Supplementary Material for Self-Supervised Facial Representation Learning with Facial Region Awareness

Zheng Gao Ioannis Patras

Queen Mary University of London, Mile End Road, London, E1 4NS
{z.gao, i.patras}@qmul.ac.uk

1. Discussions

1.1. Discussion with LEWEL

Both LEWEL [10] and our FRA seek to encourage the model to look into local regions instead of treating image as a whole. However, our FRA is fundamentally different from LEWEL [10] in the following aspects:

- **Motivation**. We learn the heatmaps from a set of learnable positional embeddings, which leverages the attention mechanism to globally look up the image for visual patterns. By contrast, LEWEL [10] only derives the heatmaps from CNN-based local features and thus lacks the capability to capture global information. Moreover, we explicitly capture the pairwise relations between the visual patterns (regions) represented by the facial mask embeddings by aligning the per-pixel assignments of each pixel feature over the facial mask embeddings (*i.e*., each pixel feature should have similar similarity distribution over the facial mask embeddings between the momentum teacher and online student) while LEWEL [10] simply matches the visual patterns across views independently, failing to look into the such pairwise relations.
- **Results**. Our FRA significantly outperforms LEWEL [10] on various downstream facial analysis tasks, *e.g*., our FRA achieves 66.16 on AffectNet facial expression recognition while LEWEL only has 61.20. The experimental results demonstrate the superiority of our method compared with LEWEL [10] on facial representation learning.

1.2. Discussion with SwAV

From the perspective of deep clustering [4], the facial mask embeddings learned in our FRA can be viewed as prototypes (clusters) and the heatmap prediction via the correlation between the pixel features and facial mask embeddings can be viewed as assigning pixel features to different prototypes. However, the differences between our FRA and deep clustering approach SwAV [4] are as follows:

- **Motivation**. Our FRA leverages the attention mechanism to obtain the facial mask embeddings (prototypes) by using queries to globally look up the face image while SwAV [4] is limited by the local information of CNN. In addition, we enforce the consistency of the local visual patterns across augmented views by applying a region-level contrastive objective over the discovered visual patterns while SwAV [4] simply treats each image as a whole by learning image-level representations and overlooks the consistency of visual patterns.
- **Results**. As shown in Tab. 1, our FRA significantly outperforms SwAV [4] on downstream facial analysis tasks under the settings of few-shot and fine-tuning. Note that Bulat *et al*. [2] is equivalent to SwAV [4] by pre-training SwAV [4] model on face images.

2. Additional experimental results

2.1. Transfer learning with limited data

In Tab. 1, we evaluate the transfer performance with limited labeled data under few-shot settings. We randomly sample subsets from the training sets of the downstream data and then evaluate the models on the full test sets. Following [2], we fine-tune both the encoder backbone and task-specific head on downstream data. For facial expression recognition on AffectNet, we use the same training recipe for 1%, 10% and 100% data. For 1% and 10% face alignment data, we fine-tune the model for 100 epochs with head learning rate 0.001, encoder backbone learning rate 5×10^{-5}, using the AdamW optimizer with step decay at 80 and 90 epochs. Our FRA achieves the best few-shot performances. In particular, our FRA outperforms SOTA few-shot face alignment approaches, 3FabRec [1] and He *et al*. [9].

2.2. Results w.r.t. pre-training epochs

In Tab. 2, we report the transfer learning performance w.r.t. pre-training epochs, from 50 ep to default 400 ep. We observe: **(1)** As a self-supervised pre-training approach, our

Table 1. **Transfer learning with limited data on facial expression recognition (AffectNet) and face alignment (WFLW)**. Bulat *et al.* is evaluated with 0.7%, 10% and 100% data on WFLW as in the original paper [2].

Methods	1%	10%	100%
Bulat *et al.* [2, 4]	27.48	51.45	60.20
FRA	**34.81**	**55.48**	**66.16**

(a) AffectNet (Acc. ↑)

Methods	1%	5%	10%	20%	100%
3FabRec [1]	-	7.68	6.73	6.51	5.62
He *et al.* [9]	-	6.22	-	5.61	5.38
Bulat *et al.* [2, 4]	7.11	-	5.44	-	4.57
FRA	**7.04**	-	**4.98**	-	**4.11**

(b) WFLW (NME ↓)

Table 2. Transfer learning performance w.r.t. pre-training epochs.

Epochs	50 ep	200 ep	400 ep
RAF-DB	88.72	89.37	89.95
CelebA	91.18	91.68	92.02
300W	3.14	2.99	2.91

FRA benefits from long time pre-training. **(2)** With 200 epoch pre-training, the performance is close to 400 ep.

3. Visualization of learned heatmaps

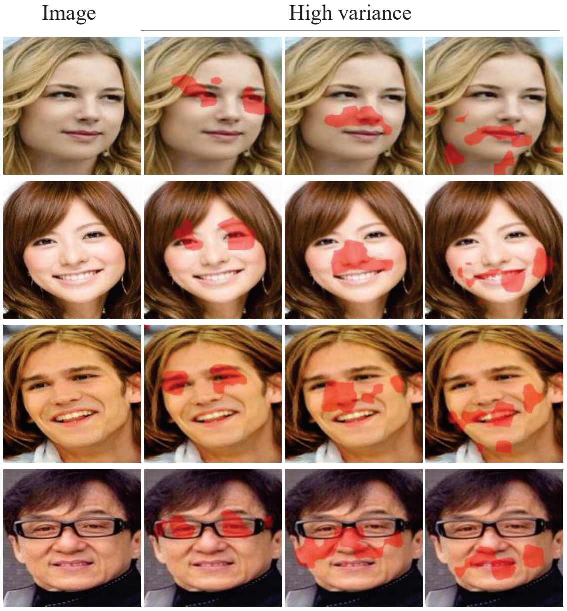

Figure 1. **Visualization of high-variance heatmaps of our method**. The first column is the original facial images. The 2-4 columns are the visualization of the top three high-variance heatmaps.

To better understand the facial mask embeddings and heatmap prediction, we visualize the learned heatmaps on VGGFace2 in Fig. 1. Following [12], we visualize the facial mask embeddings with the highest variances. We observe that the heatmaps can identify the rough location of the facial landmarks (*e.g.*, eyes, nose, mouth).

4. Training cost analysis

Table 3. **Comparison of pre-training running time with BYOL [7]**. The time relative to BYOL [7] is reported.

Method	Time/ Epoch	AffectNet	CelebA	300W
BYOL [7]	1.00	65.65	91.56	3.03
FRA	1.05	66.16	92.02	2.91

In Tab. 3, we compare pre-training cost with BYOL [7], the self-supervised pre-training framework our FRA is built upon. We report the time cost of a single training epoch ("Time/Epoch") relative to BYOL [7]. Our method outperforms BYOL [7] in all tasks with negligible training overhead (5% increase in cost).

5. Additional implementation details

5.1. Architecture

Following the common practice in self-supervised pre-training [7, 10], we use ResNet-50 [8] as the encoder backbone. The projectors H^g and H^l are implemented with a two-layer multi-layer perceptron (MLP) with Batch Normalization (BN) and ReLU activation, following [10]. The hidden/output dimension are set to 4096/256 as in BYOL [7], *i.e.*, $D = 256$. The predictors adopt the same architecture as the projectors. We use the same Transformer decoder [13] architecture as in MaskFormer [5] except only one decoder layer is employed to keep a lightweight architecture.

5.2. Pre-training

By default, we perform the pre-training on the training set of VGGFace2 [3] with 2 NVIDIA A100 GPUs. We detect the faces from the images using a face detector [6], randomly select one of the faces and then resize the cropped face to 128×128 for pre-training. Following [10], we pre-train for 400 epochs with batch size 1024, LARS optimizer [15] with 1.8 learning rate, 1.5×10^{-6} weight decay, and 0.9 momentum. The cosine annealing schedule [11] is used for learning rate decay. In ablation studies, the pre-training is performed on VGGFace2 [3] for 50 epochs for fast training.

5.3. Transfer learning

5.3.1 Facial expression recognition

We resize the cropped and aligned facial images to 224×224 before feeding them to the downstream model, following [16]. We adopt 3 variants for the evaluation: "FRA (LP)", "FRA (FT)" and "FRA (EAC)". For "FRA (LP)", "FRA (FT)", we add a linear classifier that consists of a linear fully-connected layer on top of the encoder backbone E_θ to project the latent space to the downstream task specific label space. For "FRA (EAC)", we use our pre-trained model to initialize the backbone of a SOTA facial expression recognition method, EAC [16]. We use the AdamW optimizer with 0.05 weight decay for fine-tuning. The model is fine-tuned for 100 epochs, with a batch size of 256 and cosine learning rate decay [11]. The learning rate is set to 1.0/0.005 for "FRA (LP)" and "FRA (EAC)", respectively. For "FRA (FT)", the optimal learning rates for different datasets are reported in Tab. 4.

Table 4. Facial expression recognition fine-tuning learning rate.

	FERPlus	RAF-DB	AffectNet
Backbone	2×10^{-4}	2×10^{-4}	10^{-4}
Head	2×10^{-4}	2×10^{-4}	2×10^{-3}

5.3.2 Facial attribute recognition

Following [17], we resize the cropped and aligned facial images to 224×224. We add a linear classifier (*i.e.*, classification head) on top of the encoder backbone (*i.e.*, feature extractor backbone) E_θ. Two variants are adopted: "FRA (LP)" for linear probing and "FRA (FT)" for fine-tuning. The models are fine-tuned for 100 epochs using AdamW optimizer with batch size of 256 and cosine annealing schedule [11]. For "FRA (LP)", classification head learning rate is set to 1.0 while the feature extractor backbone is fixed. For "FRA (FT)", we fine-tune the models with classification head learning rate 2×10^{-5}, feature extractor backbone learning rate 2×10^{-5}.

5.3.3 Face alignment

We evaluate the transfer learning performance on face alignment by transferring our learned general facial representation to the model of a SOTA face alignment method, STAR [18]. STAR [18] adopts hourglass network [14] as the backbone for extracting image features. In contrast, we replace the hourglass network [14] with ResNet pre-trained on VGGFace2 [3]. Three deconvolutional layers are added on top of the last ResNet layer to generate 64×64 heatmap like STAR [18]. We fine-tune the model for 100 epochs using the AdamW optimizer with 5×10^{-4} learning rate decayed at 80 and 90 epochs and 128 batch size. The other hyper-parameters are kept the same as in STAR [18] for a fair comparison.

References

[1] Björn Browatzki and Christian Wallraven. 3fabrec: Fast few-shot face alignment by reconstruction. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6109–6119, 2020. 1, 2

[2] Adrian Bulat, Shiyang Cheng, Jing Yang, Andrew Garbett, Enrique Sanchez, and Georgios Tzimiropoulos. Pre-training strategies and datasets for facial representation learning. In *Computer Vision – ECCV 2022*, pages 107–125, Cham, 2022. Springer Nature Switzerland. 1, 2

[3] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74. IEEE, 2018. 2, 3

[4] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 1, 2

[5] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. In *Advances in Neural Information Processing Systems*, pages 17864–17875. Curran Associates, Inc., 2021. 2

[6] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5202–5211, 2020. 2

[7] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap your own latent - a new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, pages 21271–21284. Curran Associates, Inc., 2020. 2

[8] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 2

[9] Xingzhe He, Gaurav Bharaj, David Ferman, Helge Rhodin, and Pablo Garrido. Few-shot geometry-aware keypoint localization. In *2023 IEEE/CVF Conference on Computer Vi-*

sion and Pattern Recognition (CVPR), pages 21337–21348, 2023. 1, 2

[10] Lang Huang, Shan You, Mingkai Zheng, Fei Wang, Chen Qian, and Toshihiko Yamasaki. Learning where to learn in cross-view self-supervised learning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14431–14440, 2022. 1, 2, 3

[11] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017. 3

[12] Yangyang Shu, Anton van den Hengel, and Lingqiao Liu. Learning common rationale to improve self-supervised representation for fine-grained visual recognition problems. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11392–11401, 2023. 2

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. 2

[14] Jing Yang, Qingshan Liu, and Kaihua Zhang. Stacked hourglass network for robust facial landmark localisation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2025–2033, 2017. 3

[15] Yang You, Igor Gitman, and Boris Ginsburg. Large batch training of convolutional networks. *arXiv preprint arXiv:1708.03888*, 2017. 3

[16] Yuhang Zhang, Chengrui Wang, Xu Ling, and Weihong Deng. Learn from all: Erasing attention consistency for noisy label facial expression recognition. In *Computer Vision – ECCV 2022*, pages 418–434, Cham, 2022. Springer Nature Switzerland. 3

[17] Yinglin Zheng, Hao Yang, Ting Zhang, Jianmin Bao, Dongdong Chen, Yangyu Huang, Lu Yuan, Dong Chen, Ming Zeng, and Fang Wen. General facial representation learning in a visual-linguistic manner. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18676–18688, 2022. 3

[18] 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, 2023. 3