General Facial Representation Learning in a Visual-Linguistic Manner

Yinglin Zheng1* Hao Yang2* Ting Zhang2 Jianmin Bao2 Dongdong Chen3 Yangyu Huang2 Lu Yuan3 Dong Chen2 Ming Zeng1† Fang Wen2
1School of Informatics, Xiamen University 2Microsoft Research Asia 3Microsoft 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

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, 50, 92] 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 vision appears to be paved by a popular learning paradigm, pre-training, due to the remarkable success of the groundbreaking models in Natural Language Processing such as BERT [24] and GPT-series [5, 74, 75], followed by a wide variety of multiple techniques [25, 45, 52, 71, 76]. 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, 69].

Meanwhile in the area of vision-language tasks involving multi-modality, there are studies [44, 53, 54, 62, 73, 87] 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 [82], 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
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
to adapt to downstream tasks.

We measure the performance on several important down-
stream 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 per-
formance 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 unexp-
lored in the literature.

2. We introduce a new framework, exploring low-level
and high-level information simultaneously for bet-
ter representation. We achieve superior transfer per-
formance 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 [49] on ImageNet recog-
nition followed by numerous improvements [38,40,85,102],
ImageNet classification [19] has been the de facto pre-
training task for visual representation learning. The back-
bones for a wide range of visual tasks including image clas-
sification, object detection, semantic segmentation and hu-
mans 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 complex-
ity of the network [47] as well as the size of the training
dataset, e.g., JFT-300M [88] and Instagram-1B [64].

As pre-training with large Transformer-based networks
on large corpora [5,24,25,52,74,75] has pushed the state of
the art forward in natural language processing, there emerge
lots of inspired advancements in visual representation learn-
ing. The vision Transformer [28, 95] 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 con-
text of images by predicting pixels whose color is indexed
by k-means clustering. On the other hand, BEiT [3] de-
digned a BERT-style loss based on the visual discrete tokens
obtained by the discrete variational auto-encoder.

Subsequently, the remarkable success of natural lan-
guage 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
the pre-trained object detector to extract image region fea-
tures as previous works [17, 43, 56, 90, 104, 105], CLIP [73]
and ALIGN [44] adopt the contrastive loss, an effective loss
in self-supervised representation learning, to pull the em-
bedding of matched image-text pairs together while pushing
those of non-matched pairs apart, followed by ALBEF [54]
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, 50, 57,
92, 93] are solved by supervised training with manually la-
beled data. Such supervised methods require a large num-
ber of training samples and may suffer from overfitting be-
cause 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 effective-
ness 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 de-
sirable for resource-limited mobile devices.

There exist several works addressing few-shot learn-
ing [4, 84] and transfer learning [1, 110] 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 learn-
ing. In contrast, we bring clarity to an unexplored regime,
weakly-supervised pre-training, on face domain as it has
been shown in [73] that leveraging massive web image-text
pairs, which provides weak supervision for images, is help-
ful for learning visual representations in few-shot scenarios.

3. FaRL

3.1. Visual Linguistic Face Data

Our goal is to learn a transferable facial representa-
tion in a visual-linguistic manner. For this purpose, we
start by collecting a sufﬁciently large dataset that con-
tains image-text pairs where the image includes a face re-
gion and the text label is natural language. Existing face
datasets [34,41,51,58,60,79] are mainly designed for a spe-
specific 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 [82], 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 [44,73], 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 [94]. 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 \(\{f^T_{\text{cls}}, f^T_1, \ldots, f^T_N\} = E_T(I), \{f^I_{\text{cos}}, f^I_1, \ldots, f^I_N\} = E_I(T)\), where \(E_I\) denotes the image Transformer-based encoder and \(E_T\) denotes the text Transformer-based encoder. \(f^T_{\text{cls}}\) is short for class token, \(e^I\) is short for end of sequence token and \(1, \ldots, N(M)\) denotes the index of visual (language) tokens. The features from the \(f^I_{\text{cos}}\) token are then fed into a projection head (a small MLP) to obtain the metric embeddings, i.e. \(e^I = P_T(f^I_{\text{cls}}), e^T = P_T(f^T_{\text{cos}})\). The contrastive loss, in the scenario of image-text pairs, is given as

\[
\begin{align*}
L_I &= -\frac{1}{B} \sum_{i=1}^{B} \frac{\sigma}{\sum_{j=1}^{B} \exp(e^I_i e^T_j / \sigma)}, \\
L_T &= -\frac{1}{B} \sum_{i=1}^{B} \frac{\sigma}{\sum_{j=1}^{B} \exp(e^T_i e^I_j / \sigma)},
\end{align*}
\]

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 \(\{I_1, \ldots, I_N\}\), the masked image \(\tilde{I}\) is also represented as \(N\) image patches \(\{\tilde{I}_1, \ldots, \tilde{I}_N\}\) with

\[
\tilde{I}_k = \begin{cases} 
I_k, & k \notin \mathcal{M} \\
\text{m}, & k \in \mathcal{M}
\end{cases},
\]

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.

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.
where \( \mathcal{M} \subset \{1, \ldots, N\} \) 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, \( \{f_{\text{cls},k}, f_1^k, \ldots, f_{14}^k\} = E_I(I) \), we feed them into a small Transformer which outputs the final hidden vectors, \( \{h_{\text{cls}}, h_1, \ldots, h_{14}\} = E_{\text{MIM}}(f_{\text{cls},k}, f_1^k, \ldots, f_{14}^k) \). The objective is to predict the masked region from the corresponding hidden vectors \( \{h_k^k, k \in \mathcal{M}\} \). Instead of directly predicting pixels requiring huge memory consumption, the discrete variational autoencoder [70] 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 \( \tilde{h}_k^k \) to predict the corresponding masked patch’s index among \( \{1, \ldots, |\mathcal{V}|\} \). The loss function is given as,

\[
L_{\text{MIM}} = - \sum_{k \in \mathcal{M}} \log p \left( q_{\text{cls}}^k(I)|\tilde{I} \right),
\]

where \( p(q_{\text{cls}}^k(I)|\tilde{I}) \) denotes the classification score of classifying the \( k \)-th hidden vector belonging to the visual token \( q_{\text{cls}}^k(I) \), where \( q_{\text{cls}} \) 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 [77].

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 \( E_{\text{MIM}} \). We implement the image encoder \( E_I \) following prior works [3,28,73] for fair comparison. Specifically, it is a 12-layer 768-width visual Transformer ViT-B/16 [28] with 87M parameters and 224 × 224 input. The input image is first split into 14 × 14 image patches, followed by a linear projection to obtain 14 × 14 patch embeddings. A learnable \( \text{cls} \) token is prepended to these 196 embeddings, which is called the [CLS] token. 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 [58] and CelebAMask-HQ [51]. 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 [58,92,93], 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: AFLFW-19 [109] with 20K images for training and 4,386 for testing, each annotated with 19 landmarks; 300W [79–81] with 3,837 images for training and 600 for testing, each annotated with 68 landmarks; WFLW [99] 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 [60] and LFSA [60]. CelebA consists of over 202K facial images, while LFSA consists of 13,143 images, both with 40 attribute annotations per image. Following [60,84], on CelebA, we use 162,770 for training and 19,962 for testing; on LFSA, 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 \( \{f_{\text{cls},k}; f_1^k; f_2^k; \ldots; f_{14}^k\} \) 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. Let \( K \) be the set of selected layers for downstream. We com-
F2D feature map of size $N \times \sqrt{N}$ responds to image patches, hence they can be reshaped to a vector for multi-label binary classification. The head is trained by layer-normalized linear to compute three feature vectors from each layer $k \in \mathcal{K}$: the $cls$ token feature $f^{l}_{cls,k}$, 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 [61] optimizer. We adopt a learning rate of $1e^{-3}$ and weight decay $1e^{-5}$. Tanh-warping [57] 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, 50, 97]. 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, 97], we simply train the head with a soft-label cross-entropy loss. UperNet [101] 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 $E_{f}$, 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)\(^1\). 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 classic vision transformer pretrained with large-scale human an-

---

\(^1\)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.
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. MIM1 means to append an additional 1-layer Transformer to $E_I$ for masked image modeling, while MIM6 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 [73] instead and do random crop to get the $224 \times 224$ input.

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. MIM1 means to append an additional 1-layer Transformer to $E_I$ for masked image modeling, while MIM6 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 [73] instead and do random crop to get the $224 \times 224$ input.

### Table 3. Ablating the different components of FaRL pre-training w.r.t downstream task performances. *ITC+MIM1+ALIGN* is the default setting of FaRL.

| Pre-training Settings | LaPa | AFLW-19 | CelebA |
|------------------------|------|---------|--------|
| ITC                    | 91.75| 1.009   | 91.31  |
| ITC+MIM                | 91.82| 1.004   | 91.21  |
| ITC+MIM1+ALIGN (FaRL)  | 92.32| 0.991   | 91.39  |
| ITC+MIM6+ALIGN         | 92.19| 1.002   | 91.38  |
| ITC+MIM1+ALIGN         | 91.99| 0.992   | 91.20  |
| ITC+ALIGN              | 91.88| 1.012   | 91.40  |
| ITC (LAION-RANDOM)     | 91.68| 1.010   | 90.76  |

Compared to ITC only, adding MIM1 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 MIM1 and ALIGN together achieves generally better scores on most downstream benchmarks (LaPa, AFLW-19). In addition, a heavier MIM6 head is not as good as MIM1, suggesting that the deeper head may weaken the effect of the MIM loss. Therefore, we select ITC+MIM1+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$_{sh}$ denotes the model that is fully fine-tuned from the vanilla FaRL for the specific downstream task. While both FaRL and FaRL$_{sh}$ accepts $224 \times 224$ inputs, we also fine-tune a model that accepts two-times larger input resolution, namely FaRL$_{448}$. FaRL$_{448}$ shares the same initial parameters with FaRL$_{sh}$, but its positional embeddings are initialized by a bi-cubic up-sampling from the positional embeddings of FaRL$_{sh}$. In order to investigate whether the gains are brought by our pre-
training method or just come from the Transformer-based framework. 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 dataset 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\(_{\text{ft}}\) brings even higher F1 scores. The ultimate performance is achieved by FaRL\(_{\text{448}}\), which outperforms the state-of-the-art method [92] 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., *nose* in *CelebAMask-HQ*). Even though, the required resolution of FaRL\(_{\text{448}}\) is lower than the resolution needed by the state-of-the-art approach [92] 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 7 and Table 8, respectively. The Transformer-based methods achieve 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\(_{\text{448}}\) 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 pre-training, by a large margin.
We freeze the pretrained backbones and only fine-tune the FairFace dataset for balanced race, gender, and age. It categorizes image dataset FairFace [46]. FairFace is a face attribute dataset and although we choose a roundabout way of filtering out face images from a general image-text pair dataset, 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 [82], 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.

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 [73] 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.

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 (%).

| Model     | White | Non-White | Average | Discrepancy |
|-----------|-------|-----------|---------|-------------|
| Age       |       |           |         |             |
| FairFace  | 60.05 | 60.63     | 60.52   | +0.58       |
| CLIP [73] | 62.25 | 61.95     | 62.00   | -0.30       |
| FaRL      | 61.49 | 61.84     | 61.78   | +0.55       |
| Gender    |       |           |         |             |
| FairFace  | 94.15 | 94.41     | 94.36   | +0.26       |
| CLIP [73] | 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.

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 [82], 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.

7. Acknowledgements

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

---

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

[1] Shivangi Aneja and Matthias Nießner. Generalized zero and few-shot transfer for facial forgery detection. arXiv preprint arXiv:2006.11863, 2020.

[2] Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. arXiv preprint arXiv:1906.00910, 2019.

[3] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. arXiv preprint arXiv:2106.08254, 2021.

[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.

[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.

[6] 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.

[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.

[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.

[9] Mathilde Caron, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features. In Proceedings of the European Conference on Computer Vision (ECCV), pages 132–149, 2018.

[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.

[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.

[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.

[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.

[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.

[15] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297, 2020.

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

[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.

[18] Arnaud Dapogny, Kevin Bailly, and Matthieu Cord. Defeafa: Deep convolutional cascade for face alignment in the wild. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 6893–6901, 2019.

[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.

[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.

[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.

[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.

[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.

[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.

[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.

[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.

[27] Xuanyi Dong and Yi Yang. Teacher supervises students 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.

[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

[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, Praveen 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 Altchê, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Dorsch, 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, Yu-Gang Jiang, Feiyeue 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

[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 Duering. 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 Jungseok 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] 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*, pages 491–507. Springer, 2020. 2

[48] Marek Kowalski, Jacek Naruniec, and Tomasz Trzciński. 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

[49] 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

[50] Abhinav Kumar, Tim K Marks, Wenzhuan Mou, Ye Wang, Michael Jones, Anoop Cherian, Toshiaki Koike-Akino, Xiaoming Liu, and Chen Feng. Luvli face alignment: Estimating 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

[51] 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

[52] 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

[53] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. Unicorn-v1: 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

[54] Junnan Li, Ramprasaath R Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caining Xiong, and Steven Hoi. Align before fuse: Vision and language representation learning with momentum distillation. arXiv preprint arXiv:2107.07651, 2021. 1, 2

[55] Jianshu Li, Fang Zhao, Jaashi Feng, Sujoy Roy, Shuicheng Yan, and Terence Sim. Landmark free face attribute prediction. IEEE Transactions on Image Processing, 27(9):4651–4662, 2018. 8

[56] Xiujuan 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

[57] 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

[58] 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

[59] 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

[60] 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

[61] Ilya Loshchilov and Frank Hutter. Decoupled weight decay. arXiv preprint arXiv:1910.13461, 2019. 1, 2

[62] Ling Luo, Dingyu Xue, and Xinglong Feng. Ehanet: An effective hierarchical aggregation network for face parsing. Applied Sciences, 10(9):3135, 2020. 6, 7

[63] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kai Ming 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

[64] Xia Ming, Xiantong Zhai, 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

[65] 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

[66] 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

[67] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748, 2018. 1

[68] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. arXiv preprint arXiv:1711.00937, 2017. 4

[69] 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

[70] 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

[71] 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

[72] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018. 1, 2

[73] 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

[74] 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

[75] 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
vision-language representations through scene graph. arXiv preprint arXiv:2006.16934, 1:12, 2020. 2

[105] 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

[106] 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

[107] Yaoyao Zhong and Weihong Deng. Face transformer for recognition. arXiv preprint arXiv:2103.14803, 2021. 5, 6

[108] Shizhan Zhu, Cheng Li, Chen-Change Loy, and Xiaou 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

[109] Shizhan Zhu, Cheng Li, Chen-Change Loy, and Xiaou 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

[110] 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

[111] 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