Pre-training strategies and datasets for facial representation learning

Adrian Bulat
Samsung AI Cambridge
adrian@adrianbulat.com

Shiyang Cheng
Samsung AI Cambridge
shiyang.c@samsung.com

Jing Yang
University of Nottingham
jing.yang2@nottingham.ac.uk

Andrew Garbett
Samsung AI Cambridge
atgarbett@gmail.com

Enrique Sanchez
Samsung AI Cambridge
e.lozano@samsung.com

Georgios Tzimiropoulos
Samsung AI Cambridge
Queen Mary University of London
g.tzimiropoulos@qmul.ac.uk

Abstract

What is the best way to learn a universal face representation? Recent work on Deep Learning in the area of face analysis has focused on supervised learning for specific tasks of interest (e.g. face recognition, facial landmark localization etc.) but has overlooked the overarching question of how to find a facial representation that can be readily adapted to several facial analysis tasks and datasets. To this end, we make the following 4 contributions: (a) we introduce, for the first time, a comprehensive evaluation benchmark for facial representation learning consisting of 5 important face analysis tasks. (b) We systematically investigate two ways of large-scale representation learning applied to faces: supervised and unsupervised pre-training. Importantly, we focus our evaluations on the case of few-shot facial learning. (c) We investigate important properties of the training datasets including their size and quality (labelled, unlabelled or even uncurated). (d) To draw our conclusions, we conducted a very large number of experiments. Our main two findings are: (1) Unsupervised pre-training on completely in-the-wild, uncurated data provides consistent and, in some cases, significant accuracy improvements for all facial tasks considered. (2) Many existing facial video datasets seem to have a large amount of redundancy. We will release code here for pre-trained models and data to facilitate future research.

1. Introduction

Supervised learning with Deep Neural Networks has been the standard approach to solving several Computer Vision problems over the recent past years [27, 53, 60, 29, 39]. Among others, this approach has been very successfully applied to several face analysis tasks including face detection [6, 81, 17, 36], recognition [58, 68, 69, 77, 16] and landmark localization [2, 3, 87, 70]. For example, face recognition was one of the domains where even very early attempts in the area of deep learning demonstrated performance of super-human accuracy [49, 62]. Beyond deep learning, this success can be largely attributed to the fact that for most face-related application domains, large scale datasets could be readily collected and annotated, see for example [7, 3].

There are several concerns related to the above approach. Firstly, from a practical perspective, collecting and annotating new large scale face datasets is still necessary; examples of this are context-dependent domains like emotion recognition [23, 63, 64] and surveillance [5, 21], or new considerations of existing problems like fair face recognition [55, 61]. Secondly, from a methodological point of view, it is unsatisfactory for each and every application to require its own large-scale dataset, although there is only one object of interest, that is the human face.

To this end, we investigate, for the first time to our knowledge, the task of large-scale learning universal facial representation in a principled and systematic manner. In particular, we shed light to the following research questions:

- *"What is the best way to learn a universal facial representation that can be readily adapted to new tasks and datasets? Which facial representation is more amenable to few-shot facial learning?"*
- *"What is the importance of different training dataset properties (including size and quality) in learning this representation? Can we learn powerful facial feature representations from uncurated facial data as well?"*

To address the aforementioned questions, **we make** the following **4 contributions**:

1. We introduce, for the first time, a comprehensive and principled evaluation benchmark for facial representation learning consisting of 5 important face analysis tasks, namely face recognition, AU recognition, emotion recog-

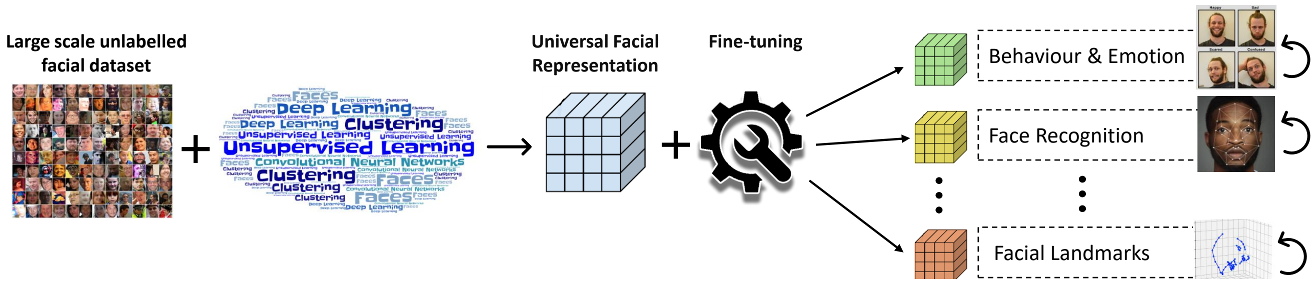

Figure 1: We advocate for a new paradigm to solving face analysis based on the following pipeline: (1) collection of large-scale unlabelled facial dataset, (2) (task agnostic) network pre-training for universal facial representation learning, and (3) facial task-specific fine-tuning. **Our main result** is that even when training on a completely in-the-wild, uncurated dataset downloaded from Flickr, this generic pipeline provides consistent and, in some cases, significant accuracy improvements for all facial tasks considered.

nition, landmark localization and 3D reconstruction.

2. Within this benchmark, and for the first time, we systematically evaluate 2 ways of large-scale representation learning applied to faces: supervised and unsupervised pre-training. Importantly, we focus our evaluations on the case of few-shot facial learning where only a limited amount of data is available for the downstream tasks.
3. We systematically evaluate the role of datasets in learning the facial feature presentations by constructing training datasets of varying size and quality. To this end, we considered ImageNet, several existing curated face datasets but also a new in-the-wild, uncurated face dataset downloaded from Flick.
4. We conducted extensive experiments to answer the aforementioned research questions and from them we were able to draw several interesting observations and conclusions.

Our main findings are: (a) Even when training on a completely in-the-wild, uncurated dataset downloaded from Flickr, unsupervised pre-training pipeline provides consistent and, in some cases, significant accuracy improvements for all facial tasks considered. (b) We found that many existing facial video datasets seem to have a large amount of redundancy. Given that unsupervised pre-training is cheap and that the cost of annotating facial datasets is often significant, some of our findings could be particularly important for researchers when collecting new facial datasets is under consideration. Finally, we will release code, pre-trained models and data to facilitate future research.

2. Related Work

Facial transfer learning: Transfer learning in Computer Vision typically consists of ImageNet pre-training followed by fine-tuning on the downstream task [53, 15, 10]. Because most recent face-related works are based on the collection of larger and larger facial datasets [24, 3, 44], the importance of transfer learning has been overlooked in face analysis and, especially, the face recognition literature. ImageNet pre-training has been applied to face analysis when training on small datasets is required, for example for emotion recognition [45], face anti-spoofing [48] and facial landmark localization [70]. Furthermore, the VGG-Face [47] or other large face datasets (e.g. [44]) have been identified as better alternatives by several works, see for example [19, 30, 78, 31, 52, 51, 48, 33, 37]. To our knowledge, we are the first to systematically evaluate supervised network pre-training using both ImageNet and VGG-Face datasets on several face analysis tasks.

Facial datasets: The general trend is to collect larger and larger facial datasets for the face-related task in hand [24, 3, 44]. Also it is known that label noise can severely impact accuracy (e.g. see Table 6 of [16]). Beyond faces, the work of [40] presents a study which shows the benefit of *weakly supervised* pre-training on much larger datasets for general image classification and object detection. Similarly, we also investigate the impact of the size of facial datasets on *unsupervised pre-training* for facial representation learning. Furthermore, one of our main results is to show that a high-quality facial representation can be learned even when a completely uncurated face dataset is used.

Few-shot face analysis: Few-shot refers to both low data and label regime. There is very little work in this area. To our knowledge, there is no prior work on few-shot face recognition where the trend is to collect large-scale datasets with millions of samples (e.g. [24]). There is no systematic study for the task of emotion recognition, too. There is only one work on few-shot learning for facial landmark localization, namely that of [1] which, different to our approach, proposes an auto-encoder approach for network pre-training. To our knowledge, our evaluation framework provides the very first comprehensive attempt to evaluate the transferability of facial representations for few-shot learning for several face analysis tasks.

Semi-supervised face analysis: Semi-supervised learning has been applied to the domain of Action Unit recognition where data labelling is extremely laborious [83, 85, 84]. Although these methods work with few labels, they are domain specific (as opposed to our work), assuming also that extra annotations are available in terms of "peak" and "valley" frames which is also an expensive operation.

Unsupervised learning: There is a very large number of recently proposed unsupervised/self-supervised learning methods, see for example [74, 8, 80, 43, 25, 11, 9, 12, 22]. To our knowledge, only very few attempts from this line of research have been applied to faces so far. The authors of [72] learn face embeddings in a self-supervised manner by predicting the motion field between two facial images. The authors of [66] propose to combine several facial representations learned using an autoencoding framework. In this work, we explore learning facial representations in an unsupervised manner using the state-of-the-art method of [9] and show how to effectively fine-tune the learned representations to the various face analysis tasks of our benchmark.

3. Method

Supervised deep learning directly applied to large labelled datasets is the de facto approach to solving the most important face analysis tasks. In this section, we propose to take a different path to solving face analysis based on the following 2-stage pipeline: (task agnostic) network pre-training followed by task adaptation. Importantly, we argue that network pre-training should be actually considered as part of the method and not just a simple initialization step. We explore two important aspects of network pre-training: (1) the method used, and (2) the dataset used. Likewise, we highlight hyper-parameter optimization for task adaptation as an absolutely crucial component of the proposed pipeline. Finally, we emphasize the importance of evaluating face analysis on low data regimes, too. We describe important aspects of the pipeline in the following sections.

3.1. Network Pre-training

Supervised pre-training of face networks on ImageNet or VGG datasets is not new. We use these networks as strong baselines. For the first time, we comprehensively evaluate their impact on the most important face analysis tasks.

Unsupervised pre-training: Inspired by [22, 43, 26, 9], we explore, for the first time in literature, large-scale unsupervised learning on facial images to learn a universal, task-agnostic facial representation. To this end, we adopt the recently proposed SwAV [9] which simultaneously clusters the data while enforcing consistency between the cluster assignments produced for different augmentations of the same image. The pretext task is defined as a "swapped" prediction problem where the code of one view is predicted from the representation of another:

$$\mathcal{L}(\mathbf{z}_0, \mathbf{z}_1) = \ell(\mathbf{z}_0, \mathbf{q}_1) + \ell(\mathbf{z}_1, \mathbf{q}_0), \quad (1)$$

where $\mathbf{z}_0, \mathbf{z}_1$ are the features produced by the network for two different views of the same image and $\mathbf{q}_0, \mathbf{q}_1$ their corresponding codes computed by matching these feature using a set of prototypes. ℓ is a cross-entropy (with temperature) loss. The full training details are available in the supplementary material.

3.2. Pre-training Datasets

With pre-training being now an important part of the face analysis pipeline, it is important to investigate what datasets can be used to this end. We argue that supervised pre-training is sub-optimal due to two main reasons: (a) the resulting models may be overly specialized to the source domain and task (*e.g*. face recognition pre-training) or be too generic (*e.g*. ImageNet pre-training), and (b) the amount of labeled data may be limited and/or certain parts of the natural data distribution may not be covered. To alleviate this, for the first time, we propose to explore 4 facial datasets of interest *within unsupervised pre-training*: (a) Large-Scale-Face ($> 5.0M$), (b) Full VGG-Face ($\sim 3.4M$), (c) Small VGG-Face ($\sim 1M$), (d) Flickr-Face ($\sim 1.5M$). We used existing datasets to create Large-Scale-Face. Flickr-Face is an uncurated dataset we created from Flickr images. See also Section 4.

3.3. Facial Task Adaptation

End facial tasks: To draw as safe conclusions as possible, we used a large variety of face tasks (5 in total) including face recognition (classification), facial Action Unit intensity estimation (regression), emotion recognition in terms of valence and arousal (regression), 2D facial landmark localization (pixel-wise regression), and 3D face reconstruction (GCN regression). For these tasks, we used, in total, 10 datasets for evaluation purposes.

Adaptation methods: We are given a pre-trained model on task m, composed of a backbone $g(.)$ and a network head $h^m(.)$. The model follows the ResNet-50 [27] architecture. We considered two widely-used methods for task adaptation: (a) *Network fine-tuning* adapts the weights of $g(.)$ to the new task m_i. The previous head is replaced with a task-specific head $h^{m_i}(.)$ that is trained from scratch. (b) *Linear layer adaptation* keeps the weights of $g(.)$ fixed and trains only the new head $h^{m_i}(.)$. Depending on the task, the structure of the head varies. This will be defined for each task in the corresponding section. See also Section 5.

Hyper-parameter optimization: We find that, without a proper hyper-parameters selection for each task and setting, the produced results are often misleading. In order to alleviate this and ensure a fair comparison, we search for the following optimal hyper-parameters: (a) learning rate, (b) scheduler duration and (c) backbone learning rate for the pre-trained ResNet-50. This search is repeated for *each data point* defined by the tuple (task, dataset, pre-training

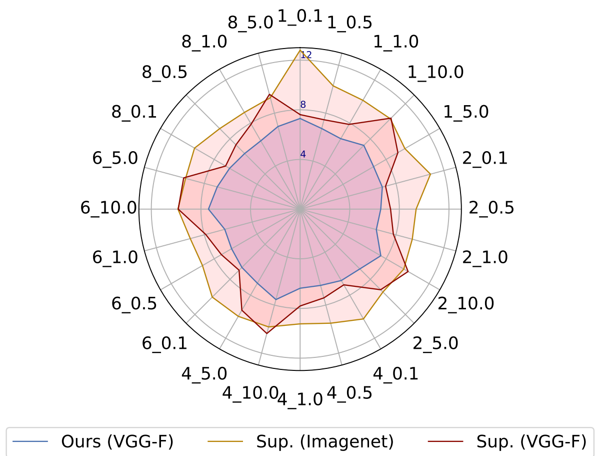

Figure 2: Facial landmark localization accuracy in terms of NME (%) of 3 different pre-training methods for selected combinations of hyperparameters. The labels on the figure's perimeter show the scheduler length (first value) and backbone relative's learning rate (second value) separated by an underscore. Each circle on the radar plot denotes a constant error level. Points located closer to the center correspond to lower error levels. Accuracy greatly varies for different hyperparameters. See also text for details.

method and $\%$ of training data). In total, this yields in an extraordinary number of experiments for discovering the optimal hyperparameters.

Fig. 2 shows the importance of hyperparameters on accuracy for the task of facial landmark localization. In particular, for 1 specific value of learning rate, about 40 different combinations of scheduler duration and backbone relative's learning rate are evaluated. 24 of those combinations are placed on the perimeter of the figure. The 3 closed curves represent the Normalized Mean Error (NME) for each hyperparameter combination for each pre-training method. We observe that accuracy greatly varies for different hyperparameters.

3.4. Few-shot Learning-based Evaluation

We explore, for the first time, evaluating the face models using a varying percentage of training data for each face analysis task. Specifically, beyond the standard evaluation using 100% of training data, we emphasize the importance of the low data regime, in particular 10% and 2%, which has a clear impact when new datasets are to be collected and annotated. The purpose of the proposed evaluation is not only to show which method works the best for this setting but also to draw interesting conclusions about the redundancy of existing facial datasets. See also Section 6.

3.5. Self-distillation for Semi-supervised Learning

The low data regime of the previous section refers to having both few data and few labels. We further propose to investigate the case of semi-supervised learning [35, 76, 75, 12] where a full facial dataset has been collected but only few labels are provided. To this end, we propose a simple self-distillation technique which fully utilizes network pre-training: we use the fine-tuned network to generate in an online manner new labels for training an identically sized student model on unlabeled data. The student is initialized from a pre-trained model trained in a fully unsupervised manner. The self-distillation process is repeated iteratively for T steps, where, at each step, the previously trained model becomes the teacher. Formally, the knowledge transfer is defined as $\mathrm{argmin}_{\theta_t} \mathcal{L}((f(x,\theta_{t-1}), f(x,\theta_t)))$, where x is the input sample, θ_{t-1} and θ_t are the parameters of the teacher and the student, respectively, and $\mathcal{L}$ is the task loss (e.g. pixel-wise ℓ_2 loss for facial landmark localization).

4. Ablation Studies

In this section, we study and answer key questions related to our approach.

Fine-tuning vs. linear adaptation: Our results, provided in the supplementary material, show that linear adaptation results in significant performance degradation. As our ultimate goal is high accuracy for the end facial task, linear adaptation is not considered for the rest of our experiments.

How much facial data are required? Unlike supervised, unsupervised pre-training does not require labels and hence it can be applied easily to all types of combinations of facial datasets. Then, a natural question arising is how much data is needed to learn a high-quality representation. To this end, we used 3 datasets of varying size. The first one, comprising $\sim 3.3M$ images, is the original VGG-Face dataset (VGG-Face). The second comprises $\sim 1M$ images randomly selected from VGGFace2 (VGG-Face-small). The last one, coined as Large-Scale-Face, comprises over 5M images, and is obtained by combining VGG-Face, 300W-LP [86], IMDb-face [67], AffectNet [44] and WiderFace [79]. We trained 3 models on these datasets and evaluated them for the tasks of facial landmark localization, AU intensity estimation and face recognition. As the results from Table 2 show, VGG-Face vs. VGG-Face-small yields small yet noticeable improvements especially for the case

Table 1: Comparison between the facial representations learned by MoCov2 and SwAV, by fine-tuning the models on 2% of 300W and DISFA.

Method	300W	DISFA
	NME	ICC
Scratch	13.5	.237
MoCov2	11.9	.280
SwAV	**4.97**	**.560**
SwAV (256)	5.00	.549

of 2% of labelled data. We did not observe further gains by training on Large-Scale-Face.

Table 2: Impact of different datasets on the facial representations learned *in an unsupervised manner* for the tasks of facial landmark localization (300W), AU intensity estimation (DISFA) and face recognition (IJB-B).

Data amount	**Unsup. Data**	**300W**	**DISFA**	**IJBB**
		NME	ICC	10^{-4}
100%	VGG-Face-small	3.91	.583	0.910
	VGG-Face	3.85	**.598**	**0.912**
	Large-Scale-Face	**3.83**	.593	**0.912**
	Flickr-Face	3.86	.590	0.911
10%	VGG-Face-small	4.37	.572	0.887
	VGG-Face	**4.25**	.592	0.889
	Large-Scale-Face	4.30	**.597**	**0.892**
	Flickr-Face	4.31	.581	0.887
2%	VGG-Face-small	5.46	.550	0.729
	VGG-Face	**4.97**	.560	**0.744**
	Large-Scale-Face	4.98	.551	0.743
	Flickr-Face	5.05	**.571**	0.740

Curated vs. uncurated datasets: While the previous section investigated the quantity of data required, it did not explore the question of data quality. While we did not use any labels during the unsupervised pre-training phase, one may argue that all datasets considered are sanitized as they were collected by human annotators with a specific task in mind. In this section, we go beyond sanitized datasets, by creating a completely uncurated, in-the-wild, dataset, coined Flick-Face, of $\sim 1.5M$ facial images by simply downloading images from Flickr (using standard search keywords like "faces", "humans", etc.) and filtering them with a face detector (the dataset will be made available). We then trained a model on it and evaluated it on the same tasks/datasets of the previous section. Table 2 shows some remarkable results: the resulting model is on par with the one trained on the full VGG-Face dataset (Section 5 shows that it outperforms all other pre-training methods, too). We believe that this result can pave a whole new way to how practitioners, both in industry and academia, collect and label facial datasets for new tasks and applications.

Pre-training task or data? In order to fully understand whether the aforementioned gains are coming from the unsupervised task alone, the data, or both, we pre-trained a model on ImageNet dataset using *both* supervised and unsupervised pre-training. Our experiments showed that both models performed similarly (e.g. 4.97% vs 5.1% on 300W@2% of data) and significantly more poorly than models trained on face datasets. We conclude that *both unsupervised pre-training and data* are required for high accuracy.

Effect of unsupervised method: Herein, we compare the results obtained by changing the unsupervised pre-training method from SwAV to Moco-v2 [26]. Table 1 shows that SwAV largely outperforms Moco-v2, emphasizing the importance of utilizing the most powerful available unsupervised method. Note, that better representation learning as measured on imagenet, doesn't equate with better representation in general [13], hence way it's important to validate the performance of different methods for faces too. Furthermore, we evaluated SwAV models using different batch-sizes which is shown to be an important hyper-parameter. We found both models to perform similarly. See SwAV (256) in Table 1 for the model trained with batch-size 256. With small batch-size training requires less resources, yet we found that it was prolonged by $2\times$.

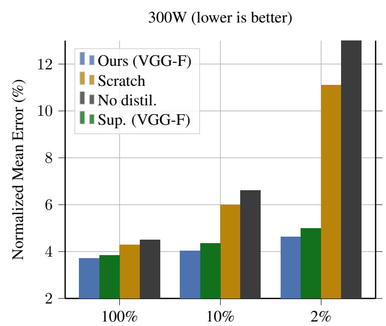

Figure 3: Effectiveness of network pre-training on self-distillation for the tasks of facial landmark localization.

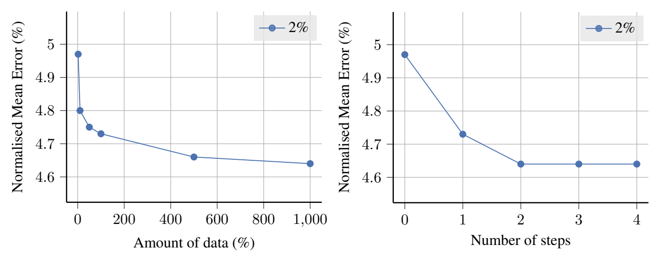

Figure 4: Self-distillation accuracy (for facial landmark localization) vs. (left) amount of unlabeled data (100% corresponds to 300W), and (right) number of distillation steps.

Self-distillation for semi-supervised learning: Herein, we evaluate the effectiveness of network pre-training on self-distillation (see Section 3.5) for the task of semi-supervised facial landmark localization (300W). See supplementary for results on AU intensity estimation (DISFA).

We compare unsupervised vs. supervised pre-training on VGG-Face as well as training from scratch. These networks are fine-tuned on 300W using 100% and, the most interesting, 10% and 2% of the data. Then, they are used as students for self-distillation. Fig. 3 clearly shows the effectiveness of unsupervised student pre-training.

Furthermore, a large pool of unlabelled data was formed by 300W, AFLW [32], WFLW [73] and COFW [4, 20]), and then used for self-distillation. Fig. 4 (left) shows the impact on the accuracy of the final model by adding more and more unlabelled data to the self-distillation process. Clearly, self-distillation based on network pre-training is capable of effectively utilizing a large amount of unlabelled data. Finally, Fig. 4 (right) shows the impact of the number of self-distillation steps on accuracy.

Table 3: Supervised pre-training applied to different datasets. The models are evaluated for AU intensity estimation on DISFA.

Data amount	**Pretrain. method**		
	Sup. (ImageNet)	Sup. (VGG-F)	Sup. (300W)
100%	.560	.575	.463
10%	.556	.560	.460
1%	.453	.542	.414

Other supervised pre-training: Our best supervised pre-trained network is that based on training CosFace [69] on VGG-Face. Herein, for completeness, we compare this to supervised pre-training on another task/dataset, namely facial landmark localization. As Table 3 shows, the supervised pre-trained model on VGG-Face outperforms it by large margin. This is expected due to the massive size of VGG-Face.

5. Main Results

In this section, we thoroughly test the generalizability of the universal facial representations by adapting the resulting models to the most important facial analysis tasks. The full training and implementation details for each of this tasks can be found in the supplementary material. Training code will also be made available.

Data $\&$ label regime: For all datasets and tasks, we used fine-tuning for network adaptation using 3 data and label regimes: full (100%), low (10%) and very low (2% or less).

Models compared: For unsupervised network pre-training, we report the results of two models, one trained on the full VGG-Face and one on Flickr-Face. These models are denoted as Ours (VGG-F) and Ours (Flickr-F). These models are compared with supervised pre-training on ImageNet and VGG-Face (denoted as VGG-F), as well as the model trained from scratch. **Comparison with SOTA:** Where possible, we also present the results reported by state-of-the-art methods for each task on the few-shot setting. Finally, for each task, and, to put our results into perspective, we report the accuracy of a state-of-the-art method for the given task. We note however, that the results are not directly comparable, due to different networks, losses, training procedure, and even training datasets.

5.1. Face Recognition

For face recognition, we fine-tuned the models on the VGGFace [7] and tested them on the IJB-B [71] and IJB-C [42] datasets. The task specific head $h(.)$ consists of a linear layer. The whole network was optimized using the CosFace loss [69]. Note that, for this experiment, since training was done on VGGFace [7], the results of supervised pre-training on VGG-Face are omitted (as meaningless).

Results are shown in Table 4. Both *Ours (VGG-F)* and *Ours (Flickr-F)* perform similarly and both they outperform the other baselines by large margin for the low (10%) and very low (2%) data regimes. For the latter case, the accuracy drops significantly for all cases.

5.2. Facial Landmark Localization

.

We fine-tuned the pre-trained models for facial landmark localization on 300W [56], AFLW-19 [32], WFLW [73] and COFW-68 [4, 20] reporting results in terms of $\text{NME}_{\text{inter-ocular}}$ [56] or NME_{diag} [32]. We followed the current best practices based on heatmap regression [3]. In order to accommodate for the pixel-wise nature of the task, the task specific head $h(.)$ is defined as a set of 3 1×1 conv. layers with 256 channels, each interleaved with bilinear upsampling operations for recovering part of the lost resolution. Additional high resolution information is brought up via skip connections and summation from the lower part of the network. Despite the simple and un-optimized architecture we found that the network performs very well, thanks to the strong facial representation learned. All models were trained using a pixel-wise MSE loss.

Results are shown in Table 5: unsupervised pre-training (both models) outperform the other baselines for all data regimes, especially for the low and very low cases. For the latter case, *Ours (VGG-F)* outperforms *Ours (Flickr-F)* probably because *Ours (VGG-F)* contains a more balanced distribution of facial poses. The best supervised pre-training method is VGG-F showing the importance of pre-training on facial datasets.

Furthermore, Table 6 shows comparison with few very recent works on few-shot face alignment. Our method scores significantly higher across all data regimes and datasets tested setting a new state-of-the-art despite the straightforward network architecture and the generic nature of our method.

5.3. Action Unit (AU) Intensity Estimation

We fine-tuned and evaluated the pre-trained models for AU intensity estimation on the corresponding partitions of BP4D [65, 82] and DISFA [41] datasets. The network head

Table 4: Face recognition results in terms of TAR for various FARs on IJB-B and IJB-C.

Data amount	**Pretrain. method**	IJB-B					IJB-C				
		10^{-6}	10^{-5}	10^{-4}	10^{-3}	10^{-2}	10^{-6}	10^{-5}	10^{-4}	10^{-3}	10^{-2}
100%	Scratch	0.389	0.835	**0.912**	0.950	0.975	0.778	0.883	0.931	0.961	0.981
	Sup. (ImageNet)	0.390	**0.843**	**0.912**	0.950	0.975	0.831	**0.891**	0.931	0.961	0.981
	Ours (Flickr-F)	0.406	0.834	0.911	**0.951**	0.975	0.807	0.880	**0.932**	**0.962**	**0.982**
	Ours (VGG-F)	**0.432**	0.835	**0.912**	0.950	**0.976**	**0.882**	0.882	**0.932**	0.961	0.981
10%	Scratch	0.326	0.645	0.848	0.926	0.965	0.506	0.7671	0.8840	0.940	0.721
	Sup. (ImageNet)	0.320	0.653	0.858	0.926	0.966	0.503	0.779	0.891	0.941	0.973
	Ours (Flickr-F)	0.334	0.758	0.887	0.940	0.970	0.715	0.834	0.909	0.952	**0.978**
	Ours (VGG-F)	**0.392**	**0.784**	**0.889**	**0.941**	**0.972**	**0.733**	**0.847**	**0.911**	**0.953**	0.977
2%	Scratch	0.086	0.479	0.672	0.800	0.909	0.400	0.570	0.706	0.829	0.922
	Sup. (ImageNet)	0.264	0.553	0.694	0.820	0.915	**0.493**	0.599	0.723	0.841	0.928
	Ours (Flickr-F)	0.282	**0.558**	0.740	0.870	0.944	0.486	**0.649**	**0.786**	0.891	0.954
	Ours (VGG-F)	**0.333**	0.547	**0.744**	**0.873**	**0.948**	0.455	0.637	**0.786**	**0.893**	**0.956**
SOTA (from paper) [16]		0.401	0.821	0.907	0.950	0.978	0.0.767	0.879	0.929	0.964	0.984

Table 5: Facial landmark localization results on 300W (test set), COFW, WFLW and AFLW in terms of $\text{NME}_{\text{inter-ocular}}$, except for AFLW where NME_{diag} is used.

Data amount	**Pretrain. method**	**300W**	**COFW**	**WFLW**	**AFLW**
100%	Scratch	4.50	4.10	5.10	1.59
	Sup. (ImageNet)	4.16	3.63	4.80	1.59
	Sup. (VGG-F)	3.97	3.51	4.70	1.58
	Ours (Flickr-F)	3.86	3.45	4.65	1.57
	Ours (VGG-F)	**3.85**	**3.32**	**4.57**	**1.55**
10%	Scratch	6.61	5.63	6.82	1.84
	Sup. (ImageNet)	5.15	5.32	6.56	1.81
	Sup. (VGG-F)	4.55	4.46	5.87	1.77
	Ours (Flickr-F)	4.31	4.27	5.45	**1.73**
	Ours (VGG-F)	**4.25**	**3.95**	**5.44**	1.74
2%	Scratch	13.52	14.7	10.43	2.23
	Sup. (ImageNet)	8.04	8.05	8.99	2.09
	Sup. (VGG-F)	5.45	5.55	6.94	2.00
	Ours (Flickr-F)	5.05	5.18	6.53	**1.86**
	Ours (VGG-F)	**4.97**	**4.70**	**6.29**	1.88
SOTA (from paper) [70]		3.85	3.45	4.60	1.57
SOTA (from paper) [34]		-	-	4.37	1.39

$h(.)$ is implemented using a linear layer. The whole network is trained to regress the intensity value of each AU using an ℓ_2 loss. We report results in terms of intra-class correlation (ICC) [59].

Results are shown in Table 7: unsupervised pre-training (both models) outperform the other baselines for all data regimes. Notably, our models achieve very high accuracy even for the case when 2% of data was used. Supervised pre-training on VGG-F also works well.

Table 6: Comparison against state-of-the-art in few-shot facial landmark localization.

300W	100%	10%	1.5%
RCN+ [28]	3.46	4.47	-
TS^3 [18]	3.49	5.03	-
3FabRec [1]	3.82	4.47	5.10
Ours (VGG-F)	**3.20**	**3.48**	**4.13**
AFLW	100%	10%	1%
RCN+ [28]	1.61	-	2.88
TS^3 [18]	-	2.14	-
3FabRec [1]	1.87	2.03	2.38
Ours (VGG-F)	**1.54**	**1.70**	**1.91**
WFLW	100%	10%	0.7%
SA [50]	4.39	7.20	-
3FabRec [1]	5.62	6.73	8.39
Ours (VGG-F)	**4.57**	**5.44**	**7.11**

Furthermore, Table 6 shows comparison with very recent works on semi-supervised AU intensity estimation. We note that these methods had access to all training data; only the amount of labels was varied. Our methods, although trained under both very low data and label regimes, outperformed them by a significant margin.

5.4. Emotion Recognition

We fine-tuned the models for valence and arousal estimation on the well-established AffectNet [44]. We report results in terms of RMSE and CCC [54] (for other metric see supplementary). The task specific head $h(.)$ is a linear layer that regresses the valence and arousal values and also predicts the basic emotion classes. The network was trained

Table 7: AU intensity estimation results in terms of ICC on BP4D and DISFA.

Data amount	Pretrain. method	DISFA		BP4D	
		finetune	linear	finetune	linear
100%	Scratch	.318	-	.617	-
	Sup. (ImageNet)	.560	.316	.708	.587
	Sup. (VGG-F)	.575	.235	.700	.564
	Ours (Flickr-F)	.590	**.373**	.715	.599
	Ours (VGG-F)	**.598**	.342	**.719**	**.610**
10%	Scratch	.313	-	.622	-
	Sup. (ImageNet)	.556	.300	.698	.573
	Sup. (VGG-F)	.560	.232	.692	.564
	Ours (Flickr-F)	.581	**.352**	.699	.603
	Ours (VGG-F)	**.592**	.340	**.706**	**.604**
1%	Scratch	.237	-	.586	-
	Sup. (ImageNet)	.453	.301	.689	.564
	Sup. (VGG-F)	.542	.187	.690	.562
	Ours (Flickr-F)	**.571**	.321	**.695**	**.596**
	Ours (VGG-F)	.560	**.326**	.694	.592
SOTA (from paper) [46]		0.57	-	0.72	-

Table 8: Comparison against the state-of-the-art in semi-supervised AU intensity estimation. Results for each individual AU can be found in the supplementary material.

Method	Data amount	Dataset	
		DISFA	BP4D
KBSS [83]	1%	.350	.645
KJRE [85]	6%	.350	.600
CLFL [84]	1%	.408	.680
SSCFL [57]	2%	.413	.680
Ours (VGG-F)	1%	**.574**	**.707**

to jointly minimise the RMSE and CCC losses for valence and arousal, and the cross-entropy loss for classification.

Results are shown in Table 9: again, *for all* data regimes, our unsupervised models outperform the supervised baselines.

5.5. 3D Face Reconstruction

We fine-tuned all models on the 300W-LP [86] dataset and tested them on AFLW2000-3D [86]. Our task specific head is implemented with a GCN based on spiral convolutions [38]. The network was trained to minimise the ℓ_1 distance between the predicted and the ground truth vertices.

Results are shown in Table 10: it can be seen that, *for all* data regimes, our unsupervised models outperform the supervised baselines. Supervised pre-training on VGG-F also works well.

Table 9: Valence and arousal estimation results in terms of RMSE and CCC on AffectNet.

Data amount	Pretrain. method	Valence		Arousal	
		RMSE	CCC	RMSE	CCC
100%	Scratch	0.370	0.695	0.339	0.611
	Sup. (ImageNet)	0.360	0.705	0.327	0.620
	Sup. (VGG-F)	0.369	0.706	0.330	0.624
	Ours (VGG-F)	**0.356**	**0.710**	**0.326**	**0.629**
10%	Scratch	0.402	0.625	0.366	0.536
	Sup. (ImageNet)	0.383	0.654	0.351	0.566
	Sup. (VGG-F)	0.401	0.636	0.372	0.532
	Ours (VGG-F)	**0.382**	**0.678**	**0.344**	**0.599**
2%	Scratch	0.453	0.515	0.400	0.422
	Sup. (ImageNet)	0.411	0.557	0.362	0.456
	Sup. (VGG-F)	0.416	**0.607**	0.384	**0.485**
	Ours (VGG-F)	**0.370**	0.593	**0.338**	0.471
SOTA (from paper) [33]		0.35	0.71	0.32	0.63

Table 10: 3D face reconstruction reconstruction in terms of NME (68 points) on AFLW2000-3D.

Data	Pretrain. method				
	Scratch	Sup. (Imagenet)	Sup. (VGG-F)	Ours (Flickr-F)	Ours (VGG-F)
100%	3.70	3.58	3.51	3.53	**3.42**
10%	4.72	4.06	3.82	3.81	**3.72**
2%	7.11	6.15	4.42	4.50	**4.31**
SOTA (from paper) [14]: 3.39					

6. Discussion and Conclusions

Several conclusions can be drawn from our results: Unsupervised pre-training followed by task-specific fine-tuning provides very strong baselines for face analysis. For example, we showed that such generically built baselines outperformed recently proposed methods for few-shot/semi-supervised learning (e.g. for facial landmark localization and AU intensity estimation) some of which are based on quite sophisticated techniques. Moreover, we showed that unsupervised pre-training largely boosts self-distillation. Hence, it might be useful for newly-proposed task-specific methods to consider such a pipeline for both development and evaluation especially when newly-achieved accuracy improvements are to be reported.

Furthermore, these results can be achieved even by simply training on uncurated facial datasets that can be readily downloaded from image repositories. The excellent results obtained by pre-training on Flickr-Face are particularly encouraging. Note that we could have probably created a better and more balanced dataset in terms of facial pose by running a method for facial pose estimation.

When new datasets are to be collected, such powerful pre-trained networks can be potentially used for minimizing data collection and label annotation labour. Our results show that many existing datasets (e.g. AFLW, DISFA, BP4D, even AffectNet) seem to have a large amount of redundancy. This is more evident for video datasets (e.g. DISFA, BP4D).

Note that by no means our results imply or suggest that all face analysis can be solved with small labelled datasets. For example, for face recognition, it was absolutely necessary to fine-tune on the whole VGG-Face in order to get high accuracy.

References

[1] Bjorn Browatzki and Christian Wallraven. 3fabrec: Fast few-shot face alignment by reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6110–6120, 2020. 2, 7

[2] Adrian Bulat and Georgios Tzimiropoulos. Two-stage convolutional part heatmap regression for the 1st 3d face alignment in the wild (3dfaw) challenge. In *European Conference on Computer Vision*, pages 616–624. Springer, 2016. 1

[3] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1021–1030, 2017. 1, 2, 6

[4] Xavier P Burgos-Artizzu, Pietro Perona, and Piotr Dollár. Robust face landmark estimation under occlusion. In *Proceedings of the IEEE international conference on computer vision*, pages 1513–1520, 2013. 6

[5] A Mike Burton, Stephen Wilson, Michelle Cowan, and Vicki Bruce. Face recognition in poor-quality video: Evidence from security surveillance. *Psychological Science*, 10(3):243–248, 1999. 1

[6] Zhaowei Cai, Quanfu Fan, Rogerio S Feris, and Nuno Vasconcelos. A unified multi-scale deep convolutional neural network for fast object detection. In *European conference on computer vision*, pages 354–370. Springer, 2016. 1

[7] 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. 1, 6

[8] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 132–149, 2018. 3

[9] 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. 3

[10] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*, 2014. 2

[11] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv*, 2020. 3

[12] 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. 3, 4

[13] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. *arXiv preprint arXiv:2011.10566*, 2020. 5

[14] Shiyang Cheng, Georgios Tzimiropoulos, Jie Shen, and Maja Pantic. Faster, better and more detailed: 3d face reconstruction with graph convolutional networks. In *Proceedings of the Asian Conference on Computer Vision*, 2020. 8

[15] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-fcn: Object detection via region-based fully convolutional networks. In *NIPS*, 2016. 2

[16] 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. 1, 2, 7

[17] 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. 1

[18] Xuanyi Dong and Yi Yang. Teacher supervises students how to learn from partially labeled images for facial landmark detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 783–792, 2019. 7

[19] Yin Fan, Xiangju Lu, Dian Li, and Yuanliu Liu. Video-based emotion recognition using cnn-rnn and c3d hybrid networks. In *ACM International Conference on Multimodal Interaction*, 2016. 2

[20] Golnaz Ghiasi and Charless C Fowlkes. Occlusion coherence: Detecting and localizing occluded faces. *arXiv preprint arXiv:1506.08347*, 2015. 6

[21] Mislav Grgic, Kresimir Delac, and Sonja Grgic. Scface–surveillance cameras face database. *Multimedia tools and applications*, 51(3):863–879, 2011. 1

[22] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020. 3

[23] Hatice Gunes and Maja Pantic. Automatic, dimensional and continuous emotion recognition. *International Journal of Synthetic Emotions (IJSE)*, 1(1):68–99, 2010. 1

[24] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *ECCV*, 2016. 2

[25] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. *arXiv*, 2019. 3

[26] 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. 3, 5

[27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 1, 3

[28] Sina Honari, Pavlo Molchanov, Stephen Tyree, Pascal Vincent, Christopher Pal, and Jan Kautz. Improving landmark localization with semi-supervised learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1546–1555, 2018. 7

[29] 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. 1

[30] Heysem Kaya, Furkan Gürpınar, and Albert Ali Salah. Video-based emotion recognition in the wild using deep transfer learning and score fusion. *Image and Vision Computing*, 65:66–75, 2017. 2

[31] Boris Knyazev, Roman Shvetsov, Natalia Efremova, and Artem Kuharenko. Convolutional neural networks pretrained on large face recognition datasets for emotion classification from video. *arXiv preprint arXiv:1711.04598*, 2017. 2

[32] Martin Koestinger, Paul Wohlhart, Peter M Roth, and Horst Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *2011 IEEE international conference on computer vision workshops (ICCV workshops)*, pages 2144–2151. IEEE, 2011. 6

[33] Jean Kossaifi, Antoine Toisoul, Adrian Bulat, Yannis Panagakis, Timothy M Hospedales, and Maja Pantic. Factorized higher-order cnns with an application to spatio-temporal emotion estimation. In *CVPR*, 2020. 2, 8

[34] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8236–8246, 2020. 7

[35] Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, 2013. 4

[36] Jian Li, Yabiao Wang, Changan Wang, Ying Tai, Jianjun Qian, Jian Yang, Chengjie Wang, Jilin Li, and Feiyue Huang. Dsfd: dual shot face detector. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5060–5069, 2019. 1

[37] Shan Li and Weihong Deng. Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing*, 2020. 2

[38] Isaak Lim, Alexander Dielen, Marcel Campen, and Leif Kobbelt. A simple approach to intrinsic correspondence learning on unstructured 3d meshes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018. 8

[39] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 1

[40] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens van der Maaten. Exploring the limits of weakly supervised pretraining. In *ECCV*, 2018. 2

[41] S Mohammad Mavadati, Mohammad H Mahoor, Kevin Bartlett, Philip Trinh, and Jeffrey F Cohn. Disfa: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, 4(2):151–160, 2013. 6

[42] Brianna Maze, Jocelyn Adams, James A Duncan, Nathan Kalka, Tim Miller, Charles Otto, Anil K Jain, W Tyler Niggel, Janet Anderson, Jordan Cheney, et al. Iarpa janus benchmark-c: Face dataset and protocol. In *IEEE International Conference on Biometrics (ICB)*, 2018. 6

[43] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. *arXiv*, 2019. 3

[44] 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, 2017. 2, 4, 7

[45] Hong-Wei Ng, Viet Dung Nguyen, Vassilios Vonikakis, and Stefan Winkler. Deep learning for emotion recognition on small datasets using transfer learning. In *Proceedings of the 2015 ACM on international conference on multimodal interaction*, pages 443–449, 2015. 2

[46] Ioanna Ntinou, Enrique Sanchez, Adrian Bulat, Michel Valstar, and Georgios Tzimiropoulos. A transfer learning approach to heatmap regression for action unit intensity estimation. *IEEE Transactions on Affective Computing*, 2021. 8

[47] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. 2015. 2

[48] Aleksandr Parkin and Oleg Grinchuk. Recognizing multimodal face spoofing with face recognition networks. In *CVPR-W*, 2019. 2

[49] P Jonathon Phillips, Amy N Yates, Ying Hu, Carina A Hahn, Eilidh Noyes, Kelsey Jackson, Jacqueline G Cavazos, Géraldine Jeckeln, Rajeev Ranjan, Swami Sankaranarayanan, et al. Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. *Proceedings of the National Academy of Sciences*, 115(24):6171–6176, 2018. 1

[50] 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 International Conference on Computer Vision*, pages 10153–10163, 2019. 7

[51] Rajeev Ranjan, Vishal M Patel, and Rama Chellappa. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *IEEE TPAMI*, 41(1):121–135, 2017. 2

[52] Rajeev Ranjan, Swami Sankaranarayanan, Carlos D Castillo, and Rama Chellappa. An all-in-one convolutional neural network for face analysis. In *IEEE FG 2017*, 2017. 2

[53] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 1, 2

[54] Fabien Ringeval, Björn Schuller, Michel Valstar, Roddy Cowie, and Maja Pantic. Avec 2015: The 5th international audio/visual emotion challenge and workshop. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1335–1336, 2015. 7

[55] Joseph P Robinson, Gennady Livitz, Yann Henon, Can Qin, Yun Fu, and Samson Timoner. Face recognition: too bias, or not too bias? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–1, 2020. 1

[56] 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. 6

[57] Enrique Sanchez, Adrian Bulat, Anestis Zaganidis, and Georgios Tzimiropoulos. Semi-supervised au intensity estimation with contrastive learning. *arXiv preprint arXiv:2011.01864*, 2020. 8

[58] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 1

[59] Patrick E Shrout and Joseph L Fleiss. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 86(2):420, 1979. 7

[60] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1

[61] Tomáš Sixta, Julio Junior, CS Jacques, Pau Buch-Cardona, Eduard Vazquez, and Sergio Escalera. Fairface challenge at eccv 2020: Analyzing bias in face recognition. *arXiv preprint arXiv:2009.07838*, 2020. 1

[62] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1701–1708, 2014. 1

[63] Panagiotis Tzirakis, George Trigeorgis, Mihalis A Nicolaou, Björn W Schuller, and Stefanos Zafeiriou. End-to-end multimodal emotion recognition using deep neural networks. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1301–1309, 2017. 1

[64] Michel Valstar, Jonathan Gratch, Björn Schuller, Fabien Ringeval, Denis Lalanne, Mercedes Torres Torres, Stefan Scherer, Giota Stratou, Roddy Cowie, and Maja Pantic. Avec 2016: Depression, mood, and emotion recognition workshop and challenge. In *Proceedings of the 6th international workshop on audio/visual emotion challenge*, pages 3–10, 2016. 1

[65] Michel F Valstar, Timur Almaev, Jeffrey M Girard, Gary McKeown, Marc Mehu, Lijun Yin, Maja Pantic, and Jeffrey F Cohn. Fera 2015-second facial expression recognition and analysis challenge. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 6, pages 1–8. IEEE, 2015. 6

[66] Valentin Vielzeuf, Alexis Lechervy, Stéphane Pateux, and Frédéric Jurie. Towards a general model of knowledge for facial analysis by multi-source transfer learning. 2020. 3

[67] Fei Wang, Liren Chen, Cheng Li, Shiyao Huang, Yanjie Chen, Chen Qian, and Chen Change Loy. The devil of face recognition is in the noise. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 765–780, 2018. 4

[68] Feng Wang, Jian Cheng, Weiyang Liu, and Haijun Liu. Additive margin softmax for face verification. *IEEE Signal Processing Letters*, 25(7):926–930, 2018. 1

[69] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5265–5274, 2018. 1, 6

[70] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 2020. 1, 2, 7

[71] Cameron Whitelam, Emma Taborsky, Austin Blanton, Brianna Maze, Jocelyn Adams, Tim Miller, Nathan Kalka, Anil K Jain, James A Duncan, Kristen Allen, et al. Iarpa janus benchmark-b face dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 90–98, 2017. 6

[72] Olivia Wiles, A Koepke, and Andrew Zisserman. Self-supervised learning of a facial attribute embedding from video. In *BMVC*, 2018. 3

[73] Wayne Wu, Chen Qian, Shuo Yang, Quan Wang, Yici Cai, and Qiang Zhou. Look at boundary: A boundary-aware face alignment algorithm. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2129–2138, 2018. 6

[74] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3733–3742, 2018. 3

[75] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10687–10698, 2020. 4

[76] I Zeki Yalniz, Hervé Jégou, Kan Chen, Manohar Paluri, and Dhruv Mahajan. Billion-scale semi-supervised learning for image classification. *arXiv preprint arXiv:1905.00546*, 2019. 4

[77] Jing Yang, Adrian Bulat, and Georgios Tzimiropoulos. Fan-face: a simple orthogonal improvement to deep face recognition. In *AAAI*, pages 12621–12628, 2020. 1

[78] Jiaolong Yang, Peiran Ren, Dongqing Zhang, Dong Chen, Fang Wen, Hongdong Li, and Gang Hua. Neural aggregation network for video face recognition. In *CVPR*, 2017. 2

[79] Shuo Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Wider face: A face detection benchmark. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 4

[80] Mang Ye, Xu Zhang, Pong C Yuen, and Shih-Fu Chang. Unsupervised embedding learning via invariant and spreading instance feature. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6210–6219, 2019. 3

[81] Shifeng Zhang, Xiangyu Zhu, Zhen Lei, Hailin Shi, Xiaobo Wang, and Stan Z Li. S3fd: Single shot scale-invariant face detector. In *Proceedings of the IEEE international conference on computer vision*, pages 192–201, 2017. 1

[82] Xing Zhang, Lijun Yin, Jeffrey F Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, Peng Liu, and Jeffrey M Girard. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, 32(10):692–706, 2014. 6

[83] Yong Zhang, Weiming Dong, Bao-Gang Hu, and Qiang Ji. Weakly-supervised deep convolutional neural network learning for facial action unit intensity estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2314–2323, 2018. 2, 8

[84] Yong Zhang, Haiyong Jiang, Baoyuan Wu, Yanbo Fan, and Qiang Ji. Context-aware feature and label fusion for facial action unit intensity estimation with partially labeled data. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 733–742, 2019. 2, 8

[85] Yong Zhang, Baoyuan Wu, Weiming Dong, Zhifeng Li, Wei Liu, Bao-Gang Hu, and Qiang Ji. Joint representation and estimator learning for facial action unit intensity estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3457–3466, 2019. 2, 8

[86] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z Li. Face alignment across large poses: A 3d solution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 146–155, 2016. 4, 8

[87] Xiangyu Zhu, Xiaoming Liu, Zhen Lei, and Stan Z Li. Face alignment in full pose range: A 3d total solution. *IEEE transactions on pattern analysis and machine intelligence*, 41(1):78–92, 2017. 1