Self-Supervised Facial Representation Learning with Facial Region Awareness

Zheng Gao Ioannis Patras

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

Abstract

Self-supervised pre-training has been proved to be effective in learning transferable representations that benefit various visual tasks. This paper asks this question: can self-supervised pre-training learn general facial representations for various facial analysis tasks? Recent efforts toward this goal are limited to treating each face image as a whole, i.e.*, learning consistent facial representations at the image-level, which overlooks the "**consistency of local facial representations**" (*i.e.*, facial regions like eyes, nose, etc). In this work, we make a **first attempt** to propose a novel self-supervised facial representation learning framework to learn consistent global and local facial representations, **F**acial **R**egion **A**wareness (FRA). Specifically, we explicitly enforce the consistency of facial regions by matching the local facial representations across views, which are extracted with learned heatmaps highlighting the facial regions. Inspired by the mask prediction in supervised semantic segmentation, we obtain the heatmaps via cosine similarity between the per-pixel projection of feature maps and "facial mask embeddings" computed from learnable positional embeddings, which leverage the attention mechanism to globally look up the facial image for facial regions. To learn such heatmaps, we formulate the learning of facial mask embeddings as a deep clustering problem by assigning the pixel features from the feature maps to them. The transfer learning results on facial classification and regression tasks show that our FRA outperforms previous pre-trained models and more importantly, using ResNet as the unified backbone for various tasks, our FRA achieves comparable or even better performance compared with SOTA methods in facial analysis tasks.*

1. Introduction

Human face understanding is an important and challenging topic in computer vision [41, 72] and supervised learning algorithms have shown promising results on a wide range of facial analysis tasks recently [5, 31, 55, 71]. Despite the impressive progress, these methods require large-scale well-annotated training data, which is expensive to collect.

Recent works in self-supervised representation learning for visual images have shown that self-supervised pre-training is effective in improving the performance on various downstream tasks such as image classification, object detection and segmentation as it can learn general representations from unlabeled data that could be transferred to downstream visual tasks, especially tasks with limited labeled data [18, 24, 28, 30, 58, 62, 65]. Among them, **instance discrimination** (including contrastive learning [10, 22, 36] and non-contrastive learning [11, 20] paradigms) has been shown to be effective in learning generalizable self-supervised features. Instance discrimination aims to learn view-invariant representations by matching the **global representations** between the augmented views generated by aggressive image augmentations, *i.e.*, the image-level representations of the augmented views should be similar [10–12, 20, 22, 36]. Another self-supervised learning paradigm, masked image modeling (MIM) [23, 54] learns visual representations by reconstructing image content from a masked image, achieving excellent performance in full model fine-tuning. This leads to the question: **can self-supervised pre-training learn general facial representations which benefit downstream facial analysis tasks**?

Several attempts have been made to learn general facial representations for facial analysis tasks [3, 41, 72]. For example, Bulat *et al*. [3] directly applies the contrastive objective to facial features. FaRL [72] and MCF [59] combine contrastive learning and mask image modeling [23]. PCL [41] proposes to disentangle the pose-related and pose-unrelated features, achieving strong performance on both pose-related (regression) and pose-unrelated (classification) tasks. However, it runs the model forward and backward three times for each training step, which is time-consuming. Despite different techniques, these methods commonly treats each face image as a whole to learn consistent global representations at the image-level and overlook the "**spatial consistency of local representations**", *i.e.*, local facial regions (*e.g.*, eyes, nose and mouth) should also be

similar across the augmented views, thus limiting the generalization to downstream tasks. This brings us to the focus: **learning consistent global and local representations for facial representation learning**.

We argue that in order to learn consistent local representations, the model needs to look into facial regions. Towards that goal, we predict a set of heatmaps highlighting different facial regions by leveraging learnable positional embeddings as facial queries (the feature maps as keys and values) to look up the facial image globally for facial regions, which is inspired by the mask prediction in supervised segmentation [13]. For visual images, the attention mechanism of Transformer allows the learnable positional embeddings to serve as object queries for visual pattern look-up [7, 13]. In our case (facial images), the learnable positional embeddings can be used as facial queries for facial regions (**see the visualization in the supplementary material**).

In this work, taking the consistency of facial regions into account, we make a **first attempt** to propose a novel self-supervised facial representation learning framework, **F**acial **R**egion **A**wareness (FRA) that learns general facial representations by enforcing consistent global and local facial representations, based on a popular instance discrimination baseline BYOL [20] for its simplicity. Specifically, we **learn consistent local facial representations** by match them across augmented views, which are extracted by aggregating the feature maps using learned heatmaps highlighting the facial regions as weights. Inspired by the mask prediction in MaskFormer [13], we produce the heatmaps from a set of learnable positional embeddings, which are used as facial queries to look up the facial image for facial regions. A Transformer decoder takes as input the feature maps from the encoder and the learnable positional embeddings to output a set of "*facial mask embeddings*", each associated with a facial region. The facial mask embeddings are used to compute cosine similarity with the per-pixel projection of feature maps to produce the heatmaps. In addition, we enforce the consistency of global representations across views simultaneously so that the image-level information is preserved. In order to learn the heatmaps (facial mask embeddings), inspired by deep clustering [8] that learns to assign samples to clusters, we treat the facial mask embeddings as clusters and learn to assign pixel features from the feature maps to them. Specifically, we **align the per-pixel cluster assignments** of each pixel feature over the facial region clusters between the online and momentum network for the same augmented view (*i.e*., each pixel feature should have similar similarity distribution over the facial mask embeddings between the momentum teacher and online student). In contrast to supervised segmentation that directly uses ground-truth masks to supervise the learning of the masks (heatmaps) with a per-pixel binary mask loss, we formulate the learning of heatmaps as a deep clustering [8] problem that learns to assign pixel features to clusters (facial mask embeddings) in a self-supervised manner.

Our contributions can be summarized as follows:

- Taking into the consistency of local facial regions into account, we make a **first attempt** to propose a novel self-supervised facial representation learning framework, **F**acial **R**egion **A**wareness (FRA) that learns consistent global and local facial representations.
- We show that the learned heatmaps can roughly discover facial regions in the supplementary material.
- In previous works, different backbones are adopted for different facial analysis tasks (*e.g*., in face alignment the common backbone is hourglass network [63] while in facial expression recognition ResNet [21] is the popular backbone). In this work, our FRA achieves SOTA results using vanilla ResNet [21] as the unified backbone for various facial analysis tasks.
- Our FRA outperforms existing self-supervised pre-training approaches (*e.g*., BYOL [20] and PCL [41]) on facial classification (*i.e*., facial expression recognition [19, 38] and facial attribute recognition [42]) and regression (*i.e*., face alignment [48–50, 60]) tasks. More importantly, our FRA achieves comparable (*e.g*., face alignment) or even better performance (*e.g*., facial expression recognition) compared with SOTA methods in the corresponding facial analysis tasks.

2. Related work

2.1. Visual representation learning

As one of the main paradigms for self-supervised pre-training, instance discrimination learns representations by treating an image as a whole and enforcing the consistency of global representations at the image-level across augmented views. Generally, instance discrimination includes two paradigms: contrastive learning [10, 22, 36] and non-contrastive learning [11, 20]. Contrastive learning considers each image and its transformations as a distinct class, *i.e*., "positive" samples are pulled together while "negative" samples are pushed apart in the latent space. Unlike contrastive learning that relies on negative samples to avoid collapse, non-contrastive learning directly maximizes the similarity of the global representations between the augmented views without involving negative samples based on techniques like stop-gradient [11] and predictor [20]. Further works perform visual-language pre-training by applying contrastive objective to image-text pairs [29, 35, 47].

Another line of work, masked image modeling (MIM) learns visual representations by reconstructing image content from a masked image [1, 23, 54], which is inspired by the masked language modeling in NLP [15]. In contrast to instance discrimination, MIM achieves strong full model fine-tuning performance with Vision Transformers

pre-trained for enough epochs. However, these works suffer from poor linear separability and are less data-efficient than instance discrimination in few-shot scenarios [1].

2.2. Facial representation learning

Recent works on facial analysis explore self-supervised learning for several face-related tasks, such as facial expression recognition [9, 53], face recognition [9, 57], facial micro-expression recognition [46], AU detection [39, 40], face alignment (facial landmark detection) [14, 64], etc. However, these methods are **task-specific**, *i.e.*, tailored for a specific facial task and thus lack the ability to generalize to various facial analysis tasks [41]. Further efforts [3, 41, 72] focus on learning general facial representations with contrastive learning and mask image modeling [23, 54]. Bulat *et al*. [3] directly apply the contrastive objective to augmented views of the same face image, showing that general facial representation learned from pre-training can benefit various facial analysis tasks. FaRL [72] performs pre-training in a visual-linguistic manner by employing image-text contrastive learning and masked image modeling. MCF [59] leverages image-level contrastive learning and masked image modeling, along with the knowledge distilled from external ImageNet pre-trained model for facial representation learning. PCL [41] argues that directly applying the contrastive objective to face images overlooks the variances of facial poses and thus leads to pose-invariant representations, limiting the performance on pose-related tasks [66, 75]. Therefore, PCL [41] disentangles the pose-related and pose-unrelated features and then performs contrastive learning on these features, achieving strong performance on both pose-related and pose-unrelated facial analysis tasks. Despite the success, it performs forward and backward three times for each input image, which brings significant increase on computational cost. These works are commonly limited by instance discrimination and overlook the consistency of local facial regions. In contrast, inspired by supervised semantic segmentation, we learn consistent global and local facial representations by learning a set of heatmaps indicating facial regions from learnable positional embeddings, which leverage the attention mechanism to look up facial image globally for facial regions.

2.3. Facial region discovery

There are some approaches leveraging facial region (landmark) discovery for facial analysis [27, 43, 61]. Some focus on landmark detection by either learning a heatmap for each landmark via image reconstruction [27, 68], or performing pixel-level matching with an equivariance loss [56, 68]. Despite different techniques, these methods are **task-specific**, *i.e.*, landmark detection with discovery of local information, while our method is **task-agnostic**, *i.e.*, learn general facial representations for various tasks by preserving global and local information in a **image, region and pixel-level contrastive manner**. MARLIN [4] applies masked image modeling to learn general representations for facial videos by utilizing an external face parsing algorithm to discover the facial regions (*e.g.*, eyes, nose and mouth), which are used to guide the masking for the masked autoencoder. A closely related work SLPT [61] leverages the attention mechanism to estimate facial landmarks from initial facial landmarks estimates of the mean face through supervised learning. These works commonly rely on external supervisory signal, whether it is from ground-truths or additional algorithms. In contrast, we learn to discover the facial regions in an **end-to-end self-supervised** manner for facial **image** representation learning.

3. Methodology

3.1. Overview

The overview of the proposed FRA is shown in Fig. 1. The goal is to learn consistent global and local facial representations. Toward this goal, we propose two objectives: **pixel-level semantic relation** and **image/region-level semantic consistency**. Semantic relation aligns the per-pixel cluster assignments of each pixel feature over the facial mask embeddings between the online and momentum network to learn heatmaps for facial regions (Sec. 3.2) while semantic consistency directly matches the global and local facial representations across augmented views (Sec. 3.3) with the learned heatmaps.

3.2. Semantic relation

As shown in Fig. 1, our method adopts the Siamese structure of BYOL [20], a popular self-supervised pre-training baseline based on instance discrimination. Following BYOL [20], we employ two branches: the online network parameterized by θ and the momentum network parameterized by ξ. The online network θ consist of an encoder E_θ, a global projector H^g_θ and a local projector H^l_θ. The momentum network ξ has the same architecture with the online network, except ξ is updated with an exponential moving average of θ. As in BYOL [20], we also adopt additional predictors G^g_θ and G^l_θ on top of the projectors in the online network. Note that this is omitted for brevity in Fig. 1.

Given an input image $\mathbf{x}$, two random augmentations are applied to generate two augmented views $\mathbf{x}_1 = \mathcal{T}_1(\mathbf{x})$ and $\mathbf{x}_2 = \mathcal{T}_2(\mathbf{x})$, following BYOL [20]. Each augmented view $\mathbf{x}_i \in \{\mathbf{x}_1, \mathbf{x}_2\}$ is fed into an encoder E to obtain a feature map $\mathbf{F}_i \in \mathbb{R}^{C\times H\times W}$ (before global average pooling), where C, H, W are the number of channels, height and width of $\mathbf{F}_i$ and a latent representation $\mathbf{h}_i \in \{\mathbf{h}_1, \mathbf{h}_2\}$ (after global average pooling), *i.e.*, $\mathbf{h}_1 = E_\theta(\mathbf{x}_1)$ and $\mathbf{h}_2 = E_\xi(\mathbf{x}_2)$. Then each latent representation $\mathbf{h}_i$ is transformed

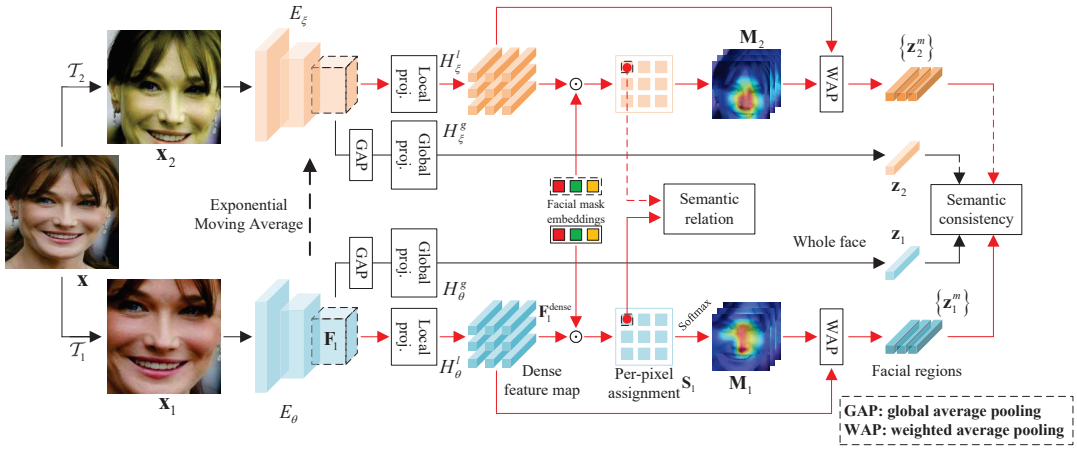

Figure 1. **Overview of the proposed FRA framework**. $\odot$ denotes cosine similarity. For each input image $\mathbf{x}$, its augmented views $\mathbf{x}_1$ and $\mathbf{x}_2$ are passed into two network branches to produce the global embeddings $\mathbf{z}_1$ and $\mathbf{z}_2$. In addition, we produce a set of heatmaps $\mathbf{M}_1$ and $\mathbf{M}_2$ indicating the local facial regions, via the correlation between the pixel features and "*facial mask embeddings*" computed from a set of learnable positional embeddings. Then we aggregate the feature map to obtain the local facial embeddings $\{\mathbf{z}^m_1\}$ and $\{\mathbf{z}^m_2\}$. The semantic consistency loss is applied to global embeddings and facial embeddings to maximize the similarity across augmented views. To learn such heatmaps, *i.e.*, *facial mask embeddings*, we treat the facial mask embeddings as facial region clusters and propose a semantic relation loss to align the cluster assignments of each pixel feature over the facial region clusters between the online and momentum network.

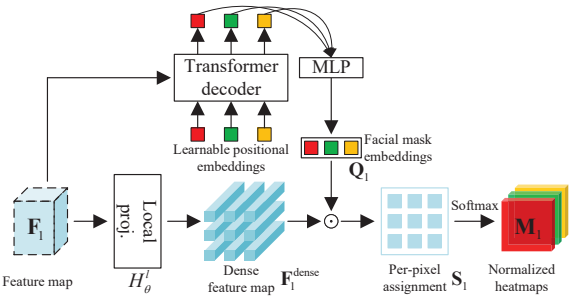

Figure 2. Generation of heatmaps using learnable positional embeddings as facial queries and the feature maps as keys and values.

by a global projector H^g to produce a global embedding $\mathbf{z}_i \in \{\mathbf{z}_1, \mathbf{z}_2\}$ of dimension $\mathbf{z}_i \in \mathbb{R}^D$, *i.e.*, $\mathbf{z}_1 = H^g_\theta(\mathbf{h}_1)$ and $\mathbf{z}_2 = H^g_\xi(\mathbf{h}_2)$.

Then we obtain a set of heatmaps $\mathbf{M}_i \in \{\mathbf{M}_1, \mathbf{M}_2\}$ highlighting the facial regions from the feature map $\mathbf{F}_i$ for each view, which is inspired by *mask classification-based supervised segmentation* [7, 13] that leverages attention mechanism to look up visual patterns globally. First the local projector (*e.g.*, H^l_θ) is applied to project the pixel features of $\mathbf{F}_i$ in a pixel-wise manner, mapping it to D dimensions to get the dense feature map $\mathbf{F}^{\text{dense}}_i \in \mathbb{R}^{D\times H\times W}$. Take view $\mathbf{x}_1$ as an example, the projected feature map can be expressed as:

$$\mathbf{F}^{\text{dense}}_1[*,u,v] = H^l_\theta(\mathbf{F}_1[*,u,v]), \tag{1}$$

where $\mathbf{F}_1[*,u,v] \in \mathbb{R}^C$ is the pixel feature at the (u,v)-th grid of $\mathbf{F}_1$. Then as shown in Fig. 2, inspired by supervised segmentation [13], we use a Transformer decoder followed by a MLP, which takes as input the feature map $\mathbf{F}_i$ and N learnable positional embeddings (*i.e.*, facial queries for looking up the facial image globally for facial regions) to predict N "*facial mask embeddings*" $\mathbf{Q}_i \in \mathbb{R}^{N\times D}$ of dimension D, where each row associated with a facial region. Next, we compute the cosine similarity between facial mask embeddings $\mathbf{Q}_i$ and dense feature map $\mathbf{F}^{\text{dense}}_i$ along the channel dimension, yielding **per-pixel cluster assignments** $\mathbf{S}_i \in \mathbb{R}^{N\times H\times W}$, where $\mathbf{S}_i[*,u,v]$ denotes the relation between the dense pixel feature $\mathbf{F}^{\text{dense}}_1[*,u,v]$ and facial mask embeddings $\mathbf{Q}_i$. Finally, we normalize $\mathbf{S}_i$ along the channel dimension with a softmax operation to encourage each channel to capture a different pattern, obtaining N heatmaps $\mathbf{M}_i \in \mathbb{R}^{N\times H\times W}$ where each vector at location (u,v) is a probabilistic similarity distribution (*i.e.*, **normalized per-pixel cluster assignments**) $\mathbf{s}^{u,v}_1$ between $\mathbf{F}^{\text{dense}}_i[*,u,v]$ and $\mathbf{Q}_i$. Note that $\mathbf{M}_i$ is a set of heatmaps where each channel of $\mathbf{M}_i$ represents a 2D heatmap $\mathbf{M}^{(m)}_i$.

To learn such heatmaps, *i.e.*, facial mask embeddings, inspired by deep clustering [8], we treat the facial mask embeddings as facial region clusters and align the per-pixel cluster assignments of each pixel feature over these clusters between the online and momentum network for the same augmented view, using the momentum network as momentum teacher [16, 73] to provide reliable target.

Following BYOL [20], we pass both augmented views

$\mathbf{x}_1$ and $\mathbf{x}_2$ through the online and momentum network. Take $\mathbf{x}_1$ as an example, the online network θ outputs the normalized per-pixel cluster assignments $\mathbf{s}_1^{u,v}$ and the momentum network outputs normalized assignments $\widehat{\mathbf{s}}_1^{u,v}$ for view $\mathbf{x}_1$. Then we learn $\mathbf{s}_1^{u,v}$ using $\widehat{\mathbf{s}}_1^{u,v}$ as guidance based on the following cross-entropy loss:

$$CE(\mathbf{s}_1^{u,v}, \widehat{\mathbf{s}}_1^{u,v}) = -\sum_{m=1}^{N} \widehat{\mathbf{s}}_1^{u,v}[m] \log \mathbf{s}_1^{u,v}[m]. \quad (2)$$

For both augmented views, we define the symmetrized semantic relation objective as:

$$\mathcal{L}_{\mathrm{r}} = \frac{1}{HW} \sum_{u,v} (CE(\mathbf{s}_1^{u,v}, \widehat{\mathbf{s}}_1^{u,v}) + CE(\mathbf{s}_2^{u,v}, \widehat{\mathbf{s}}_2^{u,v})), \quad (3)$$

where $CE(\mathbf{s}_2^{u,v}, \widehat{\mathbf{s}}_2^{u,v})$ is the cross-entropy loss for view $\mathbf{x}_2$. We apply the Sinkhorn-Knopp normalization to the target assignments from the momentum network following [8] to avoid collapse and the mean entropy maximization (ME-MAX) regularizer [1] to maximize the entropy of the prediction to encourage full use of the clusters.

3.3. Semantic consistency

In this section, we enforce the consistency of global embeddings and local facial embeddings. With the learned heatmaps $\mathbf{M}_i$, we generate the latent representations for the local facial regions through weighted average pooling:

$$\begin{aligned} \mathbf{h}_i^m &= \mathbf{M}_i^{(m)} \otimes \mathbf{F}_i \\ &= \frac{1}{\sum_{u,v} \mathbf{M}_i[m,u,v]} \sum_{u,v} \mathbf{M}_i[m,u,v] \mathbf{F}_i[*,u,v], \end{aligned} \quad (4)$$

where $\otimes$ denotes channel-wise weighted average pooling, $\mathbf{M}_i^{(m)}$ is the m-th channel (heatmap) of $\mathbf{M}_i$ and $\mathbf{h}_i^m \in \mathbb{R}^C$ is the corresponding latent representation produced with $\mathbf{M}_i^{(m)}$. The facial embeddings $\{\mathbf{z}_1^m : \mathbf{z}_1^m \in \mathbb{R}^D\}_{m=1}^N$ and $\{\mathbf{z}_2^m : \mathbf{z}_2^m \in \mathbb{R}^D\}_{m=1}^N$ are obtained accordingly with the local projector H_θ^l and H_ξ^l:

$$\begin{aligned} \mathbf{z}_1^m &= H_\theta^l(\mathbf{h}_1^m), \\ \mathbf{z}_2^m &= H_\xi^l(\mathbf{h}_2^m). \end{aligned} \quad (5)$$

We then match the global embeddings and local facial embeddings across views using the negative cosine similarity in BYOL [20]:

$$\begin{aligned} \mathcal{L}_{\mathrm{sim}}(\mathbf{z}_1, \mathbf{z}_2) = &-(\lambda_{\mathrm{c}} \times f_{\mathrm{s}}(G_\theta^g(\mathbf{z}_1), \mathbf{z}_2) + \\ &+(1-\lambda_{\mathrm{c}}) \times \frac{1}{N} \sum_{m=1}^{N} f_{\mathrm{s}}(G_\theta^l(\mathbf{z}_1^m), \mathbf{z}_2^m)), \end{aligned} \quad (6)$$

where $f_{\mathrm{s}}(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u}^\top \mathbf{v}}{\|\mathbf{u}\|_2 \|\mathbf{v}\|_2}$ denotes the cosine similarity between the vectors $\mathbf{u}$ and $\mathbf{v}$, λ_{c} is the loss weight, G_θ^g and G_θ^l are the predictors on top of the projectors H_θ^g and H_θ^l, respectively. Following BYOL [20], we symmetrize the loss $\mathcal{L}_{\mathrm{sim}}(\mathbf{z}_1, \mathbf{z}_2)$ defined in Eq. (6) by passing $\mathbf{x}_1$ through the momentum network ξ and $\mathbf{x}_2$ through the online network θ to compute $\mathcal{L}_{\mathrm{sim}}(\mathbf{z}_2, \mathbf{z}_1)$. The semantic consistency objective can be expressed as follows:

$$\mathcal{L}_{\mathrm{c}} = \mathcal{L}_{\mathrm{sim}}(\mathbf{z}_1, \mathbf{z}_2) + \mathcal{L}_{\mathrm{sim}}(\mathbf{z}_2, \mathbf{z}_1). \quad (7)$$

3.4. Overall objective

We jointly optimize the semantic relation objective Eq. (3) and the semantic consistency objective Eq. (7), leading to the following overall objective:

$$\mathcal{L} = \mathcal{L}_{\mathrm{c}} + \lambda_{\mathrm{r}} \mathcal{L}_{\mathrm{r}}, \quad (8)$$

where λ_{r} is the loss weight for balancing $\mathcal{L}_{\mathrm{c}}$ and $\mathcal{L}_{\mathrm{r}}$.

4. Experiments

4.1. Experimental setups

4.1.1 Implementation details

We use the same augmentation strategy as in [20, 25]. The number of heatmaps N is set to 8 empirically. The loss weight λ_{c} and λ_{r} are set to 0.5/0.1, respectively. For fair comparisons, the other hyper-parameters are kept the same as BYOL [20] in all experiments. The architecture and pre-training details are provided in the supplementary material.

4.1.2 Baselines

Our baselines are self-supervised pre-training approaches for visual images (*e.g.*, BYOL [20] and LEWEL [25]), and pre-training approaches for facial images (*e.g.*, Bulat *et al.* [3] and PCL [41]). Note that SwAV [8] is equivalent to Bulat *et al.* [3]. As we adopt BYOL [20] as the pre-training backbone, we compare our FRA with BYOL [20] in all experiments. We also compare our FRA with another pre-training method LEWEL [25], which learns consistent local representations for visual images. Moreover, we perform comparisons with SOTA methods in the corresponding downstream tasks.

4.2. Evaluation protocols

Following the common practice in previous works [41, 72], we evaluate the transfer performance of the self-supervised pre-trained facial representations on several popular downstream facial analysis tasks: facial expression recognition (FER) [2, 38], facial attribute recognition (FAR) [42] and face alignment (FA) [48–50, 60]. Specifically, we use

Table 1. **Comparisons with weakly-supervised pre-trained vision transformer on several downstream facial analysis tasks**, including facial expression recognition (AffectNet), facial attribute recognition (CelebA) and face alignment (300W).

Method	Arch.	Params.	Pre-training settings			Downstream performances		
			Dataset	Scale	Supervision	AffectNet Acc. ↑	CelebA Acc. ↑	300W NME ↓
FaRL [72]	ViT-B/16 [17]	86M	LAION-FACE [72]	20M	face image + text	64.85	91.88	3.08
FRA	R50 [21]	24M	VGGFace2 [6]	3.3M	face image	**66.16**	**92.02**	**2.91**

Table 2. **Comparisons on facial expression recognition**. We report the Top-1 accuracy on test set. **Text** denotes text supervision. †: our reproduction using the official codes.

Method	Text	FERPlus	RAF-DB	AffectNet
Supervised				
KTN [32]	×	90.49	88.07	63.97
RUL [69]	×	88.75	88.98	61.43
EAC [70]	×	90.05	90.35	65.32
Weakly-Supervised				
FaRL [72]†	✓	88.62	88.31	64.85
CLEF [67]	✓	89.74	90.09	65.66
Self-supervised				
MCF [59]†	×	88.17	86.86	60.98
Bulat *et al*. [3, 8]	×	-	-	60.20
BYOL [20]	×	89.25	89.53	65.65
LEWEL [25]	×	85.61	81.85	61.20
PCL [41]	×	85.87	85.92	60.77
FRA (LP)	×	78.13	73.89	57.38
FRA (FT)	×	89.78	89.95	**66.16**
FRA (EAC)	×	**90.62**	**90.76**	65.85

Table 3. **Comparisons on CelebA [42] facial attribute recognition**. We report the averaged accuracy over all attributes. †: our reproduction using the official codes. ∗: results cited from [72].

Method	Acc. (%)
Supervised	
DMM [44]	91.70
SlimCNN [51]	91.24
AFFAIR [34]	91.45
Self-supervised	
SSPL [52]	91.77
Bulat *et al*. [3, 8]∗	89.65
SimCLR [10]∗	91.08
BYOL [20]	91.56
LEWEL [25]	90.69
PCL [41]	91.48
MCF [59]†	91.33
FRA (LP)	90.86
FRA (FT)	**92.02**

the pre-trained weights to initialize the backbone of downstream tasks and then learn the backbone and task-specific head (attached to the backbone) jointly. Following [72], we report the performance with linear probe (denoted by "**LP**") and fine-tuning (denoted by "**FT**"). The details of the downstream tasks are described as follows:

Facial expression recognition is a multi-class classification task where the goal is to categorize the emotional expressions (*e.g*., anger, fear and surprise) for a given face image. Three widely-used datasets are adopted: FERPlus [2], RAF-DB [38] and AffectNet [45]. For RAF-DB, we use the basic emotion subset following [32, 41, 70]. For AffectNet, we report the results with 7 emotion classes (*i.e*., neutral, happy, sad, surprise, fear, anger, disgust) following [32, 70].

Facial attribute recognition is a multi-label classification task to predict various attributes (*e.g*., gender, age and race) of a given face image. We adopt the popular benchmark CelebA [42], which consists of more than 200K face images with 40 facial attributes per image. Following [72], we report the averaged accuracy over all attributes.

Face alignment is a regression task to predict 2D face landmark coordinates on a face image. We use two popular benchmarks: WFLW [60] and 300W [48–50]. Following the common practice [14, 26, 74], we report normalized mean error (NME), failure rate (FR) and AUC. For 300W, we report the results on full test set, common (554 images) and challenge (135 images) splits of the test set following [26, 74].

4.3. Comparisons with weakly-supervised pre-training

In Tab. 1, we compare our FRA with SOTA pre-trained Transformer FaRL [72], which is a weakly-supervised model pre-trained on 20M visual-linguistic data (face image and text) with image-text contrastive learning and mask image modeling. We fine-tune both the pre-trained feature backbone and the task-specific head on the corresponding downstream facial analysis task. Our self-supervised FRA with 24M parameter ResNet-50 achieves superior perfor-

Table 4. **Comparisons on face alignment**. †: our reproduction using the official codes.

Method	Venue	Arch.	WFLW			300W (NME ↓)		
			NME ↓	$FR_{10\%}$ ↓	$AUC_{10\%}$ ↑	Full	Comm.	Chal.
Supervised								
SLPT [61]	[CVPR'22]	ResNet [21]	4.20	3.04	0.588	3.20	2.78	4.93
DTLD [33]	[CVPR'22]	ResNet [21]	4.08	2.76	-	2.96	2.59	4.50
RePFormer [37]	[IJCAI'22]	ResNet [21]	4.11	-	-	3.01	-	-
ADNet [26]	[ICCV'21]	Hourglass [63]	4.14	2.72	0.602	2.93	2.53	4.58
STAR [74]	[CVPR'23]	Hourglass [63]	4.02	2.32	0.605	2.87	2.52	4.32
Self-supervised								
MCF [59] (concurrent work)	[ACM MM'23]	ViT [17]	**3.96**	**1.40**	**0.609**	2.98	**2.60**	4.51
Bulat *et al*. [3, 8]	[ECCV'22]	ResNet [21]	4.57	-	-	3.20	-	-
BYOL [20]	[NeurIPS'20]	ResNet [21]	4.29	2.96	0.579	3.03	2.66	4.55
LEWEL [25]	[CVPR'22]	ResNet [21]	4.52	4.50	0.563	3.09	2.70	4.71
PCL [41]†	[CVPR'23]	ResNet [21]	4.84	6.18	0.535	3.35	2.77	5.12
FRA	Ours	ResNet [21]	4.11	2.53	0.591	**2.91**	**2.60**	**4.46**

mance compared with weakly-supervised FaRL [72] with 86M parameter ViT-B/16 and text supervision on all tasks.

4.4. Transfer learning

In this section, we compare our FRA with self-supervised pre-training approaches and SOTA methods in several downstream tasks. Please refer to the supplementary material for setup details.

4.4.1 Facial expression recognition

The results on facial expression recognition are reported in Tab. 2. We observe: **(1)** With the setting of fine-tuning (FT), our FRA outperforms previous self-supervised pre-training approaches for visual images (*e.g*., BYOL and LEWEL) and pre-training approaches tailored for facial images (*e.g*., PCL, MCF). In particular, our FRA using a 24M parameter ResNet-50 surpasses the concurrent work MCF [59] with 86M parameter ViT-B/16 [17]. **(2)** By simply learning a linear classifier on top of the encoder backbone, our FRA outperforms SOTA facial expression recognition methods with sophisticated designs (*e.g*., EAC [70]) on AffectNet [45], the largest facial expression recognition dataset. **(3)** More importantly, by using our pre-trained model to initialize the backbone of SOTA facial expression recognition method EAC [70], our variant "FRA (EAC)" consistently improves EAC [70] on all datasets, which suggests "FRA (EAC)" outperforms SOTA FER methods and demonstrates the superiority of the proposed self-supervised pre-training.

4.4.2 Facial attribute recognition

As shown in Tab. 3, our FRA outperforms both self-supervised pre-training approaches for visual images and pre-training approaches tailored for facial images. The results on facial expression recognition and facial attribute recognition show that our FRA learns better facial representations for facial classification task.

4.4.3 Face alignment

As shown in Tab. 4, despite SOTA face alignment methods (*e.g*., ADNet [26] and STAR [74]) commonly rely on hourglass network [63] for feature extraction, which is tailored for regression tasks like landmark detection, our method based on ResNet backbone achieves comparable performance with these SOTA methods (*e.g*., 2.91 vs. 2.87 on 300W). The results on classification (*e.g*., facial expression recognition) and regression tasks (*e.g*., face alignment) show that **our FRA achieves SOTA results using vanilla ResNet [21] as the unified backbone for various facial analysis tasks**.

4.5. Ablation studies

We pre-train the model on VGGFace2 [6] and then evaluate it on facial expression recognition (RAF-DB) and facial attribute recognition (CelebA), as described in Sec. 4.2.

4.5.1 Effect of different modules

In Tab. 5, we investigate the contributions of the proposed semantic consistency loss (i.e, global consistency of whole face and local consistency of facial regions) and semantic

Table 5. **Effect of different modules**. **GC** denotes the global consistency for aligning images, **LC** denotes the local consistency for aligning facial regions with heatmaps and **SR** represents semantic relation for aligning pixels and heatmaps.

GC	LC	SR	RAF-DB ↑	CelebA ↑	300W ↓
✓	-	-	87.82	90.63	3.56
-	-	✓	85.34	90.66	3.25
-	✓	-	87.46	90.78	3.19
✓	✓	-	88.05	90.89	3.26
✓	✓	✓	**88.72**	**91.18**	**3.14**

Table 6. Effect of the number of heatmaps N.

N	8	32	64
RAF-DB	**88.72**	88.36	88.18
CelebA	**91.18**	91.01	90.98

Table 7. **Effect of the loss weights**. Please refer to Tab. 5 for **GC**, **LC** and **SR**.

Settings	λ_c	λ_r	RAF-DB	CelebA
GC	1.0	0	87.82	90.63
GC + LC	0.5	0	88.05	90.89
GC + LC + SR (FRA)	0.5	0.1	**88.72**	**91.18**
GC + LC + SR (FRA)	0.5	0.5	88.45	91.04
GC + LC + SR (FRA)	0.5	1.0	88.08	90.46

relation loss to our approach. Note that the global consistency (first row) is BYOL [20]. We have the following observations: **(1)** The variant using all losses achieves the best results. **(2)** GC is essential to avoid degeneration on classification task. **(3)** LC or SR alone benefits regression task (landmark). Altogether, LC and SR improve BYOL [20] (GC) on both classification and regression by capturing spatial/local information, which validates our facial region awareness.

4.5.2 Effect of the number of heatmaps

In Tab. 6, we study the effect of the number of heatmaps. We observe that the best setting is $N = 8$, which is close to the facial landmarks number 5. This suggests that given enough face images for training, a suitable N can encourage the model to learn face-specific patterns, which helps the transfer learning performance on various facial analysis tasks. Further increasing the number of heatmaps might force the model to look into fine-grained patterns that may not be suitable for facial tasks.

4.5.3 Effect of loss weights

In Tab. 7, we ablate the weights for the semantic consistency loss and semantic relation loss. We find that setting $\lambda_c = 0.5$ and $\lambda_r = 0.1$ works best. When $\lambda_c = 1.0$ and $\lambda_r = 0$, only the consistency of global representations is applied, and the model performs relatively worse, which suggests the importance of the consistency of local representations and the semantic relation loss. By using the semantic relation objective, the performance is significantly improved. However, when λ_r is too high, the performance degrades as the pixel-level consistency between the online and momentum network might affect the capture of image/object-level information.

4.5.4 Effect of Transformer decoder layers

Table 8. **Effect of Transformer decoder layers**. 0 decoder layer represents BYOL [20] where only the consistency of global representation is enforced.

# decoder layer	0	1	2	3
RAF-DB	87.82	88.72	89.01	89.06
CelebA	90.63	91.18	91.30	91.35

In Tab. 8, we study the effect of the number of decoder layers used for heatmap prediction. We observe that a single decoder layer is able to produce decent results, showing that a 1-layer decoder is large enough to capture the facial region (landmarks) relations in face images. The performance gain diminishes as the decoder depth increases. By default, we only use 1 decoder layer for fast training.

5. Conclusion

In this work, we propose a novel self-supervised facial representation learning framework to learn consistent global and local facial representations, **F**acial **R**egion **A**wareness (FRA). We learn a set of heatmaps indicating facial regions from learnable positional embeddings, which leverages the attention mechanism to look up facial image globally for facial regions. We show that our FRA outperforms previous pre-trained models on several facial classification and regression tasks. More importantly, using ResNet as the unified backbone, our FRA achieves comparable or even better performance compared with SOTA methods in facial analysis tasks.

Acknowledgement

This work was supported by the EU H2020 AI4Media No.951911 project. We thank Zengqun Zhao for his helpful comments on facial expression recognition.

References

[1] Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Florian Bordes, Pascal Vincent, Armand Joulin, Mike Rabbat, and Nicolas Ballas. Masked siamese networks for label-efficient learning. In *European Conference on Computer Vision*, pages 456–473. Springer, 2022. 2, 3, 5

[2] Emad Barsoum, Cha Zhang, Cristian Canton Ferrer, and Zhengyou Zhang. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, page 279–283, New York, NY, USA, 2016. Association for Computing Machinery. 5, 6

[3] Adrian Bulat, Shiyang Cheng, Jing Yang, Andrew Garbett, Enrique Sanchez, and Georgios Tzimiropoulos. Pre-training strategies and datasets for facial representation learning. In *Computer Vision – ECCV 2022*, pages 107–125, Cham, 2022. Springer Nature Switzerland. 1, 3, 5, 6, 7

[4] Zhixi Cai, Shreya Ghosh, Kalin Stefanov, Abhinav Dhall, Jianfei Cai, Hamid Rezatofighi, Reza Haffari, and Munawar Hayat. Marlin: Masked autoencoder for facial video representation learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1493–1504, 2023. 3

[5] Jiajiong Cao, Yingming Li, and Zhongfei Zhang. Partially shared multi-task convolutional neural network with local constraint for face attribute learning. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4290–4299, 2018. 1

[6] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74. IEEE, 2018. 6, 7

[7] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision – ECCV 2020*, pages 213–229, Cham, 2020. Springer International Publishing. 2, 4

[8] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2, 4, 5, 6, 7

[9] Jia-Ren Chang, Yong-Sheng Chen, and Wei-Chen Chiu. Learning facial representations from the cycle-consistency of face. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9660–9669, 2021. 3

[10] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, pages 1597–1607. PMLR, 2020. 1, 2, 6

[11] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15745–15753, 2021. 1, 2

[12] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 1

[13] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. In *Advances in Neural Information Processing Systems*, pages 17864–17875. Curran Associates, Inc., 2021. 2, 4

[14] Zezhou Cheng, Jong-Chyi Su, and Subhransu Maji. On equivariant and invariant learning of object landmark representations. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9877–9886, 2021. 3, 6

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019. 2

[16] Xiaoyi Dong, Jianmin Bao, Ting Zhang, Dongdong Chen, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, and Nenghai Yu. Bootstrapped masked autoencoders for vision bert pretraining. In *Computer Vision – ECCV 2022*, pages 247–264, Cham, 2022. Springer Nature Switzerland. 4

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 6, 7

[18] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9588–9597, 2021. 1

[19] Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, et al. Challenges in representation learning: A report on three machine learning contests. In *Neural Information Processing: 20th International Conference, ICONIP 2013, Daegu, Korea, November 3-7, 2013. Proceedings, Part III 20*, pages 117–124. Springer, 2013. 2

[20] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap your own latent - a new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, pages 21271–21284. Curran Associates, Inc., 2020. 1, 2, 3, 4, 5, 6, 7, 8

[21] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 2, 6, 7

[22] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9726–9735, 2020. 1, 2

[23] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15979–15988, 2022. 1, 2, 3

[24] Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Philip Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *ICLR 2019*. ICLR, 2019. 1

[25] Lang Huang, Shan You, Mingkai Zheng, Fei Wang, Chen Qian, and Toshihiko Yamasaki. Learning where to learn in cross-view self-supervised learning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14431–14440, 2022. 5, 6, 7

[26] Yangyu Huang, Hao Yang, Chong Li, Jongyoo Kim, and Fangyun Wei. Adnet: Leveraging error-bias towards normal direction in face alignment. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3060–3070, 2021. 6, 7

[27] Tomas Jakab, Ankush Gupta, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object landmarks through conditional image generation. In *Advances in Neural Information Processing Systems*, pages 4016–4027. Curran Associates, Inc., 2018. 3

[28] X. Ji, A. Vedaldi, and J. Henriques. Invariant information clustering for unsupervised image classification and segmentation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9864–9873, 2019. 1

[29] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. 2

[30] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, pages 18661–18673. Curran Associates, Inc., 2020. 1

[31] Abhinav Kumar, Tim K. Marks, Wenxuan 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 *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8233–8243, 2020. 1

[32] Hangyu Li, Nannan Wang, Xinpeng Ding, Xi Yang, and Xinbo Gao. Adaptively learning facial expression representation via c-f labels and distillation. *IEEE Transactions on Image Processing*, 30:2016–2028, 2021. 6

[33] Hui Li, Zidong Guo, Seon–Min Rhee, Seungju Han, and Jae-Joon Han. Towards accurate facial landmark detection via cascaded transformers. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4166–4175, 2022. 7

[34] 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. 6

[35] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *Advances in Neural Information Processing Systems*, pages 9694–9705. Curran Associates, Inc., 2021. 2

[36] Junnan Li, Pan Zhou, Caiming Xiong, and Steven Hoi. Prototypical contrastive learning of unsupervised representations. In *International Conference on Learning Representations*, 2021. 1, 2

[37] Jinpeng Li, Haibo Jin, Shengcai Liao, Ling Shao, and Pheng-Ann Heng. Repformer: Refinement pyramid transformer for robust facial landmark detection. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 1088–1094. International Joint Conferences on Artificial Intelligence Organization, 2022. Main Track. 7

[38] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2584–2593, 2017. 2, 5, 6

[39] Yong Li, Jiabei Zeng, Shiguang Shan, and Xilin Chen. Self-supervised representation learning from videos for facial action unit detection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10916–10925, 2019. 3

[40] Yong Li, Jiabei Zeng, and Shiguang Shan. Learning representations for facial actions from unlabeled videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):302–317, 2022. 3

[41] Yuanyuan Liu, Wenbin Wang, Yibing Zhan, Shaoze Feng, Kejun Liu, and Zhe Chen. Pose-disentangled contrastive learning for self-supervised facial representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9717–9728, 2023. 1, 2, 3, 5, 6, 7

[42] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 3730–3738, 2015. 2, 5, 6

[43] Yunfan Lu, Zipeng Wang, Minjie Liu, Hongjian Wang, and Lin Wang. Learning spatial-temporal implicit neural representations for event-guided video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1557–1567, 2023. 3

[44] 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*, 13(2):818–828, 2022. 6

[45] Ali Mollahosseini, Behzad Hasani, and Mohammad H. Mahoor. Affectnet: A database for facial expression, valence,

and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2019. 6, 7

[46] Xuan-Bac Nguyen, Chi Nhan Duong, Xin Li, Susan Gauch, Han-Seok Seo, and Khoa Luu. Micron-bert: Bert-based facial micro-expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1482–1492, 2023. 3

[47] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2

[48] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *2013 IEEE International Conference on Computer Vision Workshops*, pages 397–403, 2013. 2, 5, 6

[49] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. A semi-automatic methodology for facial landmark annotation. In *2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 896–903, 2013.

[50] 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, 5, 6

[51] 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, 2020. 6

[52] 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 *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11911–11920, 2021. 6

[53] Yuxuan Shu, Xiao Gu, Guang-Zhong Yang, and Benny P L Lo. Revisiting self-supervised contrastive learning for facial expression recognition. In *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022*. BMVA Press, 2022. 3

[54] Chenxin Tao, Xizhou Zhu, Weijie Su, Gao Huang, Bin Li, Jie Zhou, Yu Qiao, Xiaogang Wang, and Jifeng Dai. Siamese image modeling for self-supervised vision representation learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2132–2141, 2023. 1, 2, 3

[55] 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*, 30:8236–8250, 2021. 1

[56] James Thewlis, Samuel Albanie, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of landmarks by descriptor vector exchange. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6360–6370, 2019. 3

[57] Hao Wang, Min Li, Yangyang Song, Youjian Zhang, and Liying Chi. Ucol: Unsupervised learning of discriminative facial representations via uncertainty-aware contrast. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(2):2510–2518, 2023. 3

[58] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3024–3033, 2021. 1

[59] Yue Wang, Jinlong Peng, Jiangning Zhang, Ran Yi, Liang Liu, Yabiao Wang, and Chengjie Wang. Toward high quality facial representation learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, page 5048–5058, New York, NY, USA, 2023. Association for Computing Machinery. 1, 3, 6, 7

[60] Wenyan Wu, Chen Qian, Shuo Yang, Quan Wang, Yici Cai, and Qiang Zhou. Look at boundary: A boundary-aware face alignment algorithm. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2129–2138, 2018. 2, 5, 6

[61] Jiahao Xia, Weiwei Qu, Wenjian Huang, Jianguo Zhang, Xi Wang, and Min Xu. Sparse local patch transformer for robust face alignment and landmarks inherent relation learning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4042–4051, 2022. 3, 7

[62] Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, and Han Hu. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16684–16693, 2021. 1

[63] Jing Yang, Qingshan Liu, and Kaihua Zhang. Stacked hourglass network for robust facial landmark localisation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2025–2033, 2017. 2, 7

[64] Sejong Yang, Subin Jeon, Seonghyeon Nam, and Seon Joo Kim. Dense interspecies face embedding. In *Advances in Neural Information Processing Systems*, pages 33275–33288. Curran Associates, Inc., 2022. 3

[65] M. Ye, X. Zhang, P. C. Yuen, and S. Chang. Unsupervised embedding learning via invariant and spreading instance feature. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6203–6212, 2019. 1

[66] Lijun Yin, Xiaozhou Wei, Yi Sun, Jun Wang, and Matthew J Rosato. A 3d facial expression database for facial behavior research. In *7th international conference on automatic face and gesture recognition (FGR06)*, pages 211–216. IEEE, 2006. 3

[67] Xiang Zhang, Taoyue Wang, Xiaotian Li, Huiyuan Yang, and Lijun Yin. Weakly-supervised text-driven contrastive learning for facial behavior understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20751–20762, 2023. 6

[68] Yuting Zhang, Yijie Guo, Yixin Jin, Yijun Luo, Zhiyuan He, and Honglak Lee. Unsupervised discovery of object land-

marks as structural representations. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2694–2703, 2018. 3

[69] Yuhang Zhang, Chengrui Wang, and Weihong Deng. Relative uncertainty learning for facial expression recognition. In *Advances in Neural Information Processing Systems*, pages 17616–17627. Curran Associates, Inc., 2021. 6

[70] Yuhang Zhang, Chengrui Wang, Xu Ling, and Weihong Deng. Learn from all: Erasing attention consistency for noisy label facial expression recognition. In *Computer Vision – ECCV 2022*, pages 418–434, Cham, 2022. Springer Nature Switzerland. 6, 7

[71] Kaili Zhao, Wen-Sheng Chu, and Honggang Zhang. Deep region and multi-label learning for facial action unit detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3391–3399, 2016. 1

[72] Yinglin Zheng, Hao Yang, Ting Zhang, Jianmin Bao, Dongdong Chen, Yangyu Huang, Lu Yuan, Dong Chen, Ming Zeng, and Fang Wen. General facial representation learning in a visual-linguistic manner. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18676–18688, 2022. 1, 3, 5, 6, 7

[73] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. Image BERT pre-training with online tokenizer. In *International Conference on Learning Representations*, 2022. 4

[74] Zhenglin Zhou, Huaxia Li, Hong Liu, Nanyang Wang, Gang Yu, and Rongrong Ji. Star loss: Reducing semantic ambiguity in facial landmark detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15475–15484, 2023. 6, 7

[75] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z. Li. Face alignment across large poses: A 3d solution. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 146–155, 2016. 3