Supplementary material for Pre-training strategies and datasets for facial representation learning

Adrian Bulat$^{1[0000-0002-3185-4979]}$, Shiyang Cheng^{1}, Jing Yang$^{2[0000-0002-8794-4842]}$, Andrew Garbett^{1}, Enrique Sanchez$^{1[0000-0003-0196-922X]}$, and Georgios Tzimiropoulos$^{1,3[0000-0002-1803-5338]}$

[1] Samsung AI Cambridge
adrian@adrianbulat.com, {shiyang.c,a.garbett}@samsung.com,
kike.sanc@gmail.com
[2] University of Nottingham, Nottingham, UK
jing.yang2@nottingham.ac.uk
[3] Queen Mary University London, London, UK
g.tzimiropoulos@qmul.ac.uk

1 Implementation details

1.1 Unsupervised pretraining

For the unsupervised pretraining, similarly with [4] we trained our model on 64 GPUs using a batch size of 4096 and Synchronized Batch Normalization. The network was trained for 200 epochs using a weight decay of 10^{-6} and learning rate of 4.8 that was decayed toward 0.045 using a Cosine Scheduler [8]. During the first 10 epochs the learning rate is increased toward the target value using a linear scheduler. In all experiments, unless otherwise specified, we kept the temperature parameter to 0.1 and the Sinkhorn regularization parameters to 0.05. Each input sample was augmented into 2 views at a resolution of 224×224px and 6 at a resolution of 96×96px. The model was trained using the LARS [18] optimizer and was implemented in PyTorch [10].

Datasets and data preparation: All images are detected using [6] and then cropped based on the produced bounding-box so that the face will take approx. 190px on a 256×256px image. Unless otherwise specified all the data used for unsupervised pre-training were processed in the same manner.

1.2 Downstream task implementation details

Herein, we present the implementation details for each downstream task used in the main body to evaluate the efficacy of the facial representation learned. We note that in all cases the images were normalized in accordance with the training procedure of the pre-trained backbone model used as initialization.

Face recognition Following the best practices [16, 5], all images were normalized and aligned using the provided 5 landmarks. During training, the only augmentation applied was random horizontal flipping. Depending on the data regime, the models were trained between 18 and 54 epochs using a batch size of 512 and learning rate of 0.1. The weight decay was set to 0.0005 and the models were optimized using SGD with momentum (set to 0.9). For the cosface loss, the margin was set to 0.35. All models were trained on 8 GPUs.

Facial Landmark Localization The facial landmark localization pipeline was implemented following [2, 14]. During training, we applied the following augmentations randomly: rotation (between $\pm 30^o$), horizontal flipping, scaling ($0.85\times -1.15\times$) and color jittering. Depending on the data regime, dataset and pretrained model, as detailed in the main body of the work, we trained the models between 60 and 480 epochs using a learning rate of 0.0001, a batch size of 24, a weight decay of 10^{-5} and Adam optimizer [7] ($\beta_1 = 0.5, \beta_2 = 0.99$). All the models were trained using a pixel-wise ℓ_2 on a single GPU.

Action Unit (AU) Intensity Estimation For AU intensity estimation, we adopted a similar augmentation strategy with the one used for face alignment, mainly we applied random rotation ($\pm 30^o$), random horizontal flipping and scale jittering ($0.85\times -1.15\times$), Gaussian blurring with a kernel size between 5 and 10px and a probability of 0.4 and colour jittering. Depending on the setting, the models were trained between 60 and 320 epochs. The learning rate was typically set to 0.0001, the weight decay to 0.000005 and the batch size to 48. The models were optimized using Adam ($\beta_1 = 0.5, \beta_2 = 0.99$) and trained on 2 GPUs.

Emotion recognition For valence and arousal estimation, we applied the same augmentation strategies as for AU Intensity Estimation with the exception of Gaussian blurring. Depending on the setting, the models were trained between 60 and 240 epochs using a batch size of 32, a learning rate of 0.1, weight decay of 10^{-4} and Adam optimizer($\beta_1 = 0.5, \beta_2 = 0.99$). All models were trained on a single GPU.

3D Face reconstruction Since 300W-LP has a small number of identities, during training we randomly augment the data using the following transformations: scaling($0.85\times -1.15\times$), in-plane rotation ($\pm 45^o$), and random 10% translation w.r.t image width and height. Depending on the setting, we trained the model between 120 and 360 epochs using a learning rate of 0.05, a weight decay of 10^{-4} and SGD with momentum (set to 0.9). All models were trained using 2 GPUs.

1.3 Data sampling

For all low data scenarios, we randomly subsampled a set of annotated images without accounting for the labels (*i.e/* we don't attempt to balance the classes).

Once formed, the same subset is used for all subsequent experiments to avoid noise induced by different sets of images. For face recognition where the loss attempts to minimize the intra-class while maximising the inter-class distance and its sensitivity to both the number of identities and samples per identity, we deviated slightly from the above setting by enforcing that at least 1/4 of the identities are preserved for the very low data regime of 2%.

2 Curated Datasets

For unsupervised pre-training we explore 3 curated datasets, collected for various facial analysis tasks: (a) Full VGG-Face ($\sim 3.4M$), (b) Small VGG-Face ($\sim 1M$) and (c) Large-Scale-Face ($> 5.0M$), consisting of VGG-Face2 [3], 300W-LP [22], IMDb-face [15], AffectNet [9] and WiderFace [17]. During unsupervised pre-training we drop all labels using only the facial images. See supplementary material for more details.

a) *Full VGG-Face* denotes the entirety of the VGG-Face2 dataset [3], consisting of $\sim 3.4M$ facial images of 9131 identities, with an average of 362.6 images for each subject. Images are downloaded from Google Image Search and have large variations in pose, age, illumination, ethnicity and profession, although they typically depict celebrities.

b) *Small VGG-Face* is a randomly sampled subset of 1M images selected from VGG-Face2.

c) *Large-Scale-Face* is constructed by combining the facial images from VGG-Face2 [3], 300W-LP [22], IMDb-face [15], AffectNet [9] and WiderFace [17]. Therefore, the dataset combines a set of datasets originally collected for facial recognition, face alignment, emotion recognition and face detection:

300W-LP [22] is a face alignment dataset constructed by warping into large poses, from -90^o to 90^o, the ~ 4000 near-frontal images from the 300W [12] dataset. *IMDb-face* [15] is a large-scale noise-controlled dataset for face recognition, originally containing 1.7M faces with 59,000 identities which were manually cleaned by the authors from 2.0M raw images. All images were obtained by downloading data from the IMDb website. *AffectNet* [9] is a *in-the-wild* facial expression dataset consisting of more than 1M images collected by queering results from the internet using 1250 emotion related keywords. Out of this, 440,000 images were manually annotated with 7 discrete facial expressions and the intensity of valence and arousal. *WiderFace* [17] is a face detection benchmarking dataset consisting of 393,703 faces sourced from 32,203 images. The faces exhibit a high degree of variability in terms of scale, pose and occlusion.

3 Uncurated Flick-Face dataset

Herein we provide additional details regarding the collected uncurated, in-the-wild, Flickr-Face dataset. The dataset was constructed by downloading a set of images from Flickr. The facial images were then automatically localized and cropped using a face detector [6]. In order to increase the likelihood of finding

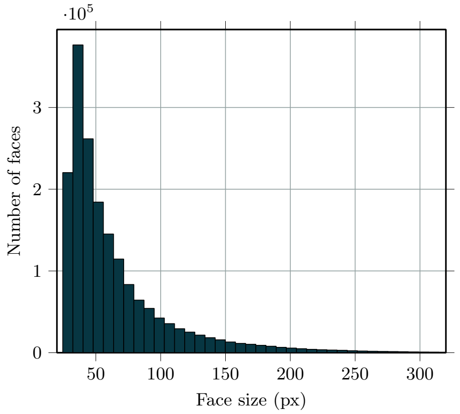


Fig. 1: Distribution on face sizes in the uncurated Flickr-Face dataset.

a face in the image we downloaded images that have one of the following 100 tags: *human, people, person, face, fashion, portrait, emotion, expression, affect, happy, sad, anger, angry, smile, laugh, joy, surprise, disgust, confused, fear, horror, adult, lady, ladies, beauty, gentleman, gentlemen, man, men, woman, women, baby, infant, toddler, kid, child, children, senior, father, mother, dad, mom, elderly, grandfather, grandmother, grandpa, grandma, grandparent, ancestor, 40s, 50s, 60s, 70s, 80s, 90s, couple, family, brother, sister, sibling, cousin, wedding, marriage, funeral, party, formal, boy, girl, teen, teenager, youth, friend, classmate, group photo, team, gathering, teacher, professor, lecturer, coach, tutor, worker, boss, celebrity, sport, self, selfie, photoshoot, concert, gigs, band, dance, marathon, passenger, army, soldier, marching, military, protest, crowds.* In total we collected 1.793.119 facial images with a bounding box size that follows the distribution shown in Fig. 1. We release the code used to download the images from Flickr thus allowing reproducing the dataset.

4 Additional results

Herein, we report results for AU intensity estimation and emotion recognition (see Section 4.1 and Tables 1, 2 and 3).

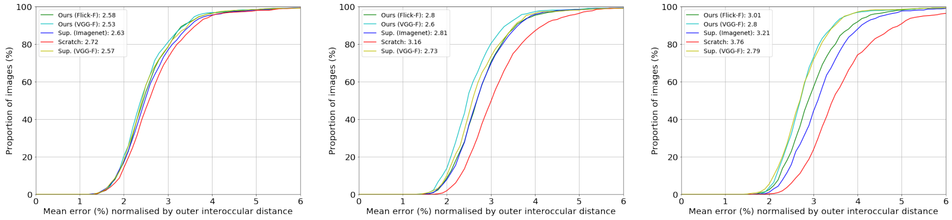
4.1 Emotion Recognition

We fine-tuned the models for valence and arousal estimation on the well-established AffectNet [9]. We report results in terms of RMSE and CCC [11], SAGR and PCC. The task specific head $h(.)$ is a linear layer that regresses the valence and arousal values and also predicts the basic emotion classes. The network was trained to jointly minimise the RMSE and CCC losses for valence and arousal, and the cross-entropy loss for classification.

Results are shown in Table 3 again, *for all* data regimes, our unsupervised models outperform the supervised baselines.

4.2 Additional 3D Face Reconstruction results

Furthermore, in Fig. 2 we report results on the Florence dataset for the task of 3D face reconstruction.



(a) Trained on 100% of the data.

(b) Trained on 10% of the data.

(c) Trained on 2% of the data.

Fig. 2: Cumulative 3D reconstruction error curves on the Florence [1] dataset for 3 different supervised data regimes: (a) using 100%, (b) 10% and (c) 2%. All models were trained on the 300W-LP dataset as detailed in the main body.

Table 1: Comparison against state-of-the-art on few-shot Facial AU intensity estimation on the DISFA dataset.

Method	**Data amount**	**AU**												**Avg.**
		1	**2**	**4**	**5**	**6**	**9**	**12**	**15**	**17**	**20**	**25**	**26**	
KBSS [19]	1%	.136	.116	.480	.169	.433	.353	.710	.154	.248	.085	.778	.536	.350
KJRE [21]	6%	.270	.350	.250	.330	.510	.310	.670	.140	.170	.200	.740	.250	.350
CLFL [20]	1%	.263	.194	.459	.354	.516	.356	.707	.183	.340	**.206**	.811	.510	.408
SSCFL [13]	2%	.327	.328	.645	.024	**.601**	.335	.783	.181	.243	.078	.882	.578	.413
Ours	1%	**.636**	**.667**	**.754**	**.367**	.549	**.535**	**.820**	**.313**	**.541**	.199	**.928**	**.608**	**.574**

Table 2: Comparison against state-of-the-art on few-shot Facial AU intensity estimation on the BU4D dataset.

Method	**Data amount**	**AU**					**Avg.**
		6	**10**	**12**	**14**	**17**	
KBSS [19]	1%	.760	.725	.840	.445	.454	.645
KJRE [21]	6%	.710	.610	.870	.390	.420	.600
CLFL [20]	1%	.766	.703	.827	.411	**.600**	.680
SSCFL [13]	2%	.766	.749	.857	.475	.553	.680
Ours	**1%**	**.789**	**.756**	**.882**	**.529**	.578	**.707**

Table 3: Results on the emotion recogntion task on the AffectNet dataset.

Data amount	**Init. method**	**Acc.**	**Valence**				**Arousal**			
			RMSE	**SAGR**	**PCC**	**CCC**	**RMSE**	**SAGR**	**PCC**	**CCC**
100%	random	0.590	0.370	0.790	0.696	0.695	0.339	0.781	0.613	0.611
	imagenet	0.592	0.360	0.789	0.705	0.705	**0.327**	0.792	0.624	0.620
	vggface	0.601	0.369	**0.798**	0.707	0.706	0.330	**0.796**	0.625	0.624
	ours	**0.602**	**0.356**	0.793	**0.711**	**0.710**	0.328	0.793	**0.634**	**0.629**
10%	random	0.493	0.402	0.752	0.626	0.625	0.366	0.753	0.536	0.536
	imagenet	0.548	0.383	**0.784**	0.655	0.654	0.351	0.767	0.569	0.566
	vggface	0.529	0.401	0.755	0.636	0.634	0.372	0.750	0.532	0.526
	ours	**0.562**	**0.382**	0.780	**0.678**	**0.678**	**0.344**	**0.803**	**0.600**	**0.599**
2%	random	0.419	0.453	0.727	0.515	0.515	0.400	0.747	0.423	0.422
	imagenet	0.479	0.411	0.740	0.562	0.557	0.362	0.769	0.465	0.456
	vggface	**0.511**	0.416	**0.778**	0.610	**0.607**	0.384	0.768	0.485	**0.485**
	ours	0.495	**0.370**	0.763	**0.620**	0.593	**0.338**	**0.794**	**0.500**	0.471

References

1. Bagdanov, A.D., Del Bimbo, A., Masi, I.: The florence 2d/3d hybrid face dataset. In: Proceedings of the 2011 joint ACM workshop on Human gesture and behavior understanding. pp. 79–80 (2011)
2. Bulat, A., Tzimiropoulos, G.: How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1021–1030 (2017)
3. Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: Vggface2: A dataset for recognising faces across pose and age. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). pp. 67–74. IEEE (2018)
4. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. arXiv preprint arXiv:2006.09882 (2020)
5. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4690–4699 (2019)
6. Deng, J., Guo, J., Zhou, Y., Yu, J., Kotsia, I., Zafeiriou, S.: Retinaface: Single-stage dense face localisation in the wild. arXiv preprint arXiv:1905.00641 (2019)
7. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
8. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016)
9. Mollahosseini, A., Hasani, B., Mahoor, M.H.: Affectnet: A database for facial expression, valence, and arousal computing in the wild. IEEE Transactions on Affective Computing **10**(1), 18–31 (2017)
10. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems 32, pp. 8024–8035. Curran Associates, Inc. (2019), http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf
11. Ringeval, F., Schuller, B., Valstar, M., Cowie, R., Pantic, M.: Avec 2015: The 5th international audio/visual emotion challenge and workshop. In: Proceedings of the 23rd ACM international conference on Multimedia. pp. 1335–1336 (2015)
12. Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., Pantic, M.: 300 faces in-the-wild challenge: The first facial landmark localization challenge. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. pp. 397–403 (2013)
13. Sanchez, E., Bulat, A., Zaganidis, A., Tzimiropoulos, G.: Semi-supervised au intensity estimation with contrastive learning. arXiv preprint arXiv:2011.01864 (2020)
14. Sun, K., Zhao, Y., Jiang, B., Cheng, T., Xiao, B., Liu, D., Mu, Y., Wang, X., Liu, W., Wang, J.: High-resolution representations for labeling pixels and regions. arXiv preprint arXiv:1904.04514 (2019)
15. Wang, F., Chen, L., Li, C., Huang, S., Chen, Y., Qian, C., Change Loy, C.: The devil of face recognition is in the noise. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 765–780 (2018)
16. Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z., Liu, W.: Cosface: Large margin cosine loss for deep face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5265–5274 (2018)

17. Yang, S., Luo, P., Loy, C.C., Tang, X.: Wider face: A face detection benchmark. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
18. You, Y., Gitman, I., Ginsburg, B.: Large batch training of convolutional networks. arXiv preprint arXiv:1708.03888 (2017)
19. Zhang, Y., Dong, W., Hu, B.G., Ji, Q.: Weakly-supervised deep convolutional neural network learning for facial action unit intensity estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2314–2323 (2018)
20. Zhang, Y., Jiang, H., Wu, B., Fan, Y., Ji, Q.: Context-aware feature and label fusion for facial action unit intensity estimation with partially labeled data. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 733–742 (2019)
21. Zhang, Y., Wu, B., Dong, W., Li, Z., Liu, W., Hu, B.G., Ji, Q.: Joint representation and estimator learning for facial action unit intensity estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3457–3466 (2019)
22. Zhu, X., Lei, Z., Liu, X., Shi, H., Li, S.Z.: Face alignment across large poses: A 3d solution. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 146–155 (2016)