General Facial Representation Learning in a Visual-Linguistic Manner Supplementary Material

Yinglin Zheng[1*] Hao Yang[2*] Ting Zhang[2] Jianmin Bao[2] Dongdong Chen[3]
Yangyu Huang[2] Lu Yuan[3] Dong Chen[2] Ming Zeng[1†] Fang Wen[2]
[1]School of Informatics, Xiamen Unversity
[2]Microsoft Research Asia [3]Microsoft Cloud+AI
{zhengyinglin@stu., zengming@}xmu.edu.cn, cddlyf@gmail.com,
{haya, tinzhan, jianbao, yangyu.huang, luyuan, doch, fangwen}@microsoft.com
https://github.com/faceperceiver/farl

1. Evaluation on Face Editing Tasks.

Here we adopt a recent text-driven face editing framework [13], which uses a pre-trained CLIP for visual-language reasoning. We replace CLIP with our FaRL (with equal model size), and show comparisons below. It can be observed that, the generated face images which are driven by FaRL are more faithful to the given text prompts.

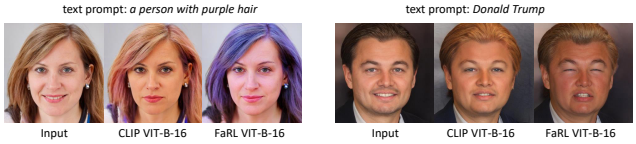


Figure 1. Comparing FaRL with CLIP in text-driven face editing.

2. Visualizing the Pre-trained Image Encoder

In Figure 2, we provide the Grad-CAM [20] visualizations for the pre-trained FaRL image encoder E_I, with different text queries fed into the text encoder E_T. Gradients are calculated in the output of the first LayerNorm within the last Transformer block of E_I. As can be seen in the figure, our image encoder successfully localizes the corresponding regions for different query texts, showing a high correlation with human attention.

3. Features on Different Backbone Levels

Instead of integrating multi-level features in downstream tasks, we also study how features on each single level of E_I affect the performances. We replace the multi-level features with repeated single feature on each level and apply the same head for downstream evaluation with backbone frozen. Figure 3a and Figure 3b illustrate the corresponding performances on LaPa (face parsing) and CelebA (face attributes recognition), respectively. In these two figures, a larger backbone level number indicates a deeper feature.

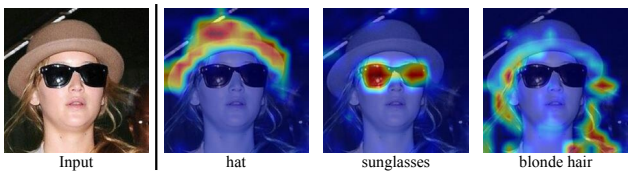
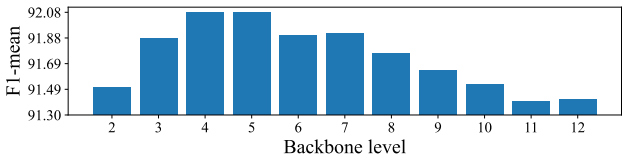


Figure 2. Grad-CAM visualizations of E_I given different text queries. Gradients are calculated in output of the first LayerNorm within the last Transformer block of E_I.

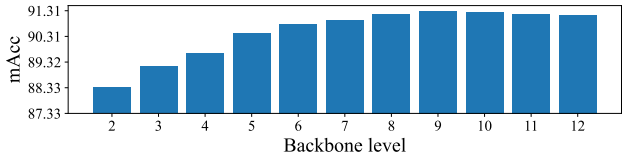
It is interesting to see that features on different backbone levels behave in quite divergent ways for different downstream tasks. The features on deep levels (e.g level 9) are the most effective for face attributes recognition on CelebA. However, they perform poorly on face parsing: the most effective feature for face parsing is on the 5-th level instead. This suggests that 1) the encoder E_I has learned different kinds of semantics on different feature levels during pre-training; 2) different kinds of downstream tasks require different kinds of semantics. Tasks like face attributes recognition rely more on high-level semantics, while face parsing is more in favor of low-level ones. In consideration of this divergence, it might not be the best way to always use the feature from just one single level for all downstream tasks.

Besides, we observe that the fusion of multi-level features, which is adopted by FaRL, achieves 92.32 (F1-mean) on LaPa and 91.39 (mAcc) on CelebA, outperforming all single-level settings. This indicates a complementary nature among features on different backbone levels.

*Equal contribution.
†Corresponding author.



(a) LaPa (face parsing).



(b) CelebA (face attributes recognition).

Figure 3. Performance of features from different backbone levels on face parsing (LaPa) and face attributes recognition (CelebA).

4. Ratios of Face Images in Pre-training

We are also curious about how the ratio of face images used in pre-training affects the performances of downstream face tasks. Therefore, we randomly sample different numbers of face images from LAION [19] and construct different pre-training datasets for investigation. All these datasets share the same size with LAION-FACE (20M), but with different ratios of face images within. We conduct pre-training on these datasets using only image-text contrastive learning.

The corresponding performances of different downstream face tasks are illustrated in Table 1. In general, a higher ratio of face images leads to a better performance. The gains are significant on tasks like face attributes recognition, but are quite subtle on tasks like face parsing and face alignment.

Such observation matches our previous hypothesis, that different downstream tasks require different kinds of semantics. Those high-level facial semantics, which are necessary for face attributes recognition, can only be learned on face images; while those low-level semantics (*e.g.* corners, edges), which are required by face parsing and face alignment, can be learned from not only face images, but also non-face images as well. This also explains why the advantages of our FaRL shown on face parsing and face alignment over those Transformers pre-trained on general images (*e.g.* ImageNet, WIT), are not as large as the advantages exhibited on face attributes recognition.

5. Comparison with Self-supervised Methods on Face Dataset

Here we compare FaRL with two recent self-supervised methods: SwAV [2] and SimCLR [3] on LAION-Face with the same network structure and fine-tuning strategy with FaRL. Note that SwAV on face images is equivalent to Bulat [1]. as it adopts SwAV for face pre-training. As shown in Table 2, FaRL achieves better performances on all tasks. We will add this comparison in the final paper.

% of Face Images	LaPa F1-mean↑	AFLW-19 NME_{diag} ↓	CelebA mAcc↑
0	91.68	1.017	89.73
12.5	91.68	1.010	90.76
50	**91.77**	**1.009**	91.17
100	91.75	**1.009**	**91.31**

Table 1. Downstream performances w.r.t different face image ratios in pre-training data. Here the pre-training only uses image-text contrastive learning.

Pre-training Settings	LaPa F1-mean↑	AFLW-19 NME_{diag} ↓	CelebA mAcc↑
(Bulat [1]) SwAV+ALIGN	90.55	1.059	89.65
SimCLR+ALIGN	91.72	0.995	91.08
(FaRL) ITC+MIM_{I}+ALIGN	**92.32**	**0.991**	**91.39**

Table 2. Comparison with self-supervised pre-training on face data.

6. More Details

Pre-training. During training, mixed-precision was used to accelerate training and save memory. Gradient checkpointing and ZeRO [15] are also used for further memory efficiency. Gradient clip with max norm of 1.0 is applied to stabilize the training process. Our implementation of image-text contrastive learning differs from CLIP [14], which computes contrastive loss using only the local batch on each GPU, our implementation gathers all logits from all GPUs and consider all of them in contrastive learning.

Computational Complexity. With the same image encoder structure (ViT-B), the computational complexity of our model is exactly the same with other pre-training methods during both downstream training and downstream inference. While during pre-training, our model has an extra text encoder and an additional MIM stage, leading to a generally doubled computation complexity comparing with CLIP; but we share comparable computation complexity with self-supervised contrastive learning methods (e.g. MoCo v3, SimCLR).

Face parsing. We adopt augmentations to face parsing tasks. On LaPa, we first compute a face alignment matrix that aligns five face landmarks retrieved from a face detector [5] to the landmarks of a mean face in a target resolution $s \in \{224, 448\}$. We then augment on the matrix with random rotation within $[-18°, 18°]$, random rescaling within $[0.9, 1.1]$ and random translation with a range of $0.01 \times s$. We transform both the image and the groundtruth label maps using the augmented matrix. The transformation is combined with the Tanh-warping [10] to ensure that the network can segment the whole face image as well as focusing on the face region. In order to better preserve linearity within face region, we modify the warping function of [10]

α	F1 ↑ Mean	Mean (w/o Hair)	Hair
0.0	91.51	91.93	87.74
0.2	92.08	92.04	92.52
0.4	92.27	92.07	94.09
0.6	92.31	92.07	**94.54**
0.8 (default)	**92.32**	**92.08**	94.53
1.0	92.11	91.84	94.49

Table 3. F1 scores on LaPa under different warping factors.

from $\tanh$ to $\tanh_\alpha$ defined as $\tanh_\alpha(x) =$

$$\begin{cases} x, & -1+\alpha \le x \le 1-\alpha \\ \alpha \tanh\left(\frac{x-1+\alpha}{\alpha}\right)+1-\alpha, & 1-\alpha < x \\ \alpha \tanh\left(\frac{x+1-\alpha}{\alpha}\right)-1+\alpha, & x < -1+\alpha \end{cases},$$

with α being a warping factor: $\tanh_{\alpha=1.0}$ equals to a vanilla $\tanh$ warping, while $\tanh_{\alpha\to 0.0}$ degenerates to a crop function that drops all peripheral pixels. Table 3 show results under different α. $\alpha = 0.8$ is selected as the default setting.

On CelebAMask-HQ, we replace the face alignment matrix to be a simple rescaling matrix that resizes the original size to $s \times s$, since all face images in CelebAMask-HQ are already aligned. We also disable Tanh-warping. The rest augmentations all remain the same with those on LaPa. All these above setups are also adopted for all other pre-trained models for fair comparison.

Face alignment. Augmentations are also applied to face alignment tasks. Random geometric transformations are first applied on the bounding boxes provided by the corresponding face alignment datasets. These transformations include random rotation within $[-10°, 10°]$, random rescaling within $[0.9, 1.1]$ and random translation with a range of $0.01 \times s$. The same transformations are imposed on the groundtruth 2D face landmarks as well. Then, we crop the original image with the transformed bounding boxes and rectify the groundtruth landmark coordinates accordingly. Finally, all these augmented groundtruth landmark points are rendered to 128×128 heatmaps for training, using an on-line approach. Random Gaussian blur, noise and occlusion are also used on the input images. These setups are also adopted for all other pre-trained models for fair comparison.

Face attributes recognition. Since our model is pretrained with aligned face images, it's important to also train the downstream task with aligned images. For CelebA dataset, we use the facial landmarks delivered with the dataset, for LFWA dataset, we do face detection with RetinaFace [5]. When training with backbone frozen, we randomly horizontal-flip the image with a probability of 0.5. When fully fine-tuning the model, apart from the random horizontal flip, random crop and Gaussian noise, we also apply random grayscale with a probability of 0.1, and impose Gaussian noise with a variance of 5 to the facial landmarks used for aligning the face. These setups are also adopted for all other pre-trained models for fair comparison.

7. Data Usage

LAION [19][1] contains 400M image-text pairs that are collected from Internet. It is licensed under Creative Common CC-BY 4.0. They don't claim copyright of the images.

LaPa [11][2] contains over 22K face images. Its license says "*this LaPa Dataset is made freely available to academic and non-academic entities for non-commercial purposes such as academic research, teaching, scientific publications, or personal experimentation. Permission is granted to use the data given that you agree to our license terms*".

CelebAMask-HQ [9][3] contains 30K face images. The usage of the dataset is restricted to non-commercial research and educational purposes.

AFLW-19 [25][4] contains around 24K face images. It is revised from the original AFLW dataset [8][5] without any claims on licensing. The original AFLW dataset is available for non-commercial research purposes only.

WFLW [22][6] contains images of about 10K faces. It does not mention any licenses.

300W [16–18][7] contains over 4K face images. The data are provided for research purposes only. Commercial use (*i.e.*, use in training commercial algorithms) is not allowed.

CelebA & LFWA [12][8]. CelebA has 202,599 face images while LFWA has 13,143. The CelebA dataset is available for non-commercial research purposes only. The LFWA dataset is based on the original LFW dataset [7][9].

8. Code Usage

ViT [6], **DeiT** [21]. We use the `timm` library[10] to load these two pre-trained Transformers. We load the ViT model with `vit_base_patch16_224_in21k`, and load DeiT with `deit_base_distilled_patch16_224`. The code of `timm` is licensed under Apache 2.0. Please refer to its website for the licensing of its pre-trained weights.

MoCo v3 [4][11]. We download its ViT-Base model and use the provided script to convert the weights to DeiT format, which is then loaded by the `timm` library. MoCo v3 is under the CC-BY-NC 4.0 license.

[1] https://laion.ai/laion-400-open-dataset/
[2] https://github.com/JDAI-CV/lapa-dataset
[3] https://github.com/switchablenorms/CelebAMask-HQ
[4] http://mmlab.ie.cuhk.edu.hk/projects/compositional.html
[5] https://www.tugraz.at/institute/icg/research/team-bischof/lrs/downloads/aflw/#license
[6] https://wywu.github.io/projects/LAB/WFLW.html
[7] https://ibug.doc.ic.ac.uk/resources/300-W/
[8] https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html
[9] http://vis-www.cs.umass.edu/lfw/
[10] https://github.com/rwightman/pytorch-image-models
[11] https://github.com/facebookresearch/moco-v3

BEiT [4][12] is under MIT License. Its BEiT-base is used.
CLIP [14][13] is under MIT license. Its ViT-B/16 is used.
FaceTransformer [24][14] does not contain a license. Its ViT-P8S8 is used.
RetinaFace [5][15] is used for face detection. It is under the MIT license.
MMSegmentation[16] is under the Apache 2.0 license. We use its UperNet [23] implementation for downstream tasks like face parsing and face alignment.

References

[1] Adrian Bulat, Shiyang Cheng, Jing Yang, Andrew Garbett, Enrique Sanchez, and Georgios Tzimiropoulos. Pre-training strategies and datasets for facial representation learning. *arXiv preprint arXiv:2103.16554*, 2021. 2

[2] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *arXiv preprint arXiv:2006.09882*, 2020. 2

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2

[4] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. *arXiv preprint arXiv:2104.02057*, 2021. 3, 4

[5] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5203–5212, 2020. 2, 3, 4

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[7] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007. 3

[8] Martin Koestinger, Paul Wohlhart, Peter M Roth, and Horst Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *2011 IEEE international conference on computer vision workshops (ICCV workshops)*, pages 2144–2151. IEEE, 2011. 3

[9] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5549–5558, 2020. 3

[10] Jinpeng Lin, Hao Yang, Dong Chen, Ming Zeng, Fang Wen, and Lu Yuan. Face parsing with roi tanh-warping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5654–5663, 2019. 2

[11] Yinglu Liu, Hailin Shi, Hao Shen, Yue Si, Xiaobo Wang, and Tao Mei. A new dataset and boundary-attention semantic segmentation for face parsing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11637–11644, 2020. 3

[12] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015. 3

[13] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2085–2094, 2021. 1

[14] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021. 2, 4

[15] Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–16. IEEE, 2020. 2

[16] Christos Sagonas, Epameinondas Antonakos, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: Database and results. *Image and vision computing*, 47:3–18, 2016. 3

[17] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 397–403, 2013. 3

[18] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. A semi-automatic methodology for facial landmark annotation. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 896–903, 2013. 3

[19] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 2, 3

[20] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 1

[21] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training

[12] https://github.com/microsoft/unilm/tree/master/beit
[13] https://github.com/openai/CLIP
[14] https://github.com/zhongyy/Face-Transformer
[15] https://github.com/biubug6/Pytorch_Retinaface
[16] https://github.com/open-mmlab/mmsegmentation

data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021. 3

[22] Wayne Wu, Chen Qian, Shuo Yang, Quan Wang, Yici Cai, and Qiang Zhou. Look at boundary: A boundary-aware face alignment algorithm. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2129–2138, 2018. 3

[23] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 418–434, 2018. 4

[24] Yaoyao Zhong and Weihong Deng. Face transformer for recognition. *arXiv preprint arXiv:2103.14803*, 2021. 4

[25] Shizhan Zhu, Cheng Li, Chen-Change Loy, and Xiaoou Tang. Unconstrained face alignment via cascaded compositional learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3409–3417, 2016. 3