

# General Facial Representation Learning in a Visual-Linguistic Manner

Yinglin Zheng<sup>1\*†</sup> Hao Yang<sup>2\*</sup> Ting Zhang<sup>2</sup> Jianmin Bao<sup>2</sup> Dongdong Chen<sup>3</sup>  
Yangyu Huang<sup>2</sup> Lu Yuan<sup>3</sup> Dong Chen<sup>2</sup> Ming Zeng<sup>1</sup> Fang Wen<sup>2</sup>

<sup>1</sup>School of Informatics, Xiamen University

<sup>2</sup>Microsoft Research Asia <sup>3</sup>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/far1>

## Abstract

*How to learn a **universal facial representation** that **boosts all face analysis tasks**? This paper takes one step toward this goal. In this paper, we study the transfer performance of pre-trained models on face analysis tasks and introduce a framework, called **FaRL**, for general facial representation learning. On one hand, the framework involves a contrastive loss to learn high-level semantic meaning from image-text pairs. On the other hand, we propose exploring low-level information simultaneously to further enhance the face representation by adding a masked image modeling. We perform pre-training on LAION-FACE, a dataset containing a large amount of face image-text pairs, and evaluate the representation capability on multiple downstream tasks. We show that FaRL achieves better transfer performance compared with previous pre-trained models. We also verify its superiority in the low-data regime. More importantly, our model surpasses the state-of-the-art methods on face analysis tasks including face parsing and face alignment.*

## 1. Introduction

Face analysis tasks are of crucial importance to social interactions and have received extensive attention over the past decades. Many existing state-of-the-art results [7, 51, 96] come from **deep neural networks** with **supervised learning**. However, such supervised models, in order to learn appropriate feature representations for each given task, are studied **separately with large-scale manually annotated data** which is **expensive** and **difficult to acquire**, especially for some face tasks such as face parsing and face alignment.

Recently, visual representation learning in computer vi-

sion appears to be paved by a popular learning paradigm, pre-training, due to the remarkable success of the ground-breaking models in Natural Language Processing such as BERT [24] and GPT-series [5, 76, 77], followed by a wide variety of multiple techniques [25, 45, 53, 73, 78]. Thereafter in vision, many attempts [13, 33, 36] along this path have been proposed, showing promising results using approaches related to contrastive loss [2, 10, 29, 35, 70].

Meanwhile in the area of vision-language tasks involving multi-modality, there are studies [44, 54, 55, 63, 75, 91] exploring learning directly from large, freely available image-text pairs up to hundreds of millions. Their results show that **natural language supervision is beneficial for visual representation learning**, bringing superior performance to most tasks on general images. Such pre-training has several advantages: 1) showing promising few-shot transfer performance, alleviating the hard-acquired labeled data issue; 2) enabling convenient deployment by extracting general feature representation once and then applying to diverse downstream tasks. Yet when it comes to the **face domain, one of the most important domains in computer vision, the effectiveness of pre-training is relatively unexplored**.

In this paper, we study the behaviors of transfer performance of pre-trained models on face analysis tasks and introduce a framework, **FaRL, to learn general facial representation in a visual-linguistic manner**. Instead of crawling images and texts from the Web by manually designed face-related queries, we create a dataset by filtering out from a large openly available **image-text-pair dataset [85]**, resulting in a subdataset containing 20 million face images, denoted as LAION-FACE.

We adopt the widely used contrastive loss to pull the embeddings of matched image-text pair together while pushing those of non-matched image-text pairs apart, which provides high-level semantic meaning. We propose exploring supplementary low-level information simultaneously to further enhance the face representation, by adding a masked



\*Equal contribution.

†Work done during internship at MSRA.

image modeling inspired from BEiT [3]. We perform pre-training on LAION-FACE and evaluate the representation capability with the frozen backbone, as our ultimate goal is to offer a general facial representation that can be quickly adapted to downstream tasks.

We measure the performance on several important downstream face tasks, including **face parsing, face alignment and face attribute prediction**, for which labeled data is usually difficult to acquire. We show that better transfer performance can be achieved compared with other pre-trained models. We also demonstrate its superiority in the low-data regime. Moreover, our model outperforms the state-of-the-art methods on face tasks including face parsing and face alignment.

In summary, our major contributions are as follows:

1. We present an extensive study about the transferable visual models learned in a visual-linguistic manner on versatile face analysis tasks, which is relatively unexplored in the literature.
2. We introduce a new framework, exploring low-level and high-level information simultaneously for better representation. We achieve superior transfer performance than previous pre-training approaches, and more importantly on face parsing and face alignment, our model has surpassed the state-of-the-art methods.

## 2. Related Works

### 2.1. Visual Representation Learning

Ever since the pioneering work [50] on ImageNet recognition followed by numerous improvements [38,40,89,106], ImageNet classification [19] has been the de facto pre-training task for visual representation learning. The backbones for a wide range of visual tasks including image classification, object detection, semantic segmentation and human pose estimation are often initialized from ImageNet pre-trained weights with the goal of less task-specific data and fewer training epochs. Supervised pre-training then becomes the predominant recipe in visual representation learning, with a scaling-up trend in the size and complexity of the network [48] as well as the size of the training dataset, *e.g.*, JFT-300M [92] and Instagram-1B [65].

As pre-training with large Transformer-based networks on large corpora [5,24,25,53,76,77] has pushed the state of the art forward in natural language processing, there emerge lots of inspired advancements in visual representation learning. The vision Transformer [28,99] applied a standard Transformer directly to image classification by splitting an image into patches, similar to the tokens in NLP. iGPT [12] defined the auto-regressive and BERT objective in the context of images by predicting pixels whose color is indexed

by k-means clustering. On the other hand, BEiT [3] designed a BERT-style loss based on the visual discrete tokens obtained by the discrete variational auto-encoder.

Subsequently, the remarkable success of natural language processing also flourished the visual-language pre-training, where large amounts of freely available image text pairs can be acquired on the Internet. Without relying on the pre-trained object detector to extract image region features as previous works [17,43,57,94,108,109], CLIP [75] and ALIGN [44] adopt the contrastive loss, an effective loss in self-supervised representation learning, to pull the embedding of matched image-text pairs together while pushing those of non-matched pairs apart, followed by ALBEF [55] which further improves the visual-language pre-training.

Another line of works [13–16,33,36] focus on learning visual representation without any supervision. Among the most successful of recent efforts, the critical core also relates to contrastive loss, measuring the similarities of augmented image pairs in a representation space.

### 2.2. Facial Representation Learning

In the area of face analysis, most tasks [7,42,51,58,96,97] are solved by supervised training with manually labeled data. Such supervised methods require a large number of training samples and may suffer from overfitting because of the massive number of model parameters. While pre-training has shown impressive performance on few-shot learning and also help reduce overfitting [39], the effectiveness of pre-training on face domain is rarely explored yet. Another major advantage of pre-training is that there will be a universal facial representation that can be well transferred to a variety of downstream tasks, which is particularly desirable for resource-limited mobile devices.

There exist several works addressing few-shot learning [4,88] and transfer learning [1,114] on a specific face task instead of examining pre-training over diverse face tasks. Closely related to our work is [6], which explored unsupervised pre-training for facial representation learning. In contrast, we bring clarity to an unexplored regime, weakly-supervised pre-training, on face domain as it has been shown in [75] that leveraging massive web image-text pairs, which provides weak supervision for images, is helpful for learning visual representations in few-shot scenarios.

## 3. FaRL

### 3.1. Visual Linguistic Face Data

Our goal is to **learn a transferable facial representation in a visual-linguistic manner**. For this purpose, we start by collecting a sufficiently large dataset that contains image-text pairs where the image includes a face region and the text label is natural language. Existing face datasets [34,41,52,59,61,82] are mainly designed for a spe-



Figure 1. Image-text pairs randomly sampled from LAION-FACE. Web texts are not always accurate but often easier to acquire.

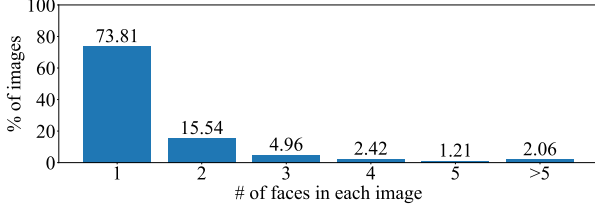


Figure 2. Distribution of #faces in each image in LAION-FACE.

cific face task. For example, the associated labels of corresponding datasets are identities, semantic masks, landmark locations, attribute tags for face recognition, face parsing, face alignment and attribute prediction respectively.

To enable learning from natural language supervision, which benefits from the large quantities of data available on the Internet, we construct a new dataset consisting of 20 million image-text pairs. Specifically, we leverage the large openly available image-text-pair dataset, LAION [85], that contains 400 million samples. To filter out those non-face images, we adopt a face detector, RetinaFace [23], to identify the presence of a face in an image. 20 million pairs are randomly sampled from those whose face detection scores are greater than 0.9. The resulting dataset is denoted as LAION-FACE. Figure 1 shows some image-text-pair samples from the dataset. The distribution of the number of faces in each image is shown in Figure 2.

### 3.2. Image-text Contrastive Learning

Following [44, 75], we adopt the image-text contrastive loss, which has been shown to be more compute-efficient than generative models [12] and learn better representation than the predictive counterpart [98]. The contrastive learning learns by comparing, according to some notion of similarity. Precisely, consider a given image-text pair  $\{T, I\}$ , the extracted feature representations are  $\{f_{cls}^I, f_1^I, \dots, f_N^I\} = E_I(I)$ ,  $\{f_{eos}^T, f_1^T, \dots, f_M^T\} = E_T(T)$ , where  $E_I$  denotes the image Transformer-based encoder and  $E_T$  denotes the text Transformer-based encoder.  $cls$  is short for class token,  $eos$  is short for end of sequence token and  $1, \dots, N(M)$  denotes the index of visual (language) tokens. The features from the  $cls$  ( $eos$ ) token are then fed into a projection head (a small MLP) to obtain the metric embeddings, i.e.  $e^I = P_I(f_{cls}^I)$ ,  $e^T = P_T(f_{eos}^T)$ . The contrastive loss, in the scenario of image-text pairs, is

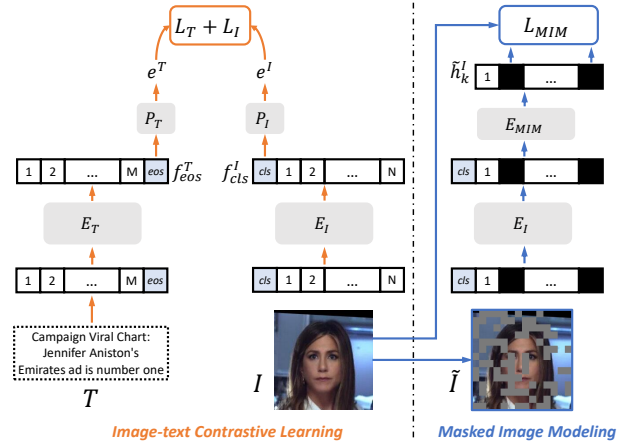


Figure 3. Illustrating our pre-training framework. We integrate masked image modeling with image-text contrastive learning. The two  $E_I$  in this figure stand for the same image encoder. After the pre-training, we use  $E_I$  to boost downstream face tasks.

given as

$$L_I = -\frac{1}{B} \sum_{i=1}^B \log \frac{\exp(e_i^I e_i^T / \sigma)}{\sum_{j=1}^B \exp(e_i^I e_j^T / \sigma)},$$

$$L_T = -\frac{1}{B} \sum_{i=1}^B \log \frac{\exp(e_i^T e_i^I / \sigma)}{\sum_{j=1}^B \exp(e_i^T e_j^I / \sigma)}, \quad (1)$$

where  $B$  is the number of image-text pairs in a mini-batch, and  $\sigma$  is the temperature to scale the logits, which is learned together with all other parameters.

### 3.3. Masked Image Modeling

It is intuitively plausible that the image-text contrastive learning facilitates to learn semantic feature representations from text about concrete or visualizable concepts. To further enhance the face representation, we add a masked image modeling task that masks some image patches in the input and predicts the visual tokens corresponding to the masked patches. This objective is similar to image inpainting aiming at filling in holes of an image, which is a representative low-level vision task. We hypothesize that this masked image modeling will help the features to capture low-level information, providing complementary information to high-level semantics.

Formally, let  $\tilde{I}$  be the masked image, where some image patches are randomly masked. That is to say, if a given input image  $I$  is split into  $N$  image patches  $\{I_1, \dots, I_N\}$ , the masked image  $\tilde{I}$  is also represented as  $N$  image patches  $\{\tilde{I}_1, \dots, \tilde{I}_N\}$  with

$$\tilde{I}_k = \begin{cases} I_k, & k \notin \mathcal{M} \\ m, & k \in \mathcal{M} \end{cases}, \quad (2)$$

where  $\mathcal{M} \subset \{1, \dots, N\}$  denotes the positions where image patches are masked, and  $m$  is the masked token, a learnable vector as same dimension as non-masked patches. After we get the features from the image encoder,  $\{\tilde{f}_{cls}^I, \tilde{f}_1^I, \dots, \tilde{f}_N^I\} = E_I(\tilde{I})$ , we feed them into a small Transformer which outputs the final hidden vectors,  $\{\tilde{h}_{cls}^I, \tilde{h}_1^I, \dots, \tilde{h}_N^I\} = E_{MIM}(\tilde{f}_{cls}^I, \tilde{f}_1^I, \dots, \tilde{f}_N^I)$ . The objective is to predict the masked region from the corresponding hidden vectors  $\{\tilde{h}_k^I, k \in \mathcal{M}\}$ . Instead of directly predicting pixels requiring huge memory consumption, the discrete variational autoencoder [71] is utilized here to first encode each image patch to one of  $|\mathcal{V}|$  possible values, with  $\mathcal{V}$  being the vocabulary of the autoencoder. Thereafter, a classification layer is attached on the hidden vector  $\tilde{h}_k^I$  to predict the corresponding masked patch's index among  $\{1, \dots, |\mathcal{V}|\}$ . The loss function is given as,

$$L_{MIM} = - \sum_{k \in \mathcal{M}} \log p\left(q_\phi^k(I) | \tilde{I}\right), \quad (3)$$

where  $p(q_\phi^k(I) | \tilde{I})$  denotes the classification score of classifying the  $k$ -th hidden vector belonging to the visual token  $q_\phi^k(I)$ , where  $q_\phi$  is the categorical distribution.

The whole framework is illustrated in Figure 3. In the experiment, we directly use the publicly available discrete variational autoencoder described in [80].

### 3.4. Pre-training Details

**Model architecture.** Our model consists of an image encoder  $E_I$ , a text encoder  $E_T$  and a masked image modeling module  $E_{MIM}$ . We implement the image encoder  $E_I$  following prior works [3, 28, 75] for fair comparison. Specifically, it is a 12-layer 768-width visual Transformer ViT-B/16 [28] with 87M parameters and  $224 \times 224$  input. The input image is first split into  $14 \times 14$  image patches, followed by a linear projection to obtain  $14 \times 14$  patch embeddings. A learnable *cls* token is prepended to these 196 embeddings, which is then added by 197 positional embeddings. We design the text encoder  $E_T$  following [75]. It is a 12-layer, 512-width and 8-head Transformer with 63M parameters. We fix the number of input text tokens to 77, and perform truncation or padding if the input length does not match. Finally we project features from both the image *cls* token and the text *eos* token to 512 dimension to calculate the contrastive loss. The masked image modeling module  $E_{MIM}$  is implemented as a one-layer Transformer for both simplicity and performance consideration.

**Pre-training setup.** We train our model from scratch with random initialized weights. We run pre-training for 16 epochs with batch-size 5120 on 32 Nvidia V100 GPUs. The AdamW [62] optimizer is adopted with a weight decay set as 0.05. The learning rate is initialized as  $1e-6$ , which is warmed up to  $1e-3$  in one epoch and then cosine decay to  $9e-4$  in 15 epochs. The softmax temperature  $\sigma$  is initialized

as 0.07. The input face image is aligned to the mean face as input. For images containing more than one face, we will choose one randomly. During pre-training, every input image will be fed into image encoder twice: one for image-text contrastive learning, one with randomly masked (at most 75 patches) image patches for masked image modeling.

### 3.5. Downstream Face Tasks

We adapt our model to multiple downstream face tasks that span various categories (segmentation, regression and classification) to evaluate its transfer performance:

**Face parsing** predicts pixel-wise regions of face components. Two popular datasets are used for this task: LaPa [59] and CelebAMask-HQ [52]. LaPa contains over 22K images, 18,176 for training and 2K for testing, each annotated with an 11-category pixel-level label map. CelebAMask-HQ consists of around 30K facial images, 24,183 for training and 2,824 for testing, each annotated with a 19-category label map including not only facial components but also body parts and accessories like eyeglasses, earrings and necklaces. Following [59, 96, 97], the F1 scores of facial components are used to measure the performance.

**Face alignment** aims to regress 2D face landmark coordinates on a face image. We use three popular datasets: AFLW-19 [113] with 20K images for training and 4,386 for testing, each annotated with 19 landmarks; 300W [82–84] with 3,837 images for training and 600 for testing, each annotated with 68 landmarks; WFLW [103] with 7,500 training data and 2,500 testing data, each annotated with 98 landmarks. Following common practice, we use normalized mean error (NME), failure rate (FR) and AUC as the metric.

**Face attributes recognition** predicts multiple attributes (e.g. gender, age and race) of a given face image, which can be viewed as a multi-label classification. Two datasets are adopted: CelebA [61] and LFW [61]. CelebA consists of over 202K facial images, while LFWA consists of 13,143 images, both with 40 attribute annotations per image. Following [61, 88], on CelebA, we use 162,770 for training and 19,962 for testing; on LFWA, we use 6,263 for training and the rest for testing. The averaged accuracy over all attributes is used as the metric.

## 4. Experiments

### 4.1. Setup

Different heads are designed for different downstream tasks. All the heads exploit features not only from the last layer of the visual Transformer  $E_I$ , but also from some intermediate layers of it. Let  $\{f_{cls,k}^I, f_{1,k}^I, f_{2,k}^I, \dots, f_{N,k}^I\}$  be the feature representations from the  $k$ -th layer, with  $1 \leq k \leq 12$ , since  $E_I$  consists of 12 layers in total.

We design a simple head for face attributes recognition.



Pre-training Settings				Downstream Performances		
Dataset	Scale	Supervision Source	Method	LaPa F1-mean $\uparrow$	AFLW-19 NME <sub>diag</sub> $\downarrow$	CelebA mAcc $\uparrow$
ImageNet-1K	1.3M	images	MoCo v3 [16]	91.86	1.007	90.23
ImageNet-22K	14M	images	BEiT [3]	91.29	1.076	89.71
ImageNet-22K	14M	images + human labels	ViT [28]	91.61	1.004	90.77
ImageNet-1K	1.3M	images + human labels	DeiT [99]	92.00	1.003	89.79
WIT	400M	images + web text	CLIP [75]	92.21	0.995	90.86
MS1MV3	5.1M	face images + human labels	Face Transformer [111]	91.09	1.031	90.77
LAION-FACE	20M	face images + web text	FaRL (Ours)	<b>92.32</b>	<b>0.991</b>	<b>91.39</b>

Table 1. Comparing the universal capability of pretrained vision transformers on multiple downstream face tasks, including face parsing (LaPa), face alignment (AFLW-19) and face attributes recognition (CelebA). All pre-trained backbones are frozen with only heads fine-tuned on downstream data.

Pre-trained Models	1%	10%	100%	Pre-trained Models	1%	10%	100%	Pre-trained Models	1%	10%	100%
MoCo v3 [16]	86.47	90.18	91.86	MoCo v3 [16]	1.41	1.19	1.007	MoCo v3 [16]	87.85	89.76	90.23
BEiT [3]	85.01	89.21	91.29	BEiT [3]	1.94	1.37	1.076	BEiT [3]	85.64	88.74	89.71
ViT [28]	86.64	89.97	91.61	ViT [28]	1.37	1.16	1.004	ViT [28]	89.20	90.21	90.77
DeiT [99]	87.24	90.45	92.00	DeiT [99]	1.41	1.17	1.003	DeiT [99]	86.73	89.00	89.79
CLIP [75]	<u>88.13</u>	<b>90.91</b>	<u>92.21</u>	CLIP [75]	<b>1.30</b>	<b>1.11</b>	<u>0.995</u>	CLIP [75]	89.09	<u>90.47</u>	<u>90.86</u>
FaceTransformer [111]	86.42	89.54	91.09	FaceTransformer [111]	1.41	1.21	1.031	FaceTransformer [111]	87.42	90.32	90.77
FaRL	<b>88.21</b>	<b>90.91</b>	<b>92.32</b>	FaRL	<u>1.35</u>	<u>1.15</u>	<b>0.991</b>	FaRL	<b>89.66</b>	<b>90.99</b>	<b>91.39</b>

(a) LaPa (F1-mean  $\uparrow$ ).

(b) AFLW-19 (NME<sub>diag</sub>  $\downarrow$ ).

(c) CelebA (mAcc  $\uparrow$ ).

Table 2. Comparing the generality of pretrained vision transformers w.r.t few-shot settings. We randomly sample subsets of the downstream training sets and evaluate the performances on the same test sets. All pre-trained backbones are frozen without fine-tuning on downstream data. Best performances are shown in bold and the second ones underlined.

Let  $\mathcal{K}$  be the set of selected layers for downstream. We compute three feature vectors from each layer  $k \in \mathcal{K}$ : the *cls* token feature  $f_{cls,k}^I$ , the mean of all non-*cls* token features and the global max-pooling of all non-*cls* token features. These  $3 \times |\mathcal{K}|$  vectors are then layer-normalized and linearly combined into one vector through learnable weights, appended with a fully connected layer to generate the logits for multi-label binary classification. The head is trained with binary-cross-entropy loss and AdamW [62] optimizer. We adopt a learning rate of 0.3 and cosinely decay to zero in 100 epochs.

Face parsing requires spatial distributions from the  $N$  non-*cls* tokens on each layer. Note these non-*cls* tokens corresponds to image patches, hence they can be reshaped to a 2D feature map of size  $\sqrt{N} \times \sqrt{N}$  ( $14 \times 14$  in our case). Following [3, 60], UperNet [105] is used to integrate multi-layer feature maps to generate a final feature map. A  $1 \times 1$  conv is appended to compute the parsing logits. The head is trained with cross-entropy loss, using AdamW with a learning rate  $1e-3$  and weight decay  $1e-5$ . Tanh-warping [58] is employed to balance the segmentation performance between the inner facial components and the hair region.

The face alignment head predicts heatmaps of the 2D landmark points, as is practiced by [42, 51, 101]. We render groundtruth landmark points as Gaussian heatmaps of size  $128 \times 128$  with a one-pixel  $\sigma$  and values  $\in [0, 1]$ . Instead of using those complex loss functions designed by [30, 42, 101], we simply train the head with a soft-label

cross-entropy loss. UperNet [105] is also employed to output the heatmap logits. We train the head using AdamW with learning rate 0.01 and weight decay  $1e-5$ .

For all downstream tasks, we follow [3] and set the selected layers  $\mathcal{K} = \{4, 6, 8, 12\}$ . The above setups are also adopted for other pre-trained models for fair comparison. Please refer to the appendix for more details.

## 4.2. Comparing with Pre-trained Transformers

We are curious about a question: *given one face image as input, is it possible that the feature output from our pre-trained model could be quickly adapted to benefit all downstream tasks?* To answer this question, we freeze our pre-trained model, extract features from input face images using the same frozen encoder  $E_I$ , and directly leverage the output features to facilitate downstream training.

We compare FaRL with other publicly available pre-trained Transformers. For fairness, all models share the same backbone structure (ViT-B/16)<sup>1</sup>. We ensure that they only differ in backbone weights, with all other settings (head structures, training hyper-parameters *etc.*) identical.

We compare with six pre-trainings: 1) MoCo v3 [16], a model pretrained on ImageNet-1K using image-wise contrastive learning; 2) BEiT [16], pretrained on ImageNet-22K with mask image modeling only; 3) ViT [28], the clas-

<sup>1</sup>Except Face Transformer, which has smaller patch size of 8 but a larger transformer layer number of 20. We set  $\mathcal{K} = \{6, 9, 13, 20\}$  for Face Transformer.

Pre-training Settings	LaPa	AFLW-19	CelebA
	F1-mean $\uparrow$	NME $_{diag}\downarrow$	mAcc $\uparrow$
ITC	91.75	1.009	91.31
ITC+MIM <sub>1</sub>	91.82	1.004	91.22
ITC+MIM <sub>1</sub> +ALIGN (FaRL)	<b>92.32</b>	<b>0.991</b>	<b>91.39</b>
ITC+MIM <sub>6</sub> +ALIGN	92.19	1.002	91.38
ITC+MIM <sub>6</sub>	91.99	0.992	91.20
ITC+ALIGN	91.88	1.012	<b>91.40</b>
ITC (LAION-RANDOM)	91.68	1.010	90.76

Table 3. Ablating the different components of FaRL pre-training w.r.t downstream task performances. *ITC+MIM<sub>1</sub>+ALIGN* is the default setting of FaRL.

sic vision transformer pretrained with large-scale human annotated labels from ImageNet-22K, in a fully supervised manner; 4) DeiT [99], an improved version of ViT that is pretrained on ImageNet-1K with distillation, also in a fully supervised manner; 5) CLIP [75], which is pretrained on 400M visual-linguistic data with image-text contrastive loss only; 6) Face Transformer [111], a transformer pretrained on a large face recognition dataset MS1MV3 [21, 22, 34] with human annotated face identity labels.

We illustrate their results on three downstream tasks in Table 1: LaPa for face parsing, AFLW-19 for face alignment and CelebA for face attributes recognition. FaRL exhibits consistent advantage over other pre-trainings on all tasks. The weakly supervised FaRL outperforms not only self-supervised pre-trainings (MoCo v3, BEiT) but also fully supervised pre-trainings (DeiT, ViT). FaRL also surpasses CLIP, a weakly-supervised model pre-trained with a significantly larger data scale (400M image-text pairs). Both pre-trained on face data, FaRL also excels the fully supervised Face Transformer by a large margin.

We also conduct evaluations under few-shot settings with limited downstream training data. Results reported in Table 2 show that FaRL achieves the best few-shot performances on both face parsing (LaPa) and face attributes recognition (CelebA). However, with 1% and 10% downstream training data, FaRL is outperformed by CLIP on face alignment (AFLW-19). We conjecture that the good generalizability of CLIP might come from their private 400M pre-training data: with an extremely large data scale (20 times the size of LAION-FACE), their dataset should involve a significant number of faces no less than LAION-FACE.

### 4.3. Ablating Pre-training Components

We evaluate the effectiveness of different training components through ablation experiments reported in Table 3, where ITC represents the *image text contrastive* learning, MIM stands for the *masked image modeling*. MIM<sub>1</sub> means to append an additional 1-layer Transformer to  $E_I$  for masked image modeling, while MIM<sub>6</sub> means to append 6 layers. ALIGN means we would follow Section 3.4 and randomly select one face from the original image, align and crop it to  $224 \times 224$  before feeding into the vision encoder. If without ALIGN, we would follow CLIP [75] instead and

Method	Skin	Hair	L-E	R-E	U-L	I-M	L-L	Nose	L-B	R-B	Mean
BASS [59]	97.2	96.3	88.1	88.0	84.4	87.6	85.7	95.5	87.7	87.6	89.8
EHANet [64]	95.8	94.3	87.0	89.1	85.3	85.6	88.8	94.3	85.9	86.1	89.2
Wei et al. [102]	96.1	95.1	88.9	87.5	83.1	89.2	83.8	96.1	86.0	87.8	89.4
EAGR [97]	97.3	96.2	89.5	90.0	88.1	90.0	89.0	97.1	86.5	87.0	91.1
AGRNet [96]	<b>97.7</b>	<b>96.5</b>	91.6	91.1	88.5	<b>90.7</b>	<b>90.1</b>	97.3	89.9	90.0	92.3
Scratch	97.18	93.06	91.61	91.50	87.22	89.44	89.13	97.26	90.12	89.69	91.62
FaRL	97.38	94.53	91.88	91.69	88.20	90.59	89.85	97.42	90.84	90.85	92.32
FaRL <sub>ft</sub>	97.52	95.11	<b>92.33</b>	<b>92.09</b>	88.69	90.70	90.05	<b>97.55</b>	<b>91.57</b>	<b>91.34</b>	<b>92.70</b>
FaRL <sub>ft</sub> <sup>448</sup>	<b>98.00</b>	<b>96.52</b>	<b>93.97</b>	<b>93.91</b>	<b>90.15</b>	<b>91.74</b>	<b>91.21</b>	<b>97.92</b>	<b>92.70</b>	<b>92.65</b>	<b>93.88</b>

Table 4. Comparing with state-of-the-art face parsing methods on LaPa test set. Results are reported in F1 scores (%).

do random crop to get the  $224 \times 224$  input.

Compared to ITC only, adding MIM<sub>1</sub> improves performances on face parsing (LaPa) and face alignment (AFLW-19), but not on face attributes recognition (CelebA). This supports our hypothesis that MIM helps capturing more low-level information, thus it is more beneficial for downstream tasks depending on relatively low-level features. Although ALIGN is the most critical component for CelebA, adding both MIM<sub>1</sub> and ALIGN together achieves generally better scores on most downstream benchmarks (LaPa, AFLW-10). In addition, a heavier MIM<sub>6</sub> head is not as good as MIM<sub>1</sub>, suggesting that the deeper head may weaken the effect of the MIM loss. Therefore, we select ITC+MIM<sub>1</sub>+ALIGN as the default setting of FaRL.

We also append a result of an ITC-only model pre-trained on a dataset called LAION-RANDOM in Table 3. LAION-RANDOM has the same size with LAION-FACE but its data is randomly sampled from LAION, thus containing lots of non-face images. The model pre-trained on LAION-FACE is consistently better than the model pre-trained on LAION-RANDOM, showing the importance of face data in pre-training. The ratio of face images is more crucial for face attributes recognition (CelebA) but relatively less important for face parsing (LaPa) and face alignment (AFLW-19). It might be because face attributes recognition require reasoning on higher-level semantics. To implicitly acquire the high-level knowledge related with facial attributes and identity, the model needs to see a lot more face images in pre-training.

### 4.4. Comparing with State-of-the-Art Face Methods

In this section, we compare FaRL with the state-of-the-art methods in multiple downstream face tasks. Different variants of FaRL are also compared. We use the name FaRL to represent our vanilla FaRL model whose pre-trained backbone is always frozen. FaRL<sub>ft</sub> denotes the model that is fully fine-tuned from the vanilla FaRL for the specific downstream task. While both FaRL and FaRL<sub>ft</sub> accepts  $224 \times 224$  inputs, we also fine-tune a model that accepts two-times larger input resolution, namely FaRL<sub>ft</sub><sup>448</sup>. FaRL<sub>ft</sub><sup>448</sup> shares the same initial parameters with FaRL<sub>ft</sub>, but its positional embeddings are initialized by a bi-cubic up-sampling from the positional embeddings of FaRL<sub>ft</sub>. In or-

Method	Face I-M	Nose U-L	Glasses L-L	L-Eye Hair	R-Eye Hat	L-B Earring	R-B Necklace	L-Ear Neck	R-Ear Cloth	Mean
Lee <i>et al.</i> [110]	95.5	85.6	92.9	84.3	85.2	81.4	81.2	84.9	83.1	80.3
EHANet [64]	96.0	93.7	90.6	86.2	86.5	83.2	83.1	86.5	84.1	84.0
Wei <i>et al.</i> [102]	93.8	88.6	90.3	93.9	85.9	67.8	30.1	88.8	83.5	82.1
EAGR [97]	96.4	91.9	89.5	87.1	85.0	80.8	82.5	84.1	83.3	85.1
AGRNet [96]	90.6	87.9	91.0	91.1	83.9	65.4	17.8	88.1	80.6	85.5
Scratch	96.2	94.0	92.3	88.6	88.7	85.7	85.2	88.0	85.7	84.74
FaRL	92.0	89.1	91.1	95.2	87.2	69.6	32.8	89.9	84.9	86.72
FaRL <sub>ft</sub>	96.17	93.77	92.28	89.04	88.97	85.32	85.36	86.88	87.32	87.55
FaRL <sub>ft</sub> <sup>448</sup>	91.66	88.10	90.04	94.94	82.73	63.05	33.52	90.76	85.92	89.56

Table 5. Comparing with state-of-the-art face parsing methods on CelebAMask-HQ test set. Results are reported in F1 scores (%).

der to investigate whether the gains are brought by our pre-training method or just come from the Transformer-based network structure, we also append a model named *Scratch*. It stands for a model which shares the same network structure with FaRL, but is specifically fully trained on the corresponding downstream task from scratch.

**Face parsing.** As illustrated in Table 4 and Table 5, our methods achieve remarkable performances on LaPa and CelebAMask-HQ. The vanilla FaRL surpasses the prior arts on both benchmarks. The refined FaRL<sub>ft</sub> brings even higher F1 scores. The ultimate performance is achieved by FaRL<sub>ft</sub><sup>448</sup>, which outperforms the state-of-the-art method [96] by 1.58 and 4.06 on LaPa and CelebAMask-HQ, respectively. We note that the input resolution plays a critical role in face parsing performance, it is especially effective for small components (*e.g.* *necklace* in CelebAMask-HQ). Even though, the required resolution of FaRL<sub>ft</sub><sup>448</sup> is lower than the resolution needed by the state-of-the-art approach [96] which is 473. It is also worth noting that the backbone-frozen FaRL achieves even better performances than the *Scratch* models on both benchmarks, showing that the representation learned from FaRL is not only widely applicable, but also sufficiently effective.

**Face alignment.** We compare with previous face alignment methods on three benchmarks: AFLW-19, 300W and WFLW, and report results in Table 6, Table 8 and Table 7, respectively. The Transformer-based methods achieves performances superior to all prior arts on AFLW-19, as shown by Table 6. Among these Transformer-based methods, the vanilla FaRL consistently outperforms *Scratch*, and our FaRL<sub>ft</sub><sup>448</sup> achieves new state-of-the-art performances on both AFLW-19 and WFLW, while being comparable with [42] on 300W. But unlike [42], our method does not assume any co-boundary relationship among landmark points. In addition, our methods outperform a previous work [6], which also leverages unified face representation

Method	NME <sub>diag</sub> ↓		NME <sub>box</sub> ↓		AUC <sub>box</sub> <sup>7</sup> ↑
	Full	Frontal	Full	Full	
CFSS [112]	3.92	2.68	-	-	-
CCL [113]	2.72	2.17	-	-	-
DAC-CSR [31]	2.27	1.81	-	-	-
LLL [81]	1.97	-	-	-	-
SAN [26]	1.91	1.85	4.04	54.0	-
DSRN [67]	1.86	-	-	-	-
LAB (w/o B) [103]	1.85	1.62	-	-	-
LAB (w/ B) [103]	1.25	1.14	-	-	-
HR-Net [93]	1.57	1.46	-	-	-
Wing [30]	-	-	3.56	53.5	-
KDN [11]	-	-	2.80	60.3	-
LUVLi [51]	1.39	1.19	2.28	68.0	-
Bulat <i>et al.</i> [6]	1.54	-	-	-	-
Scratch	1.047	0.884	1.481	79.3	-
FaRL	0.991	0.851	1.402	80.4	-
FaRL <sub>ft</sub>	0.969	0.836	1.371	80.8	-
FaRL <sub>ft</sub> <sup>448</sup>	<b>0.943</b>	<b>0.821</b>	<b>1.334</b>	<b>81.3</b>	-

Table 6. Comparing with state-of-the-art face alignment methods on two AFLW-19 test sets: the Full set and the Frontal subset.

Method	NME <sub>inter-ocular</sub> ↓							FR <sup>10</sup> ↓	AUC <sup>10</sup> ↑
	Full	Pose	Expr.	Illum.	MakeUp	Occl.	Blur		
ESR [8]	11.13	25.88	11.47	10.49	11.05	13.75	12.20	35.24	27.74
SDM [107]	10.29	24.10	11.45	9.32	9.38	13.03	11.28	29.40	30.02
CFSS [112]	9.07	21.36	10.09	8.30	8.74	11.76	9.96	20.56	36.59
DVLN [104]	6.08	11.54	6.78	5.73	5.98	7.33	6.88	10.84	45.51
LAB [103]	5.27	10.24	5.51	5.23	5.15	6.79	6.12	7.56	53.23
Wing [30]	5.11	8.75	5.36	4.93	5.41	6.37	5.81	6.00	55.04
DeCaFa [18]	4.62	8.11	4.65	4.41	4.63	5.74	5.38	4.84	56.30
AWing [101]	4.36	7.38	4.58	4.32	4.27	5.19	4.96	2.84	57.19
LUVLi [51]	4.37	-	-	-	-	-	-	3.12	57.70
ADNet [42]	4.14	6.96	4.38	4.09	4.05	5.06	4.79	2.72	60.22
Bulat <i>et al.</i> [6]	4.57	-	-	-	-	-	-	-	-
Scratch	4.80	8.78	5.09	4.74	4.99	6.01	5.35	5.72	54.54
FaRL	4.38	7.60	4.66	4.19	4.30	5.44	4.98	3.32	57.54
FaRL <sub>ft</sub>	<b>4.03</b>	<b>6.81</b>	<b>4.32</b>	<b>3.92</b>	<b>3.87</b>	<b>4.70</b>	<b>4.54</b>	<b>1.76</b>	<b>60.23</b>
FaRL <sub>ft</sub> <sup>448</sup>	<b>3.96</b>	<b>6.91</b>	<b>4.21</b>	<b>3.97</b>	<b>3.80</b>	<b>4.71</b>	<b>4.57</b>	<b>1.76</b>	<b>61.16</b>

Table 7. Comparing with state-of-the-art face alignment methods on WFLW test sets.

Method	NME <sub>inter-ocular</sub> ↓		Full
	Common	Challenge	
SAN [26]	3.34	6.60	3.98
AVS [74]	3.21	6.49	3.86
DAN [49]	3.19	5.24	3.59
LAB (w/ B) [103]	2.98	5.19	3.49
DCFE (w/ 3D) [100]	2.76	5.22	3.24
Teacher [27]	2.91	5.91	3.49
DU-Net [95]	2.97	5.53	3.47
DeCaFa [18]	2.93	5.26	3.39
HR-Net [93]	2.87	5.15	3.32
HG-HSLE [115]	2.85	5.03	3.28
AWing [101]	2.72	4.52	3.07
LUVLi [51]	2.76	5.16	3.23
ADNet [42]	<b>2.53</b>	<b>4.58</b>	<b>2.93</b>
Bulat <i>et al.</i> [6]	-	-	3.85
Scratch	2.90	5.19	3.35
FaRL	2.69	4.85	3.12
FaRL <sub>ft</sub>	2.70	4.64	3.08
FaRL <sub>ft</sub> <sup>448</sup>	<b>2.56</b>	<b>4.45</b>	<b>2.93</b>

Table 8. Comparing with state-of-the-art face alignment methods on three 300W test sets: the Common subset, the Challenge subset and the Full set.

pre-training, by a large margin.

**Face attributes recognition.** We compare with previous methods under both full-shot and few-shot settings on two benchmarks: CelebA and LFWA. As shown in Table 9, our method outperforms others reported under all few-shot settings, while ranking the second under full-shot. A possible reason is that the state-of-the-art method [7] explicitly

	CelebA					LFWA				
Proportion	0.2%	0.5%	1%	2%	100%	5%	10%	20%	50%	100%
# of training data	325	843	1,627	3,255	162,770	313	626	1,252	3,131	6,263
DMM [66]	-	-	-	-	91.70	-	-	-	-	86.56
SlimCNN [87]	79.90	80.20	80.96	82.32	91.24	70.90	71.49	72.12	73.45	76.02
AFFAIR [56]	-	-	-	-	91.45	-	-	-	-	86.13
PS-MCNN [7]	-	-	-	-	<b>92.98</b>	-	-	-	-	<b>87.36</b>
He <i>et al.</i> [37]	-	-	-	-	91.81	-	-	-	-	85.20
DeepCluster [9]	83.21	86.13	87.46	88.86	91.68	74.21	77.42	80.77	84.27	85.90
JigsawPuzzle [69]	82.88	84.71	86.25	87.77	91.57	73.90	77.01	79.56	83.29	84.86
Rot [32]	83.25	86.51	87.67	88.82	91.69	74.40	76.67	81.52	84.90	85.72
FixMath [90]	80.22	84.19	85.77	86.14	89.78	71.42	72.78	75.10	80.87	83.84
VAT [68]	81.44	84.02	86.30	87.28	91.44	72.19	74.42	76.26	80.55	84.68
SSPL [88]	86.67	88.05	88.84	89.58	91.77	78.68	81.65	83.45	85.43	86.53
FaRL	87.63	88.58	89.66	90.26	91.39	80.94	82.42	83.94	85.38	85.94
FaRL <sub>ff</sub>	<b>88.51</b>	<b>89.12</b>	<b>90.24</b>	<b>90.55</b>	<b>91.88</b>	<b>82.57</b>	<b>83.58</b>	<b>84.80</b>	<b>85.95</b>	<b>86.69</b>

Table 9. Comparing with other face attributes recognition methods under both full-shot and few-shot protocols on CelebA and LFWA. Results are reported in mean accuracy (%).

leverages extra information from downstream data, including face identity annotations and hand-designed attribute relationships.

## 5. Discussions

In this section, we discuss the pros and cons of our model. While some of these have been analyzed in previous various sections, we summarize and collect them here.

**The benefits of our model.** First and most important, our model provides a general facial representation learning framework showing excellent performance on downstream tasks. In this way, a face image can be fed into a general feature extraction module once and for all, then the feature is used in different decision modules for different respective face tasks. This is particularly useful for resource-restricted mobile devices. Second, our model is applicable for low-level face tasks as well as high-level face tasks, due to the joint learning of the contrastive learning capturing the semantics and the masked image modeling harnessing the low-level information. Last, we show that the framework achieves satisfactory results based on 20M samples, much fewer compared with general image representation pre-training [75] utilizing hundreds of millions of samples. This will facilitate its adoption.

**Ethical considerations.** Our model relies on the large-scale image-text data which is usually crawled from the Internet with a huge amount of data publicly available in such form. While crawling face images seems to be another matter. It may suffer from social biases as it would be unlikely to fully cover all different aspects of face-related queries when crawling. Although we choose a roundabout way of filtering out face images from a general image-text pair dataset, it is still possible that bias exists in the extracted subdataset.

Thereupon we provide preliminary analysis on the face image dataset FairFace [46], FairFace is a face attribute dataset for balanced race, gender, and age. It categorizes gender into 2 groups, race into 7 groups, age into 9 groups.

	Model	White	Non-White	Average	Discrepancy
Age	FairFace [46]	60.05	60.63	60.52	+0.58
	CLIP [75]	62.25	61.95	62.00	-0.30
	FaRL	61.49	61.84	61.78	+0.35
Gender	FairFace [46]	94.15	94.41	94.36	+0.26
	CLIP [75]	94.87	95.78	95.61	+0.91
	FaRL	95.16	95.77	95.65	+0.61

Table 10. Age and gender classification accuracies on FairFace. Results are reported w.r.t two race groups.

We freeze the pretrained backbones and only fine-tune classification heads on FairFace. We compare FaRL with the 400M-data-pretrained CLIP [75]<sup>2</sup>, as well as the baseline model associated with FairFace [46]. The accuracies w.r.t race groups are reported in Table 10. We follow [75] and define the group “Non-White” to include multiple race categories: “Black”, “Indian”, “East Asian”, “Southeast Asian”, “Middle Eastern” and “Latino”.

We are interested in the performance gaps between different race groups, which might associate with social bias. As shown by the right-most column of Table 10, though FaRL exhibits gaps between different race groups, the gap values of FaRL are relatively moderate among others.

**Limitations.** We are aware that there are still limitations of our model. First, as discussed above, our model presents bias to a certain degree, which might be caused by: 1) the data bias existing in the original LAION dataset [85], or 2) the performance bias of the face detector [20] we use. Second, our current work has not yet adapted to some important face tasks, *e.g.*, face detection, face anti-spoofing and face forgery detection. We expect our future updates would address these issues.

## 6. Conclusion

In this paper, we investigate the transfer performance of pre-trained models on face analysis tasks. We design a **pre-training method, called FaRL**, that leverages image-text contrastive learning as well as masked image modeling, to learn more general facial representation. We show that the face representation learned by FaRL transfers well to downstream face analysis tasks, including **face parsing**, face alignment, and face attributes recognition. Compared with previous pre-trained models, our model FaRL achieves superior transfer performance. Moreover, the proposed model surpasses the state-of-the-art methods on face parsing and face alignment.

**Acknowledgements** Yinglin Zheng and Ming Zeng were partially supported by NSFC (No. 62072382), Fundamental Research Funds for Central Universities, China (No. 20720190003).

<sup>2</sup>We compare with CLIP-ViT-B/16 under the same fine-tuning protocol for fairness.



## A. Evaluation on Face Editing Tasks.

Here we adopt a recent text-driven face editing framework [72], 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.

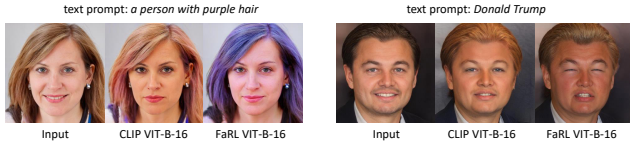


Figure 4. Comparing FaRL with CLIP in text-driven face editing.

## B. Visualizing the Pre-trained Image Encoder

In Figure 5, we provide the Grad-CAM [86] 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.

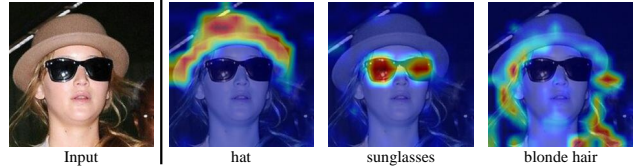
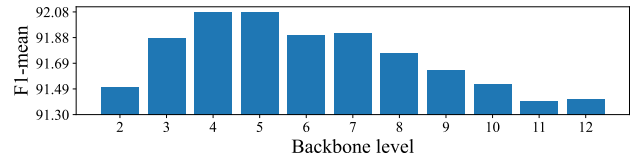
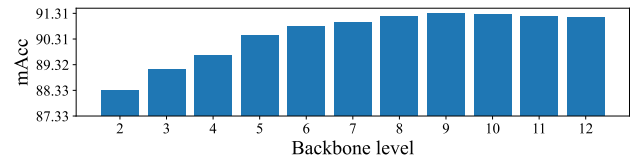


Figure 5. Grad-CAM visualizations of  $E_I$  given different text queries. Gradients are calculated in output of the first LayerNorm within the last Transformer block of  $E_I$ .



(a) LaPa (face parsing).



(b) CelebA (face attributes recognition).

Figure 6. Performance of features from different backbone levels on face parsing (LaPa) and face attributes recognition (CelebA).

## C. 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 6a and Figure 6b 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.

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.

## D. 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 [85] 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 11. 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 advan-

% of Face Images	LaPa F1-mean $\uparrow$	AFLW-19 NME $_{\text{diag}} \downarrow$	CelebA mAcc $\uparrow$
0	91.68	1.017	89.73
12.5	91.68	1.010	90.76
50	<b>91.77</b>	<b>1.009</b>	91.17
100	91.75	<b>1.009</b>	<b>91.31</b>

Table 11. Downstream performances w.r.t different face image ratios in pre-training data. Here the pre-training only uses image-text contrastive learning.

tages 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.

## E. Comparison with Self-supervised Methods on Face Dataset

Here we compare FaRL with two recent self-supervised methods: SwAV [10] and SimCLR [13] on LAION-Face with the same network structure and fine-tuning strategy with FaRL. Note that SwAV on face images is equivalent to Bulat [6]. as it adopts SwAV for face pre-training. As shown in Table 12, FaRL achieves better performances on all tasks.

Pre-training Settings	LaPa F1-mean $\uparrow$	AFLW-19 NME $_{\text{diag}} \downarrow$	CelebA mAcc $\uparrow$
(Bulat [6]) SwAV+ALIGN	90.55	1.059	89.65
SimCLR+ALIGN	91.72	0.995	91.08
(FaRL) ITC+MIM $_1$ +ALIGN	<b>92.32</b>	<b>0.991</b>	<b>91.39</b>

Table 12. Comparison with self-supervised pre-training on face data.

## F. Longer Pre-training for Finetuning

We illustrate finetuning results from longer FaRL pre-training in Table 13. Additional benefits can be observed from longer pre-training.

## G. More Details

**Pre-training.** During training, mixed-precision was used to accelerate training and save memory. Gradient checkpointing and ZeRO [79] 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 [75], 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 [20] 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  $[-18^\circ, 18^\circ]$ , 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 [58] 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 [58] from  $\tanh$  to  $\tanh_\alpha$  defined as  $\tanh_\alpha(x) =$

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

with  $\alpha$  being a warping factor:  $\tanh_{\alpha=1.0}$  equals to a vanilla  $\tanh$  warping, while  $\tanh_{\alpha \rightarrow 0.0}$  degenerates to a crop function that drops all peripheral pixels. Table 14 show results under different  $\alpha$ .  $\alpha = 0.8$  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  $[-10^\circ, 10^\circ]$ , 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  $128 \times 128$  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

	Pre-training Epoches	LaPa F1-mean $\uparrow$	CelebAMask-HQ F1-mean $\uparrow$	AFLW-19 NME <sub>diag</sub> $\downarrow$	WFLW NME <sub>inter-ocular</sub> $\downarrow$	300W NME <sub>inter-ocular</sub> $\downarrow$
FaRL <sup>448</sup> <sub>ft</sub>	16	93.88	89.56	0.943	3.96	2.93
	64	<b>94.04</b>	<b>89.57</b>	<b>0.938</b>	<b>3.88</b>	<b>2.88</b>

Table 13. Finetuning results of longer FaRL pre-training.

$\alpha$	Mean	F1 $\uparrow$ 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	<b>94.54</b>
0.8 (default)	<b>92.32</b>	<b>92.08</b>	94.53
1.0	92.11	91.84	94.49

Table 14. F1 scores on LaPa under different warping factors.

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 [20]. 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.

## H. Data Usage

**LAION** [85]<sup>3</sup> 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** [59]<sup>4</sup> 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** [52]<sup>5</sup> contains 30K face images. The usage of the dataset is restricted to non-commercial research and educational purposes.

**AFLW-19** [113]<sup>6</sup> contains around 24K face images. It is revised from the original AFLW dataset [47]<sup>7</sup> without any claims on licensing. The original AFLW dataset is available for non-commercial research purposes only.

<sup>3</sup><https://laion.ai/laion-400-open-dataset/>

<sup>4</sup><https://github.com/JDAI-CV/lapa-dataset>

<sup>5</sup><https://github.com/switchablenorms/CelebAMask-HQ>

<sup>6</sup><http://mmlab.ie.cuhk.edu.hk/projects/compositional.html>

<sup>7</sup><https://www.tugraz.at/institute/icg/research/team-bischof/lrs/downloads/aflw/#license>

**WFLW** [103]<sup>8</sup> contains images of about 10K faces. It does not mention any licenses.

**300W** [82–84]<sup>9</sup> 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** [61]<sup>10</sup>. 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 [41]<sup>11</sup>.

## I. Code Usage

**ViT** [28], **DeiT** [99]. We use the `timm` library<sup>12</sup> 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** [16]<sup>13</sup>. 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.

**BEiT** [16]<sup>14</sup> is under MIT License. Its **BEiT-base** is used.

**CLIP** [75]<sup>15</sup> is under MIT license. Its ViT-B/16 is used.

**FaceTransformer** [111]<sup>16</sup> does not contain a license. Its **ViT-P8S8** is used.

**RetinaFace** [20]<sup>17</sup> is used for face detection. It is under the MIT license.

**MMSegmentation**<sup>18</sup> is under the Apache 2.0 license. We use its UperNet [105] implementation for downstream tasks like face parsing and face alignment.

<sup>8</sup><https://wywu.github.io/projects/LAB/WFLW.html>

<sup>9</sup><https://ibug.doc.ic.ac.uk/resources/300-W/>

<sup>10</sup><https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html>

<sup>11</sup><http://vis-www.cs.umass.edu/lfw/>

<sup>12</sup><https://github.com/rwightman/pytorch-image-models>

<sup>13</sup><https://github.com/facebookresearch/moco-v3>

<sup>14</sup><https://github.com/microsoft/unilm/tree/master/beit>

<sup>15</sup><https://github.com/openai/CLIP>

<sup>16</sup><https://github.com/zhonggy/Face-Transformer>

<sup>17</sup>[https://github.com/biubug6/Pytorch\\_Retinaface](https://github.com/biubug6/Pytorch_Retinaface)

<sup>18</sup><https://github.com/open-mmlab/mmdetection>

## References

- [1] Shivangi Aneja and Matthias Nießner. Generalized zero and few-shot transfer for facial forgery detection. *arXiv preprint arXiv:2006.11863*, 2020. [2](#)
- [2] Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. *arXiv preprint arXiv:1906.00910*, 2019. [1](#)
- [3] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. [2](#), [4](#), [5](#)
- [4] Bjorn Browatzki and Christian Wallraven. 3fabrec: Fast few-shot face alignment by reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6110–6120, 2020. [2](#)
- [5] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020. [1](#), [2](#)
- [6] Adrian Bulat, Shiyang Cheng, Jing Yang, Andrew Gabbett, Enrique Sanchez, and Georgios Tzimiropoulos. Pre-training strategies and datasets for facial representation learning. *arXiv preprint arXiv:2103.16554*, 2021. [2](#), [7](#), [10](#)
- [7] Jiajiong Cao, Yingming Li, and Zhongfei Zhang. Partially shared multi-task convolutional neural network with local constraint for face attribute learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4290–4299, 2018. [1](#), [2](#), [7](#), [8](#)
- [8] Xudong Cao, Yichen Wei, Fang Wen, and Jian Sun. Face alignment by explicit shape regression. *International journal of computer vision*, 107(2):177–190, 2014. [7](#)
- [9] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 132–149, 2018. [8](#)
- [10] 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. [1](#), [10](#)
- [11] Lisha Chen, Hui Su, and Qiang Ji. Face alignment with kernel density deep neural network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6992–7002, 2019. [7](#)
- [12] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pre-training from pixels. In *International Conference on Machine Learning*, pages 1691–1703. PMLR, 2020. [2](#), [3](#)
- [13] 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. [1](#), [2](#), [10](#)
- [14] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. *arXiv preprint arXiv:2006.10029*, 2020. [2](#)
- [15] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. [2](#)
- [16] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. *arXiv preprint arXiv:2104.02057*, 2021. [2](#), [5](#), [11](#)
- [17] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *ECCV*, 2020. [2](#)
- [18] Arnaud Dapogny, Kevin Bailly, and Matthieu Cord. Decafa: Deep convolutional cascade for face alignment in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6893–6901, 2019. [7](#)
- [19] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. [2](#)
- [20] 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. [8](#), [10](#), [11](#)
- [21] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019. [6](#)
- [22] Jiankang Deng, Jia Guo, Debing Zhang, Yafeng Deng, Xiangju Lu, and Song Shi. Lightweight face recognition challenge. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. [6](#)
- [23] Jiankang Deng, Jia Guo, Yuxiang Zhou, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-stage dense face localisation in the wild. *arXiv preprint arXiv:1905.00641*, 2019. [3](#)
- [24] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. [1](#), [2](#)
- [25] Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. Unified language model pre-training for natural language understanding and generation. *arXiv preprint arXiv:1905.03197*, 2019. [2](#)
- [26] Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi Yang. Style aggregated network for facial landmark detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 379–388, 2018. [7](#)
- [27] Xuanyi Dong and Yi Yang. Teacher supervises students how to learn from partially labeled images for facial landmark detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 783–792, 2019. [7](#)
- [28] 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. 2, 4, 5, 11
- [29] Alexey Dosovitskiy, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with convolutional neural networks. *Advances in neural information processing systems*, 27:766–774, 2014. 1
- [30] Zhen-Hua Feng, Josef Kittler, Muhammad Awais, Patrik Huber, and Xiao-Jun Wu. Wing loss for robust facial landmark localisation with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2235–2245, 2018. 5, 7
- [31] Zhen-Hua Feng, Josef Kittler, William Christmas, Patrik Huber, and Xiao-Jun Wu. Dynamic attention-controlled cascaded shape regression exploiting training data augmentation and fuzzy-set sample weighting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2481–2490, 2017. 7
- [32] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018. 8
- [33] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020. 1, 2
- [34] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European conference on computer vision*, pages 87–102. Springer, 2016. 2, 6
- [35] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE, 2006. 1
- [36] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 1, 2
- [37] Keke He, Yanwei Fu, Wuhao Zhang, Chengjie Wang, Yugang Jiang, Feiyue Huang, and Xiangyang Xue. Harnessing synthesized abstraction images to improve facial attribute recognition. In *IJCAI*, pages 733–740, 2018. 8
- [38] 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. 2
- [39] Dan Hendrycks, Kimin Lee, and Mantas Mazeika. Using pre-training can improve model robustness and uncertainty. In *International Conference on Machine Learning*, pages 2712–2721. PMLR, 2019. 2
- [40] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 2
- [41] 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. 2, 11
- [42] Yangyu Huang, Hao Yang, Chong Li, Jongyoo Kim, and Fangyun Wei. Adnet: Leveraging error-bias towards normal direction in face alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3080–3090, 2021. 2, 5, 7
- [43] Zhicheng Huang, Zhaoyang Zeng, Yupan Huang, Bei Liu, Dongmei Fu, and Jianlong Fu. Seeing out of the box: End-to-end pre-training for vision-language representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12976–12985, 2021. 2
- [44] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. *arXiv preprint arXiv:2102.05918*, 2021. 1, 2, 3
- [45] Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77, 2020. 1
- [46] Kimmo Karkkainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1548–1558, 2021. 8
- [47] 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. 11
- [48] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 491–507. Springer, 2020. 2
- [49] Marek Kowalski, Jacek Naruniec, and Tomasz Trzcinski. Deep alignment network: A convolutional neural network for robust face alignment. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 88–97, 2017. 7
- [50] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012. 2
- [51] Abhinav Kumar, Tim K Marks, Wenxuan Mou, Ye Wang, Michael Jones, Anoop Cherian, Toshiaki Koike-Akino, Xiaoming Liu, and Chen Feng. Luvli face alignment: Esti-

- ating landmarks’ location, uncertainty, and visibility likelihood. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8236–8246, 2020. [1](#), [2](#), [5](#), [7](#)
- [52] 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. [2](#), [4](#), [11](#)
- [53] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019. [1](#), [2](#)
- [54] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11336–11344, 2020. [1](#)
- [55] Junnan Li, Ramprasaath R Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *arXiv preprint arXiv:2107.07651*, 2021. [1](#), [2](#)
- [56] Jianshu Li, Fang Zhao, Jiashi Feng, Sujoy Roy, Shuicheng Yan, and Terence Sim. Landmark free face attribute prediction. *IEEE Transactions on Image Processing*, 27(9):4651–4662, 2018. [8](#)
- [57] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020. [2](#)
- [58] 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](#), [5](#), [10](#)
- [59] 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. [2](#), [4](#), [6](#), [11](#)
- [60] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*, 2021. [5](#)
- [61] 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. [2](#), [4](#), [11](#)
- [62] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. [4](#), [5](#)
- [63] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv preprint arXiv:1908.02265*, 2019. [1](#)
- [64] Ling Luo, Dingyu Xue, and Xinglong Feng. Ehanet: An effective hierarchical aggregation network for face parsing. *Applied Sciences*, 10(9):3135, 2020. [6](#), [7](#)
- [65] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens Van Der Maaten. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European conference on computer vision (ECCV)*, pages 181–196, 2018. [2](#)
- [66] Longbiao Mao, Yan Yan, Jing-Hao Xue, and Hanzi Wang. Deep multi-task multi-label cnn for effective facial attribute classification. *IEEE Transactions on Affective Computing*, 2020. [8](#)
- [67] Xin Miao, Xiantong Zhen, Xianglong Liu, Cheng Deng, Vassilis Athitsos, and Heng Huang. Direct shape regression networks for end-to-end face alignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5040–5049, 2018. [7](#)
- [68] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993, 2018. [8](#)
- [69] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, pages 69–84. Springer, 2016. [8](#)
- [70] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. [1](#)
- [71] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. *arXiv preprint arXiv:1711.00937*, 2017. [4](#)
- [72] 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. [9](#)
- [73] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018. [1](#)
- [74] Shengju Qian, Keqiang Sun, Wayne Wu, Chen Qian, and Jiaya Jia. Aggregation via separation: Boosting facial landmark detector with semi-supervised style translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10153–10163, 2019. [7](#)
- [75] 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. [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [8](#), [10](#), [11](#)
- [76] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018. [1](#), [2](#)
- [77] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models

- are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 1, 2
- [78] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*, 2019. 1
- [79] 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. 10
- [80] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. *arXiv preprint arXiv:2102.12092*, 2021. 4
- [81] Joseph P Robinson, Yuncheng Li, Ning Zhang, Yun Fu, and Sergey Tulyakov. Laplace landmark localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10103–10112, 2019. 7
- [82] 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. 2, 4, 11
- [83] 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. 4, 11
- [84] 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. 4, 11
- [85] 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. 1, 3, 8, 9, 11
- [86] 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. 9
- [87] Ankit Kumar Sharma and Hassan Foroosh. Slim-cnn: A light-weight cnn for face attribute prediction. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 329–335. IEEE, 2020. 8
- [88] Ying Shu, Yan Yan, Si Chen, Jing-Hao Xue, Chunhua Shen, and Hanzi Wang. Learning spatial-semantic relationship for facial attribute recognition with limited labeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11916–11925, 2021. 2, 4, 8
- [89] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2
- [90] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv preprint arXiv:2001.07685*, 2020. 8
- [91] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vi-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019. 1
- [92] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pages 843–852, 2017. 2
- [93] 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. 7
- [94] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019. 2
- [95] Zhiqiang Tang, Xi Peng, Kang Li, and Dimitris N Metaxas. Towards efficient u-nets: A coupled and quantized approach. *IEEE transactions on pattern analysis and machine intelligence*, 42(8):2038–2050, 2019. 7
- [96] Gusi Te, Wei Hu, Yinglu Liu, Hailin Shi, and Tao Mei. Agrnet: Adaptive graph representation learning and reasoning for face parsing. *IEEE Transactions on Image Processing*, 2021. 1, 2, 4, 6, 7
- [97] Gusi Te, Yinglu Liu, Wei Hu, Hailin Shi, and Tao Mei. Edge-aware graph representation learning and reasoning for face parsing. In *European Conference on Computer Vision*, pages 258–274. Springer, 2020. 2, 4, 6, 7
- [98] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 776–794. Springer, 2020. 3
- [99] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021. 2, 5, 6, 11
- [100] Roberto Valle, Jose M Buenaposada, Antonio Valdes, and Luis Baumela. A deeply-initialized coarse-to-fine ensemble of regression trees for face alignment. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 585–601, 2018. 7
- [101] Xinyao Wang, Liefeng Bo, and Li Fuxin. Adaptive wing loss for robust face alignment via heatmap regression. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6971–6981, 2019. 5, 7
- [102] Zhen Wei, Si Liu, Yao Sun, and Hefei Ling. Accurate facial image parsing at real-time speed. *IEEE Transactions on Image Processing*, 28(9):4659–4670, 2019. 6, 7

- [103] 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. 4, 7, 11
- [104] Wenyan Wu and Shuo Yang. Leveraging intra and inter-dataset variations for robust face alignment. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 150–159, 2017. 7
- [105] 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. 5, 11
- [106] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017. 2
- [107] Xuehan Xiong and Fernando De la Torre. Supervised descent method and its applications to face alignment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 532–539, 2013. 7
- [108] Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. Ernie-vil: Knowledge enhanced vision-language representations through scene graph. *arXiv preprint arXiv:2006.16934*, 1:12, 2020. 2
- [109] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588, 2021. 2
- [110] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. 7
- [111] Yaoyao Zhong and Weihong Deng. Face transformer for recognition. *arXiv preprint arXiv:2103.14803*, 2021. 5, 6, 11
- [112] Shizhan Zhu, Cheng Li, Chen Change Loy, and Xiaoou Tang. Face alignment by coarse-to-fine shape searching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4998–5006, 2015. 7
- [113] 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. 4, 7, 11
- [114] Nan Zhuang and Cheng Yang. Few-shot knowledge transfer for fine-grained cartoon face generation. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2021. 2
- [115] Xu Zou, Sheng Zhong, Luxin Yan, Xiangyun Zhao, Jiahuan Zhou, and Ying Wu. Learning robust facial landmark detection via hierarchical structured ensemble. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 141–150, 2019. 7