Pre-training strategies and datasets for facial representation learning

Adrian Bulat[1][0000−0002−3185−4979], Shiyang Cheng[1], Jing Yang[2][0000−0002−8794−4842], Andrew Garbett[1], Enrique Sanchez[1][0000−0003−0196−922X], and Georgios Tzimiropoulos[1,3][0000−0002−1803−5338]

[1] Samsung AI Cambridge
adrian@adrianbulat.com, {shiyang.c,a.garbett}@samsung.com,
kike.sanc@gmail.com
[2] University of Nottingham, Nottingham, UK
jing.yang2@nottingham.ac.uk
[3] Queen Mary University London, London, UK
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, and pre-trained models to facilitate future research.

Keywords: Face recognition, face alignment, emotion recognition, 3D face reconstruction, representation learning

1 Introduction

Supervised learning with Deep Neural Networks has been the standard approach to solving several Computer Vision problems over the recent past years [28, 57, 65, 30, 41]. Among others, this approach has been very successfully applied to

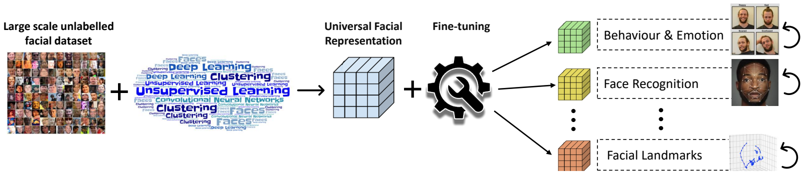


Fig. 1: We advocate for a new paradigm to solving face analysis based on the following pipeline: (1) collection of large-scale unlabelled facial dataset, (2) (task agnostic) network pre-training for universal facial representation learning, and (3) facial task-specific fine-tuning. **Our main result** 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.

several face analysis tasks including face detection [7, 88, 18, 38], recognition [63, 74, 75, 84, 17] and landmark localization [3, 4, 94, 76]. 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 [53, 68]. 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 [8, 4].

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 [24, 69, 70] and surveillance [6, 22], or new considerations of existing problems like fair face recognition [59, 66]. Secondly, from a methodological point of view, it is unsatisfactory for each application to require its own large-scale dataset, although there is only one object of interest - 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 this, **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 recognition, 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 Flickr.
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 and pre-trained models 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 [57, 16, 11]. Because most recent face-related works are based on the collection of larger and larger facial datasets [25, 4, 47], 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 [48], face anti-spoofing [51] and facial landmark localization [76]. Furthermore, the VGG-Face [50] or other large face datasets (e.g. [47]) have been identified as better alternatives by several works, see for example [20, 31, 83, 33, 56, 55, 51, 35, 39]. 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 [25, 4, 47]. Also it is known that label noise can severely impact accuracy (e.g. see Table 6 of [17]). Beyond faces, the work of [43] 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. [25]). 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 [2] 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 [90, 92, 91]. 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 [80, 9, 86, 46, 26, 12, 10, 13, 23]. To our knowledge, only very few attempts from this line of research have been applied to faces so far. The authors of [78] learn face embeddings in a self-supervised manner by predicting the motion field between two facial images. The authors of [72] 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 [10] 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 [23, 46, 27, 10], 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 [10] 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: $\mathcal{L}(\mathbf{z}_0, \mathbf{z}_1) = \ell(\mathbf{z}_0, \mathbf{q}_1) + \ell(\mathbf{z}_1, \mathbf{q}_0)$, 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. See supplementary material for training details.

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 large scale unsupervised pre-training on 4 facial datasets of interest, under two settings: using curated and uncurated data. The later departs from the common paradigm that uses carefully collected data that already includes some forms of explicit annotations and post-processing. In contrast, in the later case, all acquired facial images are used.

Curated Datasets For unsupervised pre-training we explore 3 curated datasets, collected for various facial analysis tasks: (a) Full VGG-Face ($\sim 3.4M$), (b) Small VGG-Face ($\sim 1M$) and (c) Large-Scale-Face ($> 5.0M$), consisting of VGG-Face2 [8], 300W-LP [93], IMDb-face [73], AffectNet [47] and WiderFace [85]. During unsupervised pre-training we drop all labels using only the facial images. See supplementary material for more details.

Uncurated Datasets For a more realistic and practical scenario, we go beyond sanitized datasets, by creating a completely uncurated, in-the-wild, dataset, coined Flickr-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 [18] (the dataset will be made available). In total we collected 1.793.119 facial images. For more details, see supp. material.

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.

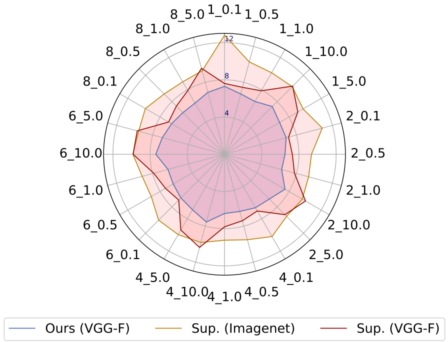


Fig. 2: Facial landmark localization accuracy in terms of NME (%) of 3 different pre-training methods for selected combinations of hyperparameters. The labels on the figure's perimeter show the scheduler length (first value) and backbone relative's learning rate (second value) separated by an underscore. 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.

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 [28] 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 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.

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

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 [37, 82, 81, 13] 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 Table 7, 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 is 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 [93], IMDb-face [73], AffectNet [47] and WiderFace [85]. For more details regarding the datasets see Section 3.2. 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 of 2% of labelled data. We did not observe further gains by training on Large-Scale-Face.

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 experimenting with the newly completely uncurated, in-the-wild, dataset, coined Flickr-Face, introduced in Section 3.2.

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

We 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 [27]. 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 [14], 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×.

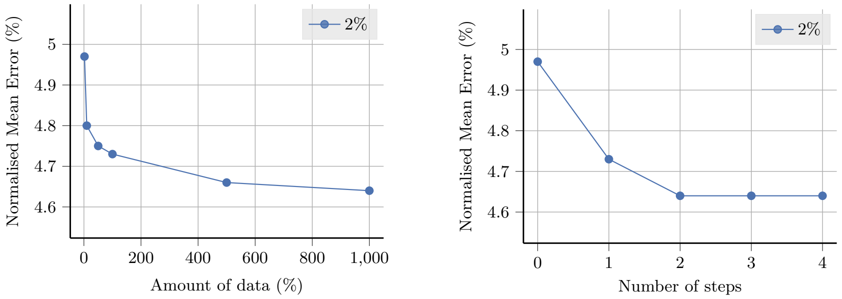


Fig. 3: Self-distillation accuracy for facial landmark vs. (left) amount of unlabeled data (100% corresponds to 300W), and (right) number of distillation steps.

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).

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. 4 clearly shows the effectiveness of unsupervised student pre-training.

Furthermore, a large pool of unlabelled data was formed by 300W, AFLW [34], WFLW [79] and COFW [5, 21]), and then used for self-distillation. Fig. 3 (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. 3 (right) shows the impact of the number of self-distillation steps on accuracy.

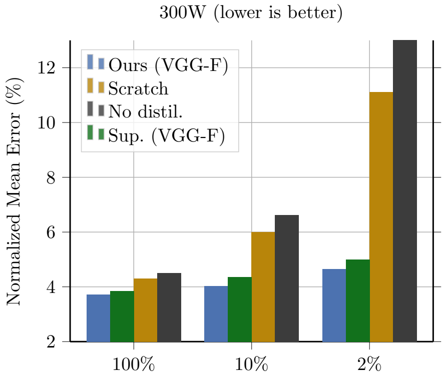


Fig. 4: Effectiveness of network pre-training on self-distillation for the tasks of facial landmark localization.

Other supervised pre-training: Our best supervised pre-trained network is that based on training CosFace [75] 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.

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

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 is detailed in the corresponding sub-section. Training code will 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). For all low data scenarios, we randomly sub-sampled a set of annotated images without accounting for the labels (i.e. we don't attempt to balance the classes). Once formed, the same subset is used for all subsequent experiments to avoid noise induced by different sets of images. For face recognition, we deviated slightly from the above setting by enforcing that at least 1/4 of the identities are preserved for the very low data regime of 2%. This is a consequence of the training objective used for face recognition that is sensitive to both the number of identities and samples per identity.

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 [8] and tested them on the IJB-B [77] and IJB-C [45] datasets. The task specific head $h(.)$

Table 4: Face recognition results in terms of TAR 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) [17]		0.401	0.821	0.907	0.950	0.978	0.0.767	0.879	0.929	0.964	0.984

consists of a linear layer. The whole network was optimized using the CosFace loss [75]. Note that, for this experiment, since training was done on VGGFace [8], the results of supervised pre-training on VGG-Face are omitted (as meaningless). For training details, see supplementary material.

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 [60], AFLW-19 [34], WFLW [79] and COFW-68 [5, 21] reporting results in terms of $\text{NME}_{\text{i-o}}$ [60] or NME_{diag} [34]. We followed the current best practices based on heatmap regression [4]. 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. For full training details, see supp. material.

Table 5: Comparison against state-of-the-art in few-shot facial landmark localization.

300W	100%	10%	1.5%
RCN+ [29]	3.46	4.47	-
TS^3 [19]	3.49	5.03	-
3FabRec [2]	3.82	4.47	5.10
Ours (VGG-F)	**3.20**	**3.48**	**4.13**
AFLW	100%	10%	1%
RCN+ [29]	1.61	-	2.88
TS^3 [19]	-	2.14	-
3FabRec [2]	1.87	2.03	2.38
Ours (VGG-F)	**1.54**	**1.70**	**1.91**
WFLW	100%	10%	0.7%
SA [54]	4.39	7.20	-
3FabRec [2]	5.62	6.73	8.39
Ours (VGG-F)	**4.57**	**5.44**	**7.11**

Table 6: Facial landmark localization results on 300W (test set), COFW, WFLW and AFLW in terms of NME$_{\text{inter-ocular}}$, except for AFLW where NME$_{\text{diag}}$ 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) [76]		3.85	3.45	4.60	1.57
SOTA (from paper) [36]		-	-	4.37	1.39

Results are shown in Table 6: 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 5 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 [71, 89] and DISFA [44] datasets. The network head $h(.)$ 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) [64]. For training details, see supplementary material.
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.

Furthermore, Table 8 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.

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) [49]		0.57	-	0.72	-

Table 8: Comparison against state-of-the-art on few-shot Facial AU intensity estimation on the BP4D dataset.

Method	**Data amount**	**AU**					**Avg.**
		6	**10**	**12**	**14**	**17**	
KBSS [90]	1%	.760	.725	.840	.445	.454	.645
KJRE [92]	6%	.710	.610	.870	.390	.420	.600
CLFL [91]	1%	.766	.703	.827	.411	**.600**	.680
SSCFL [62]	2%	.766	.749	.857	.475	.553	.680
Ours	**1%**	**.789**	**.756**	**.882**	**.529**	.578	**.707**

5.4 Emotion Recognition

We observe similar behaviour on the well-established AffectNet [47] for emotion recognition. For details and results, see supplementary material.

5.5 3D Face Reconstruction

We fine-tuned all models on the 300W-LP [93] dataset and tested them on AFLW2000-3D [93]. Our task specific head is implemented with a GCN based on spiral convolutions [40]. The network was trained to minimise the ℓ_1 distance between the predicted and the ground truth vertices.

Training details: Since 300W-LP has a small number of identities, during training we randomly augment the data using the following transformations: scaling($0.85\times -1.15\times$), in-plane rotation ($\pm 45^o$), and random 10% translation w.r.t image width and height. Depending on the setting, we trained the model

Table 9: 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) [15]: 3.39					

between 120 and 360 epochs using a learning rate of 0.05, a weight decay of 10^{-4} and SGD with momentum (set to 0.9). All models were trained using 2 GPUs. **Results** are shown in Table 9: 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. For more results, see supplementary material.

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. Bagdanov, A.D., Del Bimbo, A., Masi, I.: The florence 2d/3d hybrid face dataset. In: Proceedings of the 2011 joint ACM workshop on Human gesture and behavior understanding. pp. 79–80 (2011)
2. Browatzki, B., Wallraven, C.: 3fabrec: Fast few-shot face alignment by reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6110–6120 (2020)
3. Bulat, A., Tzimiropoulos, G.: Two-stage convolutional part heatmap regression for the 1st 3d face alignment in the wild (3dfaw) challenge. In: European Conference on Computer Vision. pp. 616–624. Springer (2016)
4. Bulat, A., Tzimiropoulos, G.: 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. pp. 1021–1030 (2017)
5. Burgos-Artizzu, X.P., Perona, P., Dollár, P.: Robust face landmark estimation under occlusion. In: Proceedings of the IEEE international conference on computer vision. pp. 1513–1520 (2013)
6. Burton, A.M., Wilson, S., Cowan, M., Bruce, V.: Face recognition in poor-quality video: Evidence from security surveillance. Psychological Science **10**(3), 243–248 (1999)
7. Cai, Z., Fan, Q., Feris, R.S., Vasconcelos, N.: A unified multi-scale deep convolutional neural network for fast object detection. In: European conference on computer vision. pp. 354–370. Springer (2016)
8. Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: Vggface2: A dataset for recognising faces across pose and age. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). pp. 67–74. IEEE (2018)
9. Caron, M., Bojanowski, P., Joulin, A., Douze, M.: Deep clustering for unsupervised learning of visual features. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 132–149 (2018)
10. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. arXiv preprint arXiv:2006.09882 (2020)
11. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Semantic image segmentation with deep convolutional nets and fully connected crfs. arXiv preprint arXiv:1412.7062 (2014)
12. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. arXiv (2020)
13. Chen, T., Kornblith, S., Swersky, K., Norouzi, M., Hinton, G.: Big self-supervised models are strong semi-supervised learners. arXiv preprint arXiv:2006.10029 (2020)
14. Chen, X., He, K.: Exploring simple siamese representation learning. arXiv preprint arXiv:2011.10566 (2020)
15. Cheng, S., Tzimiropoulos, G., Shen, J., Pantic, M.: Faster, better and more detailed: 3d face reconstruction with graph convolutional networks. In: Proceedings of the Asian Conference on Computer Vision (2020)
16. Dai, J., Li, Y., He, K., Sun, J.: R-fcn: Object detection via region-based fully convolutional networks. In: NIPS (2016)
17. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4690–4699 (2019)

18. Deng, J., Guo, J., Zhou, Y., Yu, J., Kotsia, I., Zafeiriou, S.: Retinaface: Single-stage dense face localisation in the wild. arXiv preprint arXiv:1905.00641 (2019)
19. Dong, X., Yang, Y.: Teacher supervises students how to learn from partially labeled images for facial landmark detection. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 783–792 (2019)
20. Fan, Y., Lu, X., Li, D., Liu, Y.: Video-based emotion recognition using cnn-rnn and c3d hybrid networks. In: ACM International Conference on Multimodal Interaction (2016)
21. Ghiasi, G., Fowlkes, C.C.: Occlusion coherence: Detecting and localizing occluded faces. arXiv preprint arXiv:1506.08347 (2015)
22. Grgic, M., Delac, K., Grgic, S.: Scface–surveillance cameras face database. Multimedia tools and applications **51**(3), 863–879 (2011)
23. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P.H., Buchatskaya, E., Doersch, C., Pires, B.A., Guo, Z.D., Azar, M.G., et al.: Bootstrap your own latent: A new approach to self-supervised learning. arXiv preprint arXiv:2006.07733 (2020)
24. Gunes, H., Pantic, M.: Automatic, dimensional and continuous emotion recognition. International Journal of Synthetic Emotions (IJSE) **1**(1), 68–99 (2010)
25. Guo, Y., Zhang, L., Hu, Y., He, X., Gao, J.: Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In: ECCV (2016)
26. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. arXiv (2019)
27. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9729–9738 (2020)
28. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
29. Honari, S., Molchanov, P., Tyree, S., Vincent, P., Pal, C., Kautz, J.: Improving landmark localization with semi-supervised learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1546–1555 (2018)
30. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–4708 (2017)
31. Kaya, H., Gürpınar, F., Salah, A.A.: Video-based emotion recognition in the wild using deep transfer learning and score fusion. Image and Vision Computing **65**, 66–75 (2017)
32. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
33. Knyazev, B., Shvetsov, R., Efremova, N., Kuharenko, A.: Convolutional neural networks pretrained on large face recognition datasets for emotion classification from video. arXiv preprint arXiv:1711.04598 (2017)
34. Koestinger, M., Wohlhart, P., Roth, P.M., Bischof, H.: 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). pp. 2144–2151. IEEE (2011)
35. Kossaifi, J., Toisoul, A., Bulat, A., Panagakis, Y., Hospedales, T.M., Pantic, M.: Factorized higher-order cnns with an application to spatio-temporal emotion estimation. In: CVPR (2020)
36. Kumar, A., Marks, T.K., Mou, W., Wang, Y., Jones, M., Cherian, A., Koike-Akino, T., Liu, X., Feng, C.: Luvli face alignment: Estimating landmarks' location, uncertainty, and visibility likelihood. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8236–8246 (2020)

37. Lee, D.H.: Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In: Workshop on challenges in representation learning, ICML. vol. 3 (2013)
38. Li, J., Wang, Y., Wang, C., Tai, Y., Qian, J., Yang, J., Wang, C., Li, J., Huang, F.: Dsfd: dual shot face detector. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5060–5069 (2019)
39. Li, S., Deng, W.: Deep facial expression recognition: A survey. IEEE Transactions on Affective Computing (2020)
40. Lim, I., Dielen, A., Campen, M., Kobbelt, L.: A simple approach to intrinsic correspondence learning on unstructured 3d meshes. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 0–0 (2018)
41. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3431–3440 (2015)
42. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016)
43. Mahajan, D., Girshick, R., Ramanathan, V., He, K., Paluri, M., Li, Y., Bharambe, A., van der Maaten, L.: Exploring the limits of weakly supervised pretraining. In: ECCV (2018)
44. Mavadati, S.M., Mahoor, M.H., Bartlett, K., Trinh, P., Cohn, J.F.: Disfa: A spontaneous facial action intensity database. IEEE Transactions on Affective Computing **4**(2), 151–160 (2013)
45. Maze, B., Adams, J., Duncan, J.A., Kalka, N., Miller, T., Otto, C., Jain, A.K., Niggel, W.T., Anderson, J., Cheney, J., et al.: Iarpa janus benchmark-c: Face dataset and protocol. In: IEEE International Conference on Biometrics (ICB) (2018)
46. Misra, I., van der Maaten, L.: Self-supervised learning of pretext-invariant representations. arXiv (2019)
47. Mollahosseini, A., Hasani, B., Mahoor, M.H.: Affectnet: A database for facial expression, valence, and arousal computing in the wild. IEEE Transactions on Affective Computing **10**(1), 18–31 (2017)
48. Ng, H.W., Nguyen, V.D., Vonikakis, V., Winkler, S.: Deep learning for emotion recognition on small datasets using transfer learning. In: Proceedings of the 2015 ACM on international conference on multimodal interaction. pp. 443–449 (2015)
49. Ntinou, I., Sanchez, E., Bulat, A., Valstar, M., Tzimiropoulos, G.: A transfer learning approach to heatmap regression for action unit intensity estimation. IEEE Transactions on Affective Computing (2021)
50. Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition (2015)
51. Parkin, A., Grinchuk, O.: Recognizing multi-modal face spoofing with face recognition networks. In: CVPR-W (2019)
52. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems 32, pp. 8024–8035. Curran Associates, Inc. (2019), http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf
53. Phillips, P.J., Yates, A.N., Hu, Y., Hahn, C.A., Noyes, E., Jackson, K., Cavazos, J.G., Jeckeln, G., Ranjan, R., Sankaranarayanan, S., et al.: Face recognition ac-

curacy of forensic examiners, superrecognizers, and face recognition algorithms. Proceedings of the National Academy of Sciences **115**(24), 6171–6176 (2018)

54. Qian, S., Sun, K., Wu, W., Qian, C., Jia, J.: Aggregation via separation: Boosting facial landmark detector with semi-supervised style translation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 10153–10163 (2019)
55. Ranjan, R., Patel, V.M., Chellappa, R.: Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. IEEE TPAMI **41**(1), 121–135 (2017)
56. Ranjan, R., Sankaranarayanan, S., Castillo, C.D., Chellappa, R.: An all-in-one convolutional neural network for face analysis. In: IEEE FG 2017 (2017)
57. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in neural information processing systems. pp. 91–99 (2015)
58. Ringeval, F., Schuller, B., Valstar, M., Cowie, R., Pantic, M.: Avec 2015: The 5th international audio/visual emotion challenge and workshop. In: Proceedings of the 23rd ACM international conference on Multimedia. pp. 1335–1336 (2015)
59. Robinson, J.P., Livitz, G., Henon, Y., Qin, C., Fu, Y., Timoner, S.: Face recognition: too bias, or not too bias? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 0–1 (2020)
60. Sagonas, C., Antonakos, E., Tzimiropoulos, G., Zafeiriou, S., Pantic, M.: 300 faces in-the-wild challenge: Database and results. Image and vision computing **47**, 3–18 (2016)
61. Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., Pantic, M.: 300 faces in-the-wild challenge: The first facial landmark localization challenge. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. pp. 397–403 (2013)
62. Sanchez, E., Bulat, A., Zaganidis, A., Tzimiropoulos, G.: Semi-supervised au intensity estimation with contrastive learning. arXiv preprint arXiv:2011.01864 (2020)
63. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 815–823 (2015)
64. Shrout, P.E., Fleiss, J.L.: Intraclass correlations: uses in assessing rater reliability. Psychological bulletin **86**(2), 420 (1979)
65. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
66. Sixta, T., Junior, J., Jacques, C., Buch-Cardona, P., Vazquez, E., Escalera, S.: Fairface challenge at eccv 2020: Analyzing bias in face recognition. arXiv preprint arXiv:2009.07838 (2020)
67. Sun, K., Zhao, Y., Jiang, B., Cheng, T., Xiao, B., Liu, D., Mu, Y., Wang, X., Liu, W., Wang, J.: High-resolution representations for labeling pixels and regions. arXiv preprint arXiv:1904.04514 (2019)
68. Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: Deepface: Closing the gap to human-level performance in face verification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1701–1708 (2014)
69. Tzirakis, P., Trigeorgis, G., Nicolaou, M.A., Schuller, B.W., Zafeiriou, S.: End-to-end multimodal emotion recognition using deep neural networks. IEEE Journal of Selected Topics in Signal Processing **11**(8), 1301–1309 (2017)
70. Valstar, M., Gratch, J., Schuller, B., Ringeval, F., Lalanne, D., Torres Torres, M., Scherer, S., Stratou, G., Cowie, R., Pantic, M.: Avec 2016: Depression, mood, and emotion recognition workshop and challenge. In: Proceedings of the 6th international workshop on audio/visual emotion challenge. pp. 3–10 (2016)

71. Valstar, M.F., Almaev, T., Girard, J.M., McKeown, G., Mehu, M., Yin, L., Pantic, M., Cohn, J.F.: Fera 2015-second facial expression recognition and analysis challenge. In: 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG). vol. 6, pp. 1–8. IEEE (2015)
72. Vielzeuf, V., Lechervy, A., Pateux, S., Jurie, F.: Towards a general model of knowledge for facial analysis by multi-source transfer learning (2020)
73. Wang, F., Chen, L., Li, C., Huang, S., Chen, Y., Qian, C., Change Loy, C.: The devil of face recognition is in the noise. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 765–780 (2018)
74. Wang, F., Cheng, J., Liu, W., Liu, H.: Additive margin softmax for face verification. IEEE Signal Processing Letters **25**(7), 926–930 (2018)
75. Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z., Liu, W.: Cosface: Large margin cosine loss for deep face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5265–5274 (2018)
76. Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., et al.: Deep high-resolution representation learning for visual recognition. IEEE transactions on pattern analysis and machine intelligence (2020)
77. Whitelam, C., Taborsky, E., Blanton, A., Maze, B., Adams, J., Miller, T., Kalka, N., Jain, A.K., Duncan, J.A., Allen, K., et al.: Iarpa janus benchmark-b face dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 90–98 (2017)
78. Wiles, O., Koepke, A., Zisserman, A.: Self-supervised learning of a facial attribute embedding from video. In: BMVC (2018)
79. Wu, W., Qian, C., Yang, S., Wang, Q., Cai, Y., Zhou, Q.: Look at boundary: A boundary-aware face alignment algorithm. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2129–2138 (2018)
80. Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised feature learning via non-parametric instance discrimination. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3733–3742 (2018)
81. Xie, Q., Luong, M.T., Hovy, E., Le, Q.V.: Self-training with noisy student improves imagenet classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10687–10698 (2020)
82. Yalniz, I.Z., Jégou, H., Chen, K., Paluri, M., Mahajan, D.: Billion-scale semi-supervised learning for image classification. arXiv preprint arXiv:1905.00546 (2019)
83. Yang, J., Ren, P., Zhang, D., Chen, D., Wen, F., Li, H., Hua, G.: Neural aggregation network for video face recognition. In: CVPR (2017)
84. Yang, J., Bulat, A., Tzimiropoulos, G.: Fan-face: a simple orthogonal improvement to deep face recognition. In: AAAI. pp. 12621–12628 (2020)
85. Yang, S., Luo, P., Loy, C.C., Tang, X.: Wider face: A face detection benchmark. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
86. Ye, M., Zhang, X., Yuen, P.C., Chang, S.F.: Unsupervised embedding learning via invariant and spreading instance feature. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6210–6219 (2019)
87. You, Y., Gitman, I., Ginsburg, B.: Large batch training of convolutional networks. arXiv preprint arXiv:1708.03888 (2017)
88. Zhang, S., Zhu, X., Lei, Z., Shi, H., Wang, X., Li, S.Z.: S3fd: Single shot scale-invariant face detector. In: Proceedings of the IEEE international conference on computer vision. pp. 192–201 (2017)
89. Zhang, X., Yin, L., Cohn, J.F., Canavan, S., Reale, M., Horowitz, A., Liu, P., Girard, J.M.: Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. Image and Vision Computing **32**(10), 692–706 (2014)

90. Zhang, Y., Dong, W., Hu, B.G., Ji, Q.: 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. pp. 2314–2323 (2018)
91. Zhang, Y., Jiang, H., Wu, B., Fan, Y., Ji, Q.: 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. pp. 733–742 (2019)
92. Zhang, Y., Wu, B., Dong, W., Li, Z., Liu, W., Hu, B.G., Ji, Q.: Joint representation and estimator learning for facial action unit intensity estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3457–3466 (2019)
93. Zhu, X., Lei, Z., Liu, X., Shi, H., Li, S.Z.: Face alignment across large poses: A 3d solution. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 146–155 (2016)
94. Zhu, X., Liu, X., Lei, Z., Li, S.Z.: Face alignment in full pose range: A 3d total solution. IEEE transactions on pattern analysis and machine intelligence **41**(1), 78–92 (2017)

A Implementation details

A.1 Unsupervised pretraining

For the unsupervised pretraining, similarly with [10] we trained our model on 64 GPUs using a batch size of 4096 and Synchronized Batch Normalization. The network was trained for 200 epochs using a weight decay of 10^{-6} and learning rate of 4.8 that was decayed toward 0.045 using a Cosine Scheduler [42]. During the first 10 epochs the learning rate is increased toward the target value using a linear scheduler. In all experiments, unless otherwise specified, we kept the temperature parameter to 0.1 and the Sinkhorn regularization parameters to 0.05. Each input sample was augmented into 2 views at a resolution of 224×224px and 6 at a resolution of 96×96px. The model was trained using the LARS [87] optimizer and was implemented in PyTorch [52].

Datasets and data preparation: All images are detected using [18] and then cropped based on the produced bounding-box so that the face will take approx. 190px on a 256×256px image. Unless otherwise specified all the data used for unsupervised pre-training were processed in the same manner.

A.2 Downstream task implementation details

Herein, we present the implementation details for each downstream task used in the main body to evaluate the efficacy of the facial representation learned. We note that in all cases the images were normalized in accordance with the training procedure of the pre-trained backbone model used as initialization.

Face recognition Following the best practices [75, 17], all images were normalized and aligned using the provided 5 landmarks. During training, the only augmentation applied was random horizontal flipping. Depending on the data regime, the models were trained between 18 and 54 epochs using a batch size of 512 and learning rate of 0.1. The weight decay was set to 0.0005 and the models were optimized using SGD with momentum (set to 0.9). For the cosface loss, the margin was set to 0.35. All models were trained on 8 GPUs.

Facial Landmark Localization The facial landmark localization pipeline was implemented following [4, 67]. During training, we applied the following augmentations randomly: rotation (between $\pm 30^o$), horizontal flipping, scaling $(0.85\times -1.15\times)$ and color jittering. Depending on the data regime, dataset and pretrained model, as detailed in the main body of the work, we trained the models between 60 and 480 epochs using a learning rate of 0.0001, a batch size of 24, a weight decay of 10^{-5} and Adam optimizer [32] $(\beta_1 = 0.5, \beta_2 = 0.99)$. All the models were trained using a pixel-wise ℓ_2 on a single GPU.

Action Unit (AU) Intensity Estimation For AU intensity estimation, we adopted a similar augmentation strategy with the one used for face alignment, mainly we applied random rotation $(\pm 30^o)$, random horizontal flipping and scale jittering $(0.85\times -1.15\times)$, Gaussian blurring with a kernel size between 5 and 10px and a probability of 0.4 and colour jittering. Depending on the setting, the models were trained between 60 and 320 epochs. The learning rate was typically set to 0.0001, the weight decay to 0.000005 and the batch size to 48. The models were optimized using Adam $(\beta_1 = 0.5, \beta_2 = 0.99)$ and trained on 2 GPUs.

Emotion recognition For valence and arousal estimation, we applied the same augmentation strategies as for AU Intensity Estimation with the exception of Gaussian blurring. Depending on the setting, the models were trained between 60 and 240 epochs using a batch size of 32, a learning rate of 0.1, weight decay of 10^{-4} and Adam optimizer$(\beta_1 = 0.5, \beta_2 = 0.99)$. All models were trained on a single GPU.

3D Face reconstruction Since 300W-LP has a small number of identities, during training we randomly augment the data using the following transformations: scaling$(0.85\times -1.15\times)$, in-plane rotation $(\pm 45^o)$, and random 10% translation w.r.t image width and height. Depending on the setting, we trained the model between 120 and 360 epochs using a learning rate of 0.05, a weight decay of 10^{-4} and SGD with momentum (set to 0.9). All models were trained using 2 GPUs.

A.3 Data sampling

For all low data scenarios, we randomly subsampled a set of annotated images without accounting for the labels (*i.e.* we don't attempt to balance the classes).

Once formed, the same subset is used for all subsequent experiments to avoid noise induced by different sets of images. For face recognition where the loss attempts to minimize the intra-class while maximising the inter-class distance and its sensitivity to both the number of identities and samples per identity, we deviated slightly from the above setting by enforcing that at least 1/4 of the identities are preserved for the very low data regime of 2%.

B Curated Datasets

For unsupervised pre-training we explore 3 curated datasets, collected for various facial analysis tasks: (a) Full VGG-Face ($\sim 3.4M$), (b) Small VGG-Face ($\sim 1M$) and (c) Large-Scale-Face ($> 5.0M$), consisting of VGG-Face2 [8], 300W-LP [93], IMDb-face [73], AffectNet [47] and WiderFace [85]. During unsupervised pre-training we drop all labels using only the facial images. See supplementary material for more details.

a) *Full VGG-Face* denotes the entirety of the VGG-Face2 dataset [8], consisting of $\sim 3.4M$ facial images of 9131 identities, with an average of 362.6 images for each subject. Images are downloaded from Google Image Search and have large variations in pose, age, illumination, ethnicity and profession, although they typically depict celebrities.

b) *Small VGG-Face* is a randomly sampled subset of 1M images selected from VGG-Face2.

c) *Large-Scale-Face* is constructed by combining the facial images from VGG-Face2 [8], 300W-LP [93], IMDb-face [73], AffectNet [47] and WiderFace [85]. Therefore, the dataset combines a set of datasets originally collected for facial recognition, face alignment, emotion recognition and face detection:

300W-LP [93] is a face alignment dataset constructed by warping into large poses, from -90^o to 90^o, the $\sim$ 4000 near-frontal images from the 300W [61] dataset. *IMDb-face* [73] is a large-scale noise-controlled dataset for face recognition, originally containing 1.7M faces with 59,000 identities which were manually cleaned by the authors from 2.0M raw images. All images were obtained by downloading data from the IMDb website. *AffectNet* [47] is a *in-the-wild* facial expression dataset consisting of more than 1M images collected by queering results from the internet using 1250 emotion related keywords. Out of this, 440,000 images were manually annotated with 7 discrete facial expressions and the intensity of valence and arousal. *WiderFace* [85] is a face detection benchmarking dataset consisting of 393,703 faces sourced from 32,203 images. The faces exhibit a high degree of variability in terms of scale, pose and occlusion.

C Uncurated Flick-Face dataset

Herein we provide additional details regarding the collected uncurated, in-the-wild, Flickr-Face dataset. The dataset was constructed by downloading a set of images from Flickr. The facial images were then automatically localized and cropped using a face detector [18]. In order to increase the likelihood of finding

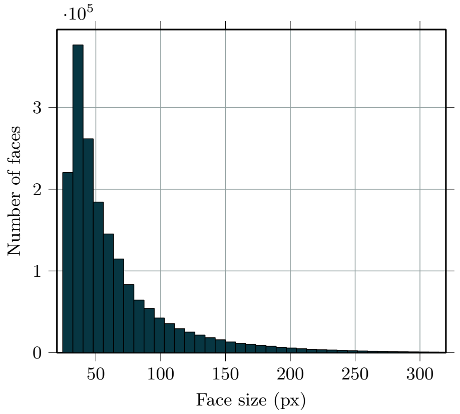


Fig. 5: Distribution on face sizes in the uncurated Flickr-Face dataset.

a face in the image we downloaded images that have one of the following 100 tags: *human, people, person, face, fashion, portrait, emotion, expression, affect, happy, sad, anger, angry, smile, laugh, joy, surprise, disgust, confused, fear, horror, adult, lady, ladies, beauty, gentleman, gentlemen, man, men, woman, women, baby, infant, toddler, kid, child, children, senior, father, mother, dad, mom, elderly, grandfather, grandmother, grandpa, grandma, grandparent, ancestor, 40s, 50s, 60s, 70s, 80s, 90s, couple, family, brother, sister, sibling, cousin, wedding, marriage, funeral, party, formal, boy, girl, teen, teenager, youth, friend, classmate, group photo, team, gathering, teacher, professor, lecturer, coach, tutor, worker, boss, celebrity, sport, self, selfie, photoshoot, concert, gigs, band, dance, marathon, passenger, army, soldier, marching, military, protest, crowds*. In total we collected 1.793.119 facial images with a bounding box size that follows the distribution shown in Fig. 5. We release the code used to download the images from Flickr thus allowing reproducing the dataset.

D Additional results

Herein, we report results for AU intensity estimation and emotion recognition (see Section D.1 and Tables 10, 11 and 12).

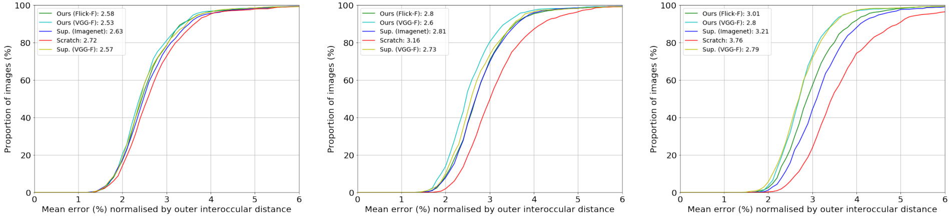
D.1 Emotion Recognition

We fine-tuned the models for valence and arousal estimation on the well-established AffectNet [47]. We report results in terms of RMSE and CCC [58], SAGR and PCC. 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 to jointly minimise the RMSE and CCC losses for valence and arousal, and the cross-entropy loss for classification.

Results are shown in Table 12 again, *for all* data regimes, our unsupervised models outperform the supervised baselines.

D.2 Additional 3D Face Reconstruction results

Furthermore, in Fig. 6 we report results on the Florence dataset for the task of 3D face reconstruction.



(a) Trained on 100% of the data.

(b) Trained on 10% of the data.

(c) Trained on 2% of the data.

Fig. 6: Cumulative 3D reconstruction error curves on the Florence [1] dataset for 3 different supervised data regimes: (a) using 100%, (b) 10% and (c) 2%. All models were trained on the 300W-LP dataset as detailed in the main body.

Table 10: Comparison against state-of-the-art on few-shot Facial AU intensity estimation on the DISFA dataset.

Method	**Data amount**	**AU**												**Avg.**
		1	**2**	**4**	**5**	**6**	**9**	**12**	**15**	**17**	**20**	**25**	**26**	
KBSS [90]	1%	.136	.116	.480	.169	.433	.353	.710	.154	.248	.085	.778	.536	.350
KJRE [92]	6%	.270	.350	.250	.330	.510	.310	.670	.140	.170	.200	.740	.250	.350
CLFL [91]	1%	.263	.194	.459	.354	.516	.356	.707	.183	.340	**.206**	.811	.510	.408
SSCFL [62]	2%	.327	.328	.645	.024	**.601**	.335	.783	.181	.243	.078	.882	.578	.413
Ours	1%	**.636**	**.667**	**.754**	**.367**	.549	**.535**	**.820**	**.313**	**.541**	.199	**.928**	**.608**	**.574**

Table 11: Comparison against state-of-the-art on few-shot Facial AU intensity estimation on the BU4D dataset.

Method	**Data amount**	**AU**					**Avg.**
		6	**10**	**12**	**14**	**17**	
KBSS [90]	1%	.760	.725	.840	.445	.454	.645
KJRE [92]	6%	.710	.610	.870	.390	.420	.600
CLFL [91]	1%	.766	.703	.827	.411	**.600**	.680
SSCFL [62]	2%	.766	.749	.857	.475	.553	.680
Ours	**1%**	**.789**	**.756**	**.882**	**.529**	.578	**.707**

Table 12: Results on the emotion recogntion task on the AffectNet dataset.

Data amount	**Init. method**	**Acc.**	**Valence**				**Arousal**			
			RMSE	**SAGR**	**PCC**	**CCC**	**RMSE**	**SAGR**	**PCC**	**CCC**
100%	random	0.590	0.370	0.790	0.696	0.695	0.339	0.781	0.613	0.611
	imagenet	0.592	0.360	0.789	0.705	0.705	**0.327**	0.792	0.624	0.620
	vggface	0.601	0.369	**0.798**	0.707	0.706	0.330	**0.796**	0.625	0.624
	ours	**0.602**	**0.356**	0.793	**0.711**	**0.710**	0.328	0.793	**0.634**	**0.629**
10%	random	0.493	0.402	0.752	0.626	0.625	0.366	0.753	0.536	0.536
	imagenet	0.548	0.383	**0.784**	0.655	0.654	0.351	0.767	0.569	0.566
	vggface	0.529	0.401	0.755	0.636	0.634	0.372	0.750	0.532	0.526
	ours	**0.562**	**0.382**	0.780	**0.678**	**0.678**	**0.344**	**0.803**	**0.600**	**0.599**
2%	random	0.419	0.453	0.727	0.515	0.515	0.400	0.747	0.423	0.422
	imagenet	0.479	0.411	0.740	0.562	0.557	0.362	0.769	0.465	0.456
	vggface	**0.511**	0.416	**0.778**	0.610	**0.607**	0.384	0.768	0.485	**0.485**
	ours	0.495	**0.370**	0.763	**0.620**	0.593	**0.338**	**0.794**	**0.500**	0.471