebase of [12]. Specifically, we follow [12] and report the so-called forward error on the whole MAFL & AFLW datasets in Table 5. The forward error is a measure of landmark stability, its pipeline is to train a regressor with predicted landmarks as the training data and 5 manually labelled landmarks on the MAFL & AFLW datasets as the test set. The more stable the predicted landmarks are, the better they map to the ground truth (for details and forward error definition, please see[12]). As it can be observed our method offers competitive results with recently proposed methods which are exclusively designed for unsupervised landmark localization.

	Method	MAFL	AFLW
Supervised	TCDCN [19]	7.95	7.65
	MTCNN [18]	5.39	6.90
Unsupervised	Thewlis [15]	7.15	-
	Jakab [6]	3.19	6.86
	Zhang [17]	3.46	7.01
	Shu [13]	5.45	-
	Sahasrabudhe [11]	6.07	-
	Sanchez [12]	3.99	6.69
	Mallis [9]	4.12	7.37
	Li [7]	**3.08**	**6.20**
Ours	Landmark CNN	4.87	10.22
	Landmark CNN ($R = 49$)	3.37	7.16
	Landmark CNN ($R = 16$)	3.88	7.69

Table 5: Comparison on unsupervised landmark discovery. Forward error results [12] are reported on the whole MAFL & AFLW datasets.

References

[1] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 702–703, 2020.

[2] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning

for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[4] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1314–1324, 2019.

[5] Tomas Jakab, Ankush Gupta, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object landmarks through conditional image generation. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 4020–4031, 2018.

[6] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Fractalnet: Ultra-deep neural networks without residuals. *arXiv preprint arXiv:1605.07648*, 2016.

[7] Weijian Li, Haofu Liao, Shun Miao, Le Lu, and Jiebo Luo. Unsupervised learning of facial landmarks based on inter-intra subject consistencies. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 4077–4082. IEEE, 2021.

[8] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[9] Dimitrios Mallis, Enrique Sanchez, Matthew Bell, and Georgios Tzimiropoulos. Unsupervised learning of object landmarks via self-training correspondence. *Advances in Neural Information Processing Systems*, 33, 2020.

[10] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[11] Mihir Sahasrabudhe, Zhixin Shu, Edward Bartrum, Riza Alp Guler, Dimitris Samaras, and Iasonas Kokkinos. Lifting autoencoders: Unsupervised learning of a fully-disentangled 3d morphable model using deep non-rigid structure from motion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[12] Enrique Sanchez and Georgios Tzimiropoulos. Object landmark discovery through unsupervised adaptation. *arXiv preprint arXiv:1910.09469*, 2019.

[13] Zhixin Shu, Mihir Sahasrabudhe, Riza Alp Guler, Dimitris Samaras, Nikos Paragios, and Iasonas Kokkinos. Deforming autoencoders: Unsupervised disentangling of shape and appearance. In *Proceedings of the European conference on computer vision (ECCV)*, pages 650–665, 2018.

[14] Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your vit? data, augmentation, and regularization in vision transformers. *arXiv preprint arXiv:2106.10270*, 2021.

[15] James Thewlis, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object landmarks by factorized spatial embeddings. In *Proceedings of the IEEE international conference on computer vision*, pages 5916–5925, 2017.

[16] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[17] Yuting Zhang, Yijie Guo, Yixin Jin, Yijun Luo, Zhiyuan He, and Honglak Lee. Unsupervised discovery of object landmarks as structural representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2694–2703, 2018.

[18] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Facial landmark detection by deep multi-task learning. In *European conference on computer vision*, pages 94–108. Springer, 2014.

[19] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Learning deep representation for face alignment with auxiliary attributes. *IEEE transactions on pattern analysis and machine intelligence*, 38(5):918–930, 2015.