Image and Vision Computing 156 (2025) 105453

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

FRoundation: Are foundation models ready for face recognition?

Tahar Chettaoui [a,*], Naser Damer [a,b], Fadi Boutros [a]

[a] Fraunhofer Institute for Computer Graphics Research, Darmstadt, 64283, Germany
[b] Department of Computer Science, TU Darmstadt, Darmstadt, 64283, Hessen, Germany

ARTICLE INFO

MSC:
0000
1111

Keywords:
Face recognition
Foundation models
Computer vision
Face and gesture analyses
Human analyses
Fine-tuning

ABSTRACT

Foundation models are predominantly trained in an unsupervised or self-supervised manner on highly diverse and large-scale datasets, making them broadly applicable to various downstream tasks. In this work, we investigate for the first time whether such models are suitable for the specific domain of face recognition (FR). We further propose and demonstrate the adaptation of these models for FR across different levels of data availability, including synthetic data. Extensive experiments are conducted on multiple foundation models and datasets of varying scales for training and fine-tuning, with evaluation on a wide range of benchmarks. Our results indicate that, despite their versatility, pre-trained foundation models tend to underperform in FR in comparison with similar architectures trained specifically for this task. However, fine-tuning foundation models yields promising results, often surpassing models trained from scratch, particularly when training data is limited. For example, after fine-tuning only on 1K identities, DINOv2 ViT-S achieved average verification accuracy on LFW, CALFW, CPLFW, CFP-FP, and AgeDB30 benchmarks of 87.10%, compared to 64.70% achieved by the same model and without fine-tuning. While training the same model architecture, ViT-S, from scratch on 1k identities reached 69.96%. With access to larger-scale FR training datasets, these performances reach 96.03% and 95.59% for the DINOv2 and CLIP ViT-L models, respectively. In comparison to the ViT-based architectures trained from scratch for FR, fine-tuned same architectures of foundation models achieve similar performance while requiring lower training computational costs and not relying on the assumption of extensive data availability. We further demonstrated the use of synthetic face data, showing improved performances over both pre-trained foundation and ViT models. Additionally, we examine demographic biases, noting slightly higher biases in certain settings when using foundation models compared to models trained from scratch. We release our code and pre-trained models' weights at github.com/TaharChettaoui/FRoundation.

1. Introduction

Significant progress has been made in developing foundation models trained on extensive and diverse datasets recently [1–3]. Once these models are trained, they can be adapted to a wide array of downstream tasks, providing a versatile basis. Inspired by the success of large language models (LLM) [4–8], similar large-scale foundation models have been explored for various perception tasks [1–3,9]. Often based on the ViT architecture [10], which has been shown to match or exceed the performance of traditional methods in large-scale image classification tasks, visual foundation models are becoming increasingly popular due to their strong generalization capabilities when trained on larger datasets. These models can be optimized to perform template extraction, using zero or few-shot learning approaches, making them highly versatile for biometric applications, where collecting a large set of training data, e.g. face images, is technically and legally challenging. Although foundation models hold considerable promise for a wide range of applications, their adaptation for face recognition (FR) has not been explored in previous research, which motivates this study. In this work, we utilize visual foundation models to enhance performance for the downstream task of FR while minimizing reliance on extensive amounts of data, leveraging their pre-training on diverse datasets.

To explore the potential of foundation models in FR, we evaluate the performance of different versions of two widely used foundation models, namely DINOv2 [1] and CLIP [9]. Although foundation models demonstrate strong generalization capabilities, our experiments show that they perform poorly compared to state-of-the-art (SOTA) FR models [11–14] on various benchmarks. To enhance their effectiveness for the downstream task of FR, we propose to fine-tune the considered foundation models using low-rank adaptation (LoRA) [15]. LoRA integrates trainable low-rank decomposition matrices into each transformer

* Corresponding author.
E-mail addresses: tahar.chettaoui@igd.fraunhofer.de (T. Chettaoui), naser.damer@igd.fraunhofer.de (N. Damer), fadi.boutros@igd.fraunhofer.de (F. Boutros).

https://doi.org/10.1016/j.imavis.2025.105453
Received 30 October 2024; Received in revised form 21 January 2025; Accepted 5 February 2025
Available online 19 February 2025

block while keeping the pre-trained model weights frozen. After fine-tuning on dedicated datasets, the LoRA layers adapt to the downstream task of FR, resulting in a significant increase in performance. For example, the smallest released pre-trained models of DINOv2 and CLIP achieve average verification accuracies of 64.70% and 82.64%, respectively. After fine-tuning on CASIA-WebFace [16], their accuracies increase to 90.94% and 92.13%, respectively.

In addition to examining the performance of fine-tuned foundation models for FR, we compare them to a Vision Transformer (ViT) trained from scratch on different subsets of a small-scale training dataset, namely CASIA-WebFace. The goal of this experiment is to leverage the versatility of foundation models, which are trained on extensive and diverse datasets, making them strong candidates for small-scale fine-tuning, as they benefit from the data on which they were previously trained. Additionally, we aim to address the challenges in data collection for the downstream task of FR, which can be tedious due to legal privacy regulations requiring strict consent, ethical considerations regarding individual rights, and technical limitations in collecting large and diverse training datasets. Our experimental results demonstrate that, under conditions of low data availability, fine-tuned foundation models significantly outperform models trained from scratch, leveraging their pre-trained knowledge. On the other hand, as data availability increases, the performance of models trained from scratch becomes competitive. To provide insight into foundation models performances when large-scale finetuning datasets are available and to provide comparison results to recent SOTA FR models, we finetuned/trained different instances of CLIP, DINOv2, and ViT on two large-scale datasets, MS1MV2 and WebFace4M, that are widely used in the literature. Our results validated that models trained from scratch can eventually compete with or outperform fine-tuned foundation models when using large-scale datasets, highlighting the critical importance of selecting the appropriate training strategy based on dataset size.

With the growing use of synthetic data to develop FR models [17–19], which enables privacy-preserving training, we explore the performance of these solutions and demonstrate that foundation models outperform models trained from scratch when both are trained on the same synthetic data. The presented models in this paper are evaluated under cross-validation settings on mainstream challenging benchmarks, including ones with cross-age, cross-pose, and large-scale verification benchmarks. These evaluations are also enriched with demographic bias evaluation on racial face in the wild (RFW) by reporting verification accuracies as well as bias assessing metrics, e.g. skewed error ratio (SER) and standard deviations (STD). Our results on RFW, which are aligned with previous works [18,20–23], report demographic bias in deep learning-based FR models.

2. Related work

Face recognition:. Recent advances in FR performance have been primarily driven by advancing development in neural network architectures [10,24], innovations in training loss functions [12], and the availability of large-scale training datasets [25–27]. The majority of recent FR studies [11–13,28–30] employ ResNet-like [24] architectures, with some recent works [31,32] exploring ViT-based [10] architectures. Most of these studies [11–13,28–30] focus on innovations in training loss functions. FR training losses can be broadly categorized into two groups: metric learning (e.g., Triplet loss [33]) and multi-class classification learning (e.g., Softmax loss [34]). Metric learning losses [33,35], such as Triplet loss, explicitly encourage the model to learn discriminative feature representations by minimizing distances (e.g., Euclidean distance) between samples of the same label while maximizing distances between samples of different labels. In contrast, margin-penalty softmax loss employs cross-entropy over a softmax layer to implicitly guide the model in learning discriminative features. This is achieved by deploying a margin penalty on the angle or cosine angle between samples and their respective class centers, aiming at pushing samples closer to their respective class centers and further away from other class centers. Innovations in margin-penalty softmax loss have focused on the geometric deployment of penalty margins, whether fixed [11,12] or adaptive [13,28–30], yielding significant recognition improvements on mainstream benchmarks. These advancements in FR research have been made possible by large-scale training datasets [25–27], which enable the training of deep networks with millions of parameters. Notable datasets include CASIA-WebFace [16], MS-Celeb-1M [25] and its subsets (MS1MV2 [11] and MS1MV3 [36]), VGGFace2 [27], and WebFace260M [26] with its subsets (WebFace42M, WebFace12M, and WebFace4M). However, these datasets are collected from the internet, raising discussion about the ethical use of such data in FR development [17]. Such concerns, combined with the technical challenges of assembling large-scale datasets, have motivated researchers [37–40] to explore the use of synthetically generated data in FR development. These challenges in collecting large-scale face datasets are among the key motivations for this work. As detailed later in this paper, we studied the impact of fine-tuning dataset size on the foundation model performances, providing insights into the model performance when extremely small subsets are available (e.g., 82k images of 1k identities) and in the case where large-scale datasets (e.g., 5.8M images of 85k identities) are accessible. We also opted, as detailed later in this paper, to fine-tune foundation models on synthetically generated data, exploring the potential of such data in adapting foundation models for the FR task.

Foundation models:. Foundation models are defined by a substantial number of trainable parameters and are pre-trained on a large and diverse dataset, enabling them to adapt to a wide range of downstream tasks across various domains [41]. Vision foundation models are commonly structured around the use of Vision Transformers (ViTs) [10] and rely on self-supervised learning (SSL) [42]. SSL is a technique that trains models to learn representations from unlabeled data and is essential for training ViTs, which tend to perform poorly on small datasets [43]. Several high-performing vision foundation models specialized in various tasks have been released in recent years. Building on the success of the Segment Anything Model (SAM) [44], which introduced a foundation model for object segmentation, SAM2 [45] presents a model for real-time, promptable object segmentation in images and videos, achieving SOTA performance. CLIP [9] learns visual concepts from natural language supervision and can be applied to any visual classification by providing the names of the visual categories to be recognized. DINOv2 [1] foundation models generate universal features suitable for both image-level visual tasks, such as image classification and video understanding, as well as pixel-level visual tasks, including depth estimation and semantic segmentation.

Vision foundation models are characterized by their generalization ability due to massive training data, but they tend to show poor performance when applied to domain-specific settings [46]. To adapt them to a downstream task, multiple approaches were considered in the literature. An example of that is the AdaptFormer [47] that replaces the MLP block in the transformer encoder with two identical MLP branches, where one mirrors the original network and the other introduces a lightweight module for task-specific fine-tuning, demonstrating significant improvements compared to fully fine-tuned models on downstream tasks. Another example is the ViT-Adapter [48] that proposes a method to adapt plain ViTs for dense prediction tasks by injecting image-related inductive biases into the ViT and reconstructing fine-grained multi-scale features, yielding SOTA results on COCO test-dev. Another approach is to insert trainable rank decomposition matrices, called Low-rank adaptation (LoRA) [15] layers while freezing the pre-trained model weights. In this work, we choose LoRA as our foundation model adapter, as recent studies [46,49–52] highlight the significant potential of LoRA for this purpose. For instance, integrating LoRA layers into DINOv2 has been successfully applied in the medical domain for two distinct tasks: capsule endoscopy diagnosis [49] and depth estimation in endoscopic surgery [50], both of which have

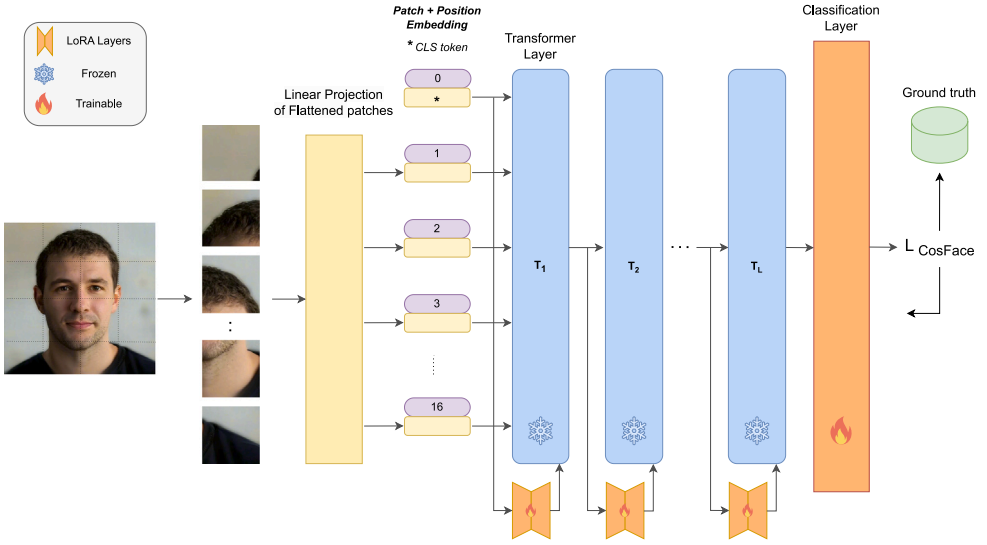

Fig. 1. ViT training pipeline with LoRA adapters. The facial image is divided into patches, which are mapped to patch embeddings via a linear projection. A class token is added, and positional embeddings are incorporated to maintain spatial information. This sequence of embedding vectors is then input into the encoder. The transformer layers remain fixed during training, while trainable LoRA layers are introduced to fine-tune the model. Each LoRA layer operates independently within the transformer layers and possesses its own set of weights.

proven superior performance in their respective fields. In another work on adapting foundation models for multiple plant phenotyping [51], Chen et al. demonstrated that LoRA consistently outperforms decoder tuning in leaf counting and disease classification, with their method achieving high performance in both tasks, approaching the results of SOTA models. Zanella et al. [52] explored few-shot adaptation of Vision Language Models (VLMs) using LoRA, showing that their CLIP-based method not only achieves substantial improvements but also reduces training times.

The application of foundation models in biometrics is still very limited, with only a few recent works starting to investigate their potential. For example, Iris-SAM [53], which is based on the foundation model SAM [44], fine-tunes it on ocular images for iris segmentation, achieving an average segmentation accuracy that surpasses the best baseline by a substantial margin of 10% on the ND-IRIS-0405 dataset. Arc2Face [54] is an identity-conditioned face foundation model that generates photo-realistic images based on the ArcFace [11] embedding of a person. To showcase the performance of the generated data, they train a FR model on synthetic images from their model, achieving superior performance compared to existing synthetic datasets [17]. Recognizing the immense potential of foundation models across diverse tasks, this study uncovers new perspectives by exploring their adaptation for FR, a path that, to the best of our knowledge, has been unexplored until now.

3. Methodology

This section presents our approach FRoundation, for optimizing vision foundation models for FR. This section starts by providing details on the selected baseline foundation models, CLIP [9] and DINOv2 [1]. Then, we provide details on the adaptation mechanism used to optimize foundation models for downstream tasks. Finally, we conclude by describing the extension of pre-trained foundation models with LoRA for FR.

3.1. Baseline foundation models

We selected two SOTA foundation models, CLIP and DINOv2, to conduct our studies in this paper. These models proved to be generalizable across different downstream tasks [1,9] and achieved very competitive results with zero-shot learning. Previous works [49,50] also showed that fine-tuning these models using, for example, LoRA, could lead to SOTA performance on specific downstream tasks.

3.1.1. DINOv2

DINOv2 [1] is a self-supervised image encoder trained on a large curated dataset, namely the LVD-142M dataset. The dataset was created as part of this initiative, using an automated pipeline to assemble a dedicated, diverse, and curated collection of images. The model network architecture follows a student–teacher framework based on vision transformers (ViT) [10] that learns features at the image and patch levels by combining DINO [55] and iBOT [56] losses. The image-level objective, inspired by DINO, is derived from the cross-entropy loss between features extracted from the student and teacher networks. These features are taken from the ViT class token and are based on different crops of the same image. For the patch-level objective, random input patches are masked and sent to the student, while remaining visible to the teacher. The student's iBOT head processes the masked tokens while the teacher's iBOT head processes the corresponding visible tokens, leading to the calculation of the loss term. Additionally, several contributions were made to accelerate and stabilize large-scale training. As a result, a ViT model with 1B parameters was trained and distilled into a series of smaller models that surpass the best available general-purpose features on most of the benchmarks at image and pixel levels [1], making it a top candidate as a foundation model for FR.

3.1.2. CLIP

Introduced by Radford et al. [9], Contrastive Language-Image Pre-training (CLIP) is a multimodal foundation model that jointly learns from visual and textual modalities. It leverages a large dataset comprising images paired with text description, enabling the model to learn and relate visual information to textual context, and vice versa. The architecture consists of two separate encoders to process image and text inputs, respectively. During training, CLIP employs a contrastive learning approach, maximizing the cosine similarity between feature representations obtained from image-text pairs while minimizing it for negative samples. This allows the model to effectively capture the relationship between images and their corresponding textual descriptions. The training process involves a large-scale training dataset, facilitating the model's ability to generalize across a variety of visual and textual

tasks. In this work, we will focus exclusively on employing the image encoder as we aim to obtain feature representation from face images for the face verification task.

3.2. Fine-tuning with LoRA

In this work, we utilize Low-rank adaptation (LoRA) [15] to fine-tune the considered foundation model. LoRA was initially developed to fine-tune LLMs for specific downstream tasks, but it can be applied to any neural network with dense layers. Its development was inspired by [57], which demonstrates that a low-dimensional reparameterization can be as effective for fine-tuning as the full parameter space. This indicates that pre-trained models actually reside on a low intrinsic dimension. Building on this concept, LoRA freezes the pre-trained model weights and inserts trainable rank decomposition matrices into each layer of the pre-trained transformer architecture. For a pre-trained weight matrix $W_0 \in \mathbb{R}^{d\times k}$, LoRA utilizes a low-rank decomposition to restrict its update by $W_0 + \Delta W = W_0 + BA$ where $B \in \mathbb{R}^{d\times r}$ and $A \in \mathbb{R}^{r\times k}$ with the rank $r \ll min(d, k)$. W_0 does not receive gradient updates during the training process, while only A and B are updated. When fine-tuning, this approach significantly reduces the number of trainable parameters compared to fine-tuning all the model parameters, while also not introducing any additional inference latency. The latter is achieved by computing $W = W_0 + BA$. After fine-tuning the model, the adapter weights BA are merged with the base model weight W_0 to compute the final model weights W.

During the forward pass, the low-rank matrix product BA is scaled by $\frac{\alpha}{r}$ where α is a constant. When optimizing with Adam [58], tuning α is roughly the same as tuning the learning rate [15]. This scaling factor causes gradient collapse as the rank increases, resulting in larger ranks performing no better than smaller ones [59]. To tackle this issue, the rank-stabilized LoRA (rsLoRA) [59] method proposes to scale the low-rank matrix with $\frac{\alpha}{\sqrt{r}}$. Gradients do not collapse with rsLoRA, and training with higher ranks has been experimentally validated to improve performance. This method allows for an effective compute/performance trade-off, where higher ranks can be used to achieve higher performance at the cost of increased training computation.

3.3. FRoundation

Utilizing pre-trained foundation models for FR leads to suboptimal results, as will be presented in Section 5.1. Thus, we propose to fine-tune the considered foundation models using LoRA. This requires extending the pre-trained ViT models with LoRA layers before fine-tuning them on dedicated datasets. It is possible to apply LoRA to the q, k, v, and o projection layers, which refer respectively to the query, key, value, and output projection matrices in the self-attention module of ViT architecture [10]. Following [15], we adapt the weight matrices of q and v only, as it was shown that adapting their respective weights W^q and W^v yields the best results on different downstream tasks [15].

The transformer encoder consists of alternating layers of Multi-headed Self-Attention (MSA) and multilayer perceptron (MLP) blocks. Layer Normalization (LN) is applied before every block and residual connections after every block. LoRA is applied exclusively to the attention weights, leaving the MLP unchanged for both simplicity and parameter efficiency [15]. As illustrated in Fig. 1, given a facial image, the image is divided into non-overlapping patches which are then mapped into patch embeddings using a linear projection layer. Additionally, a learnable embedding known as the class (CLS) token [10] is appended to the sequence of embedded patches. The goal of the CLS token is to serve as the image representation, which we utilize to obtain feature representation from input face samples. Then, position embeddings are incorporated into the patch embeddings to preserve positional information. The resulting sequence of embedding vectors

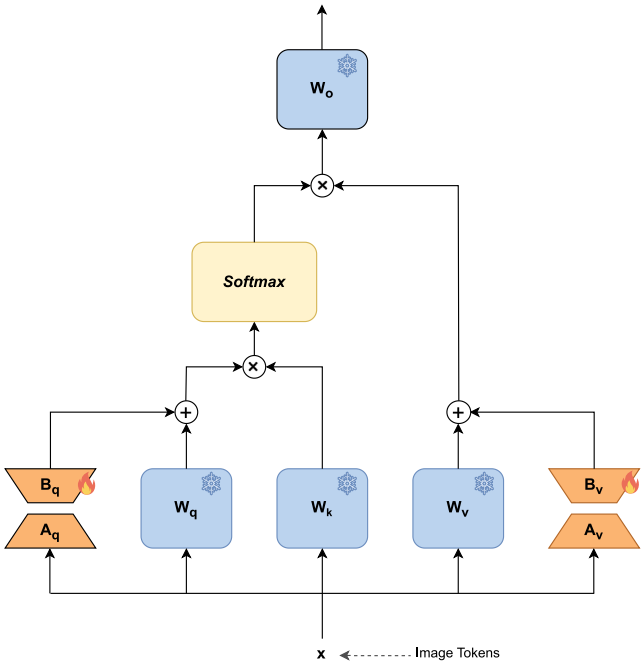

Fig. 2. LoRA integration in multi-headed self-attention block. We implement LoRA to q and v projection layers in each transformer block.

is used as input for the encoder. In MSA, we run k heads in parallel, each with its own set of q, k, and v. In each head, LoRA layers operate separately and have their own distinct weights. As shown in Fig. 2, for an embedding x, the computation of the q, k, and v projection layers in head i are:

$$\begin{aligned} Q_i &= W_i^q x + B_i^q A_i^q x, \\ K_i &= W_i^k x, \\ V_i &= W_i^v x + B_i^v A_i^v x \end{aligned} \tag{1}$$

W_i^q, W_i^k, and W_i^v are frozen projection layers for q, k, and v, respectively, while A_i^q, B_i^q, A_i^v, and B_i^v are the trainable LoRA layers. Then the attention scores are computed for all heads using a scaled dot-product attention mechanism. The attention output for head i is:

$$Attention(Q_i, K_i, V_i) = Softmax\left(\frac{Q_i K_i^T}{\sqrt{d_k}}\right) V_i, \tag{2}$$

where the scaling factor d_k represents the dimension of the key vectors. The output of all heads is concatenated along the feature dimension and passed through the projection layer O:

$$MutliHead(Q, K, V) = Concat(head_1, .., head_k) W^o. \tag{3}$$

The output of the projection layer O is then passed to the MLP, completing the execution of a single transformer block. L transformer layers are used to transform the image tokens into feature representations t^l where l denotes the output of the lth transformer block. Each lth transformer layer processes the output vectors of the previous layer. The final hidden state of the CLS token is employed as feature representation. To optimize the considered foundation models for FR, we proposed to fine-tune the models with the widely adapted margin-penalty softmax loss for FR [11,12]. Specifically, we extended the architecture with an additional multi-class classification layer and utilized CosFace loss [12] as a margin-penalty softmax loss.

Table 1
The achieved verification performances by DINOv2 and CLIP on several evaluation benchmarks. The results are reported in (%) on the small benchmarks and as average accuracies. The results of the first row are achieved by the ViT-S model trained from scratch on MS1MV2 and provided as a reference. The DINOv2 dataset blurred faces during training, as discussed in Section 5.1, which causes the gap in evaluation accuracies between the two models.

Method	Backbone	LFW	CFP-FP	AgeDB30	CALFW	CPLFW	Avg.	IJB-B			IJB-C		
								10^{-3}	10^{-4}	10^{-5}	10^{-3}	10^{-4}	10^{-5}
Baseline (MS1MV2)	ViT-S	99.73	95.44	97.42	95.95	92.35	96.18	95.89	92.85	79.68	96.76	94.51	88.30
DINOv2	ViT-S	78.73	71.15	54.70	59.47	59.48	64.70	14.66	5.90	2.33	18.10	7.44	2.87
	ViT-B	80.22	**72.64**	56.45	59.77	63.10	66.44	15.13	5.77	2.44	18.21	6.90	2.59
	ViT-L	**80.37**	71.97	55.25	**60.52**	**64.33**	66.49	17.18	6.44	2.52	**20.36**	7.84	2.77
	ViT-G	80.32	71.97	**58.62**	59.53	64.20	**66.93**	**17.59**	**6.76**	**2.79**	19.94	**8.05**	**3.23**
CLIP	ViT-B/32	94.03	84.91	70.72	76.13	77.47	80.65	39.36	20.58	9.38	43.73	25.02	11.98
	ViT-B/16	93.33	88.86	74.67	77.13	79.23	82.64	49.19	27.79	11.93	52.21	32.40	16.34
	ViT-L/14	95.90	90.66	**79.82**	**83.10**	**82.73**	**86.44**	**62.72**	**40.90**	**20.52**	**64.74**	**44.69**	**25.23**
	ViT-L/14@336	**96.30**	**91.26**	79.80	81.83	82.30	86.30	61.88	39.69	18.86	64.30	43.68	24.13

4. Experimental setups

Evaluation datasets. To explore the capabilities of foundation models for FR, we assess their performance using the following benchmarks: Labeled Faces in the Wild (LFW) [60], Celebrities in Frontal-Profile in the Wild (CFP-FP) [61], AgeDB30 [62], Cross-age LFW (CA-LFW) [63], CrossPose LFW (CP-LFW) [64]. The results are reported on each benchmark as verification accuracies in (%) and following their official evaluation protocol. In addition, we evaluated on large-scale evaluation benchmarks, IARPA Janus Benchmark-B (IJB-B) [65], and IARPA Janus Benchmark-C (IJB-C) [66]. For IJB-C and IJB-B, we used the official 1:1 mixed verification protocol and reported the verification performance as true acceptance rates (TAR) at false acceptance rates (FAR) of 1e−3, 1e−4, and 1e−5. All these benchmarks were chosen as they are widely used to benchmark the latest FR developments and provide a comprehensive variation of use-cases [11–14]. All results in the paper are reported using cross-validation settings, ensuring no overlap between the training and testing datasets. This approach aligns with SOTA works [11,13,28] in FR, by utilizing datasets such as CASIA-WebFace, MS1MV2, and WebFace4M for training/finetuning while reporting verification results on mainstream evaluation benchmarks, including LFW, CFP-FP, AgeDB30, CALFW, CPLFW, CFP-FP, IJB-C and IJB-B.

Training and fine-tuning datasets. To fine-tune the considered foundation models, we employ the CASIA-WebFace [16] dataset, consisting of 0.5 million images and 10K identities. We also conduct large-scale fine-tuning on the MS1MV2 [11] and WebFace4M [26] datasets. MS1MV2 is a refined version of MS-Celeb-1M [25] by [11], containing 5.8M images of 85K identities. WebFace4M is a subset of the WebFace260M dataset [26], consisting of 200K identities and 4 million images. We also investigate the performance of the selected foundation models when trained using synthetic data, focusing on a latent diffusion models-based approach, IDiff-Face [40], and a generative adversarial network (GAN)-based approach, SFace2 [39]. For both synthetic datasets, we conducted training on 10K identities, using 50 images per identity. The images in the training and testing dataset are aligned and cropped to 112 × 112 as described in [11] using five landmark points extracted by the Multi-task Cascaded Convolutional Networks (MTCNN) [67].

Model architectures. DINOv2 [1] officially released four ViT architectures, small (ViT-S), base (ViT-B), large (ViT-L), and giant (ViT-G). ViT-S, ViT-B, ViT-L, and ViT-G contain 22M, 86M, 0.3B, and 1.1B parameters, respectively. All models use a patch size of 14 but differ in embedding dimensions and the number of attention heads, with ViT-S having an embedding dimension of 384 and 6 heads, ViT-B having an embedding dimension of 768 and 12 heads, ViT-L having an embedding dimension of 1024 and 16 heads, and ViT-g having an embedding dimension of 1536 and 24 heads. On the other hand, CLIP [9] offers 4 different models with 2 architectures: base and large. The base model ViT-B of CLIP contains 86M parameters and has 2 variants with different patch sizes, 16 (ViT-B/16) and 32 (ViT-B/32). The large model ViT-L/14 has 0.3 billion parameters and includes a variant, namely ViT-L/14@336, that was pre-trained at a higher resolution of 336 pixels for one additional epoch to boost performance [68]. The ViT-B model has a width of 768 and 12 attention heads, while the ViT-L model has a width of 1024 and 16 attention heads. We will start by evaluating the performance of all models in Section 5.1. For further detailed experiments, we focus on the following models: ViT-S for DINOv2 and ViT-B/16 for CLIP, as these are the smallest models released by the corresponding authors, which makes our detailed experiments viable (given hardware and time limitations). We also investigate the performance of the larger models in Section 5.2 and choose ViT-L for CLIP and DINOv2 for a fairer comparison, as the largest model, namely the giant model of DINOv2, has 1.1 billion parameters compared to 0.3 billion parameters for DINOv2 and CLIP ViT Large.

Training settings. We utilize the CosFace [12] loss function to train all models presented in this paper with a margin penalty of 0.3 and scale factor of 64, following [12], as well as other works analyzing different building blocks of FR [39]. During the fine-tuning, we used AdamW [69] optimizer with a weight decay of 0.05. We train for 40 epochs (on CASIA-WebFace) and for 30 epochs (on MS1MV2 and WebFace4M), with a batch size set to 512 [11]. The initial learning rate is set to 0.0001 and is updated using a Cosine Learning Rate scheduler [70]. Additionally, the LoRA rank is set to 16 for all applicable experiments. The images are resized to 224 × 224 pixels to adapt to the image resolution initially used by DINOv2 and CLIP during training. For data augmentation, we apply horizontal flipping and RandAug [71] with 4 operations and a magnitude of 16, following [72]. During training and following [11,13], we explore efficient face verification datasets (e.g. LFW, CALFW, CPLFW, CFP-FP, AgeDB) to track and check the convergence status of the model. This has been performed after each epoch. It is important to note that for the RFW [23] evaluation, models trained on MS1MV2 are not subject to full cross-validation due to the identity overlap between RFW and MS1MV2, which is a cleaned version derived from MS-Celeb-1M [25], as stated in [23]. We additionally trained 12 instances of ViT, ViT-S, and ViT-L, from scratch on different subsets of CASIA-WebFace as well as on MS1MV2 and WebFace4M. These models are noted as the baseline.

Computational cost. We utilized two foundation models in this paper, each with different base architectures. For DINOv2, we utilized ViT-S and ViT-L, containing 22M and 0.3B parameters, respectively. For CLIP, we utilized ViT-B and ViT-L, containing 86M and 0.3B parameters, respectively. Given that these models utilize 32 floating-point (4-byte) values to represent each parameter, the required memory footprint for each model is 4 times the number of parameters. To evaluate computational time during inference, we measure the models' latency on a Quadro RTX 6000 GPU. Specifically, the ViT-S and ViT-L variants of DINOv2 require 6.993 ms and 13.614 ms, respectively, to process a single image on this hardware. When fine-tuning the models, we

Table 2
The achieved verification performances by the baseline model (trained from scratch) and fine-tuned DINOv2 and CLIP on different subsets of CASIA-WebFace. The results are reported in (%) on the small benchmarks and as average accuracies. On IJB-B and IJB-C, the results are reported as TAR at FAR of 1e−3, 1e−4 and 1e−5. The results of the first two rows are obtained from DINOv2 and CLIP models without fine-tuning. It is worth noting that fine-tuning DINOv2 and CLIP achieved higher recognition accuracies than training ViT models from scratch.

# Identities	# Images	Method	LFW	CFP-FP	AgeDB30	CALFW	CPLFW	Avg.	IJB-B			IJB-C		
									10^{-3}	10^{-4}	10^{-5}	10^{-3}	10^{-4}	10^{-5}
–	–	DINOV2	78.73	71.15	54.70	59.47	59.48	64.70	14.66	5.90	2.33	18.10	7.44	2.87
–	–	CLIP	93.33	88.86	74.67	77.13	79.23	82.64	49.19	27.79	11.93	52.21	32.40	16.34
		Baseline	88.33	65.21	61.07	73.35	61.85	69.96	2.44	0.93	0.54	2.69	1.08	0.55
1K	82 425	DINOv2	96.82	87.31	82.20	85.92	83.27	87.10	65.87	45.28	25.54	70.82	51.32	32.66
		CLIP	98.55	93.11	85.28	88.98	87.83	90.75	70.56	43.43	16.36	75.70	51.01	24.86
		Baseline	93.17	74.70	69.93	78.32	71.13	77.45	5.38	1.23	0.51	5.04	1.12	0.41
2.5K	163 945	DINOv2	97.80	89.60	84.25	87.72	85.15	88.90	72.21	52.88	25.12	77.05	59.88	37.23
		CLIP	98.87	93.51	86.12	90.07	88.78	91.47	71.03	44.12	18.42	76.38	52.58	26.97
		Baseline	95.78	82.89	78.33	83.25	77.72	83.59	30.37	4.81	1.00	26.91	4.04	0.83
2.5K (S)	289 228	DINOv2	98.50	90.81	87.25	88.68	85.93	90.23	77.28	61.69	38.96	80.68	66.97	49.67
		CLIP	98.63	94.23	86.28	89.50	88.20	91.37	72.16	41.36	13.18	77.46	51.09	20.48
		Baseline	96.32	81.71	78.25	84.23	78.25	83.75	28.63	5.91	1.11	24.17	4.79	1.00
5K	280 215	DINOv2	98.43	90.81	86.83	89.17	85.55	90.16	76.58	61.34	38.03	80.81	66.96	48.81
		CLIP	99.13	94.27	86.85	90.50	88.87	91.92	75.65	53.65	26.92	80.25	59.21	35.83
		Baseline	97.07	85.71	81.55	86.20	81.55	86.42	56.90	18.05	3.03	55.15	17.40	3.80
7.5K	389 007	DINOv2	98.40	91.94	87.20	89.63	85.60	90.55	74.36	46.42	15.00	77.96	52.97	22.29
		CLIP	99.13	94.49	87.33	90.43	88.45	91.97	77.92	56.26	25.84	81.77	62.78	34.74
		Baseline	98.02	88.04	84.70	88.25	83.78	88.56	73.00	39.61	10.48	73.03	36.77	10.79
10K	490 623	DINOv2	98.38	91.57	88.22	89.87	86.67	90.94	79.27	63.24	34.63	82.46	69.50	48.45
		CLIP	98.97	94.29	87.62	90.62	89.13	92.13	80.50	61.45	33.48	83.96	67.45	42.45

attach LoRA layers to the various considered model architectures. These layers introduce computational overhead, as the forward pass of the model becomes more complex due to the need to process the additional weights. As a result, the forward pass time for the DINOv2 ViT-S and ViT-L increases to 9.862 ms and 19.502 ms, respectively. For the CLIP variants, namely ViT-B and ViT-L, they require 8.657 ms and 12.74 ms, respectively, which increase to 12.498 and 18.817 ms when LoRA layers are added. While models fine-tuned with LoRA exhibit a slower forward pass during training, they benefit from the fact that only the additional LoRA layers are updated during the backpropagation process. For DINOv2, ViT-S (22M parameters) and ViT-L (0.3B parameters) incorporate an additional 0.3M and 1.5M parameters due to LoRA, respectively, which corresponds to only 1.32% and 0.51% of the total trainable parameters. For CLIP, ViT-B (86M parameters) and ViT-L (0.3B parameters) also add 0.5M and 1.5M parameters from LoRA, respectively, which account for only 0.68% and 0.51% of the total trainable parameters. It is important to note that LoRA adapter weights are merged with the base model weights once training is complete, as discussed in Section 3.2. This merging reduces memory usage and allows for inference speeds that are comparable to those of the original model, eliminating the need to maintain separate layers [15]. As a result, the extra computational load and memory usage during training is removed during inference.

5. Results

5.1. Are foundation models ready for face recognition?

We first evaluate the face verification performances of the considered foundation models, DINOv2 and CLIP, on several challenging benchmarks described in Section 4. For DINOv2 and CLIP, we utilized the official pre-trained models released by the corresponding authors [1,9]. The models, in this case, are utilized as feature extractors without fine-tuning the model weights.

The results achieved by the different pre-trained architectures of DINOv2 and CLIP are presented in Table 1. All network architectures are based on vision-transformer (ViT). For details on network architectures, one can refer to the corresponding papers [1,9]. The results of the first row (noted as baseline (MS1MV2)) are achieved by a ViT-S model trained from scratch on MS1MV2. These evaluation results are provided in this table as a reference. One can clearly observe that different pre-trained CLIP model versions outperformed DINOv2 models on the considered benchmarks. This result might be attributed to the fact that the curated dataset LVD-142M [1], on which DINOv2 was trained, involved a post-processing step that blurred identifiable faces [1]. Although the considered foundation models are not trained and optimized to perform feature extraction for face verification, they achieved relatively high accuracies on the considered benchmarks. For example, the achieved verification accuracy by CLIP (ViT-L/14@336), the largest model from CLIP, was 96.30% on LFW. On the benchmarks with cross-age evaluation protocol, the best-achieved accuracies on AgeDB30 and CALLFW were 79.80% and 81.83%, respectively. All considered models achieved relatively low TAR at the reported thresholds of FAR on large and challenging benchmarks, IJB-B and IJB-C.

From the reported results in Table 1, one can conclude that the achieved results by foundation models on face verification are far from being random. They achieved relatively high accuracies on less challenging benchmarks (LFW). However, this performance significantly drops when the evaluation benchmark contains challenging pairs such as AgeDB30, CALFW, IJB-B, and IJB-C, especially when considering the baseline model trained from scratch and the current performances of SOTA FR solutions [11–14].

5.2. Froundation: Fine-tuning foundation models

Table 2 presents the verification accuracies achieved by considered foundation models, DINOv2 and CLIP, fine-tuned with LoRA (Section 3.3) on different subsets of CASIA-WebFace. The results noted as the baseline are achieved by ViT-S trained from scratch and provided as a reference. In this experiment, we utilized the smallest released pre-trained ViT architecture from DINOv2 (ViT-S) and CLIP (ViT-B). The results of the first two rows in Table 2 are achieved by pre-trained DINOv2 (ViT-S) and CLIP (ViT-B) without fine-tuning. Driven by possible technical limitations and ethical concerns in practice to collect large-scale face (or other biometric) datasets [17], we propose to fine-tune the foundation models on small subsets from CAISA-WebFace and compare them to the case where the model was trained from scratch. Specifically, we provided a border evaluation of the impact of dataset size in terms of number of identities (dataset width) on the model performance by utilizing (n)K identities from CASIA-WebFace,

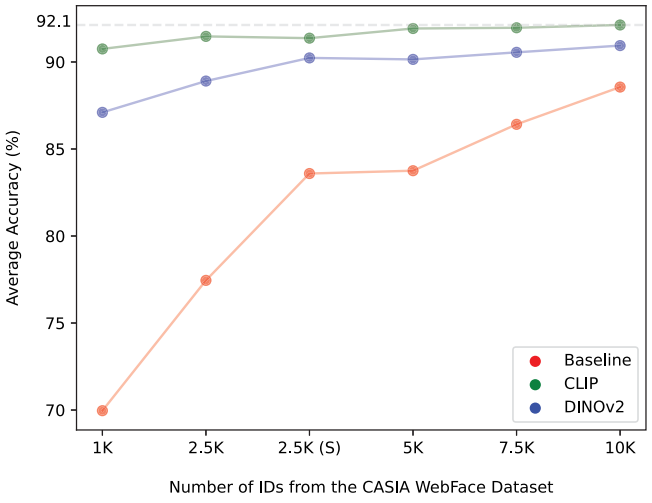

Fig. 3. Average verification accuracies on five benchmarks, LFW, CFP-FP, AgeDB30, CALFW, and CPLFW on the y-axis vs. training/fine-tuning dataset size, in terms of the number of identities, on the x-axis. The results correspond to the ones reported in Table 2 and plotted for ViT (baseline) as well as fine-tuned DINOv2 and CLIP. Increasing the training/fine-tuning dataset width (number of identities) improved the model recognition performances.

where $n \in [1, 2.5, 5, 7.5, 10]$. Additionally, we investigate the impact of the dataset depth (number of images) on the model performance by comparing the case where the 2.5K identities are randomly selected to the case where these 2.5K identities are selected (noted as 2.5K(S)) with the largest number of images per identity. The 2.5K and 2.5K(S) resulted in a total of 163 945 and 289 228 images, respectively. Based on the reported results in Table 2, we made the following observations:

- Fine-tuning DINOv2 and CLIP on different subsets from CASIA-WebFace significantly improved the achieved verification accuracies on all evaluation benchmarks, in comparison to the case where DINOv2 and CLIP are utilized without fine-tuning. Initially, without fine-tuning, DINOv2 (ViT-S) and CLIP (ViT-B) achieved average accuracies of 64.70% and 82.64%, respectively. These average accuracies increased to 87.10% and 90.75% by fine-tuning DINOv2 and CLIP, respectively, on a small subset of only 1K identities from CAISA-WebFace.
- As expected, using a larger subset for fine-tuning DINOv2 and CLIP consistently resulted in higher verification accuracies across all considered benchmarks. The same observation can also be made for the model trained from scratch (baseline).
- Fine-tuning ViT of pre-trained DINOv2 and CLIP led to higher verification accuracies, in comparison to the case where the model is trained from scratch. The superiority of foundation models over the model trained from scratch on CASIA-WebFace (or a subset of CASIA-WebFace) can be observed on the small evaluation benchmarks and the large-scale, IJB-B and IJB-C, benchmarks. This observation can be seen also in Fig. 3.
- Fig. 3 presents average accuracies (y-axis) of baseline and foundation models fine-tuned on different subsets (x-axis) from CASIA-WebFace. One can notice that the average accuracies slightly improved when a larger subset of CASIA-WebFace is utilized. On the other hand, the performance of the baseline significantly increases with more data, starting at an average accuracy of 69.96% when trained on 1K and reaching 88.56% on the full dataset with an increase of 18.6%. The same observation can also be made for the IJB-B and IJB-C benchmarks (Table 2).
- The impact of the dataset depth (number of images) can be observed by comparing the achieved results of models trained/fine-tuned on 2.5K (163 945) and 2.5K (S) (289 228). One can observe that increasing the dataset depth leads to higher verification accuracies in most settings.

5.3. Large-scale fine-tuning

Table 3 presents the achieved recognition performances by models trained from scratch (baseline) and fine-tuned (CLIP and DINOV2) on a relatively small dataset, CASIA-WebFace (0.5M images) and larger datasets, MS1MV2 (5.8M images) and WebFace4M (4M images). All results are reported using large (ViT-L) and small (ViT-S of DINOv2 and baseline and ViT-B of CLIP) architectures. We made the following observation from the reported results in Table 3:

- Using CASIA-WebFace (0.5M images), fine-tuning DINOv2 (ViT-S and ViT-L architectures) and CLIP (ViT-B and ViT-L architectures) led to higher verification accuracies, in comparison to the case where ViT-S and ViT-L are trained from scratch on the same data. For example, the average verification accuracies on the small benchmarks, LFW, CFP-FP, AgeDB30, CALFW, and CPLFW, was 87.39% by baseline (ViT-L) which is lower than 94.28% and 94.26% achieved by the fine-tuned DINOv2 and CLIP, respectively.
- The models that are trained/fine-tuned on large datasets, MS1MV2 and WebFace4M, achieved higher verification accuracies than the one trained on the relatively smaller dataset, CASIA-WebFace.
- Using small architectures (ViT-S and ViT-B) and large-scale training datasets (MS1MV2 or WebFace4M), the model trained from scratch achieved higher verification accuracies than the fine-tuned DINOv2 and CLIP. This observation can be noticed from the achieved results on small benchmarks as well as on large-scale, IJB-B and IJB-C, benchmarks.
- Using large architectures (ViT-L) and large-scale training/fine-tuning datasets (MS1MV2 or WebFace4M), the fine-tuned DINOv2 and CLIP achieved slightly higher recognition performance than the models trained from scratch on most of the evaluation benchmarks.

To conclude, for the case where only a relatively small training dataset is accessible, fine-tuning pre-trained foundation models can achieve higher recognition accuracy than training the same model architecture from scratch. In the case that one has access to large training datasets, training a model from scratch for FR can achieve very competitive results to fine-tuning foundation models. However, this, next to the technical and legal limitations of collecting or maintaining the data, comes with a high training time cost. For example, training ViT-L from scratch on MS1MV2 using the settings described in Section 4 requires around 70 GPU hours, in comparison to around 40 GPU hours for fine-tuning DINOv2 with LoRA on 8 Nvidia A100 SXM4 40 GB GPUs. Additionally, to put the performances achieved by FRoundation in the context of some of the major works in FR, we present in Table 4 results of the most influential works in the field [11–13,28–30]. However, this comparison is not direct, as such works commonly use the ResNet-100 [24] architecture and the MS1MV2 or WebFace4M dataset [11,26] for training. The various solutions outperform the trained from scratch models presented in Table 3. However, the results achieved by our FRoundation enhance the performance to a very close level to the works of the solutions presented in Table 4.

5.4. Learning from synthetic data

Table 5 presents the verification accuracies obtained by models trained from scratch (baseline) on IDiff-Face [40] and SFace2 [39] alongside the considered foundation models, DINOv2 and CLIP, which were fine-tuned using LoRA on IDiff-Face and SFace2. All results are reported using the experimental setups presented in Section 4. From the results presented in Table 5, we draw the following insights:

Table 3
The achieved verification accuracies by baseline models, as well as fine-tuned DINOv2 and CLIP, are presented on several challenging benchmarks. Different architectures of DINOv2 and CLIP are fine-tuned on different datasets, CASIA-WebFace, MS1MV2, and WebFace4M. The results of the baseline refer to models trained from scratch.

Method	Backbone	Train data	LFW	CFP-FP	AgeDB30	CALFW	CPLFW	Avg.	IJB-B			IJB-C		
									10^{-3}	10^{-4}	10^{-5}	10^{-3}	10^{-4}	10^{-5}
Baseline	ViT-S	CASIA-WebFace	98.02	88.04	84.70	88.25	83.78	88.56	73.00	39.61	10.48	73.03	36.77	10.79
		MS1MV2	99.73	95.44	97.42	95.95	92.35	96.18	95.89	92.85	79.68	96.76	94.51	88.30
		WebFace4M	99.63	96.57	96.20	95.55	92.92	96.17	96.34	93.86	88.30	97.44	95.62	92.46
	ViT-L	CASIA-WebFace	97.87	87.84	81.53	87.03	82.67	87.39	76.45	58.96	37.88	79.26	62.51	43.95
		MS1MV2	99.73	95.31	96.48	95.68	92.22	95.88	94.76	90.07	74.38	95.85	92.49	82.71
		WebFace4M	99.68	96.47	94.60	94.85	92.63	95.65	95.82	92.92	85.94	97.14	94.79	90.90
DINOv2	ViT-S	CASIA-WebFace	98.38	91.57	88.22	89.87	86.67	90.94	79.27	63.24	34.63	82.46	69.50	48.45
		MS1MV2	99.02	89.71	91.42	92.90	86.83	91.98	89.52	81.70	69.10	91.61	85.25	76.95
		WebFace4M	98.95	91.43	87.77	91.43	87.73	91.46	89.71	81.13	68.34	92.08	85.13	76.10
	ViT-L	CASIA-WebFace	99.33	95.77	92.77	92.33	91.20	94.28	89.98	78.81	61.77	92.44	84.45	72.91
		MS1MV2	99.63	95.93	96.22	95.50	92.78	96.01	95.31	91.95	83.50	96.41	93.94	89.89
		WebFace4M	99.65	96.84	95.10	94.80	93.75	96.03	95.64	92.65	86.42	96.96	94.79	91.40
CLIP	ViT-B	CASIA-WebFace	98.97	94.29	87.62	90.62	89.13	92.13	80.50	61.45	33.48	83.96	67.45	42.45
		MS1MV2	99.43	93.51	92.02	93.37	90.40	93.75	90.82	82.39	65.17	92.82	86.31	75.32
		WebFace4M	99.30	93.93	88.90	92.75	90.67	93.11	90.71	81.52	68.05	92.85	85.63	75.73
	ViT-L	CASIA-WebFace	99.55	95.73	91.73	92.58	91.70	94.26	90.33	78.71	56.96	92.59	83.12	68.87
		MS1MV2	99.68	96.76	93.20	94.60	93.73	95.59	95.73	91.22	82.78	96.80	93.66	88.62
		WebFace4M	99.65	96.50	93.72	94.37	93.73	95.59	95.12	90.72	82.76	96.62	93.40	88.55

Table 4
The achieved verification accuracies by notable previous works, namely CosFace, ArcFace, CurricularFace, MagFace, ElasticFace, and AdaFace, are reported on IJB-B and IJB-C as TAR at a FAR of 1e−4, along with various small benchmarks.

Method	Train data	LFW	CFP-FP	AgeDB30	CALFW	CPLFW	Avg.	IJB-B	IJB-C
CosFace [12,26] (CVPR 2018)	MS1MV2	99.81	98.18	98.34	96.18	92.76	97.05	–	96.01
ArcFace [11,26] (CVPR2019)	MS1MV2	99.78	98.54	98.21	96.05	92.93	97.10	–	96.03
CurricularFace [26,29] (CVPR2020)	MS1MV2	99.83	98.67	98.32	96.28	93.05	97.23	–	96.21
MagFace [28] (CVPR2021)	MS1MV2	99.83	98.46	98.17	96.15	92.87	97.09	94.51	95.97
ElasticFace-Arc [13] (CVPRW2022)	MS1MV2	99.80	98.67	98.35	96.17	93.27	97.25	95.22	96.49
AdaFace [30] (CVPR2022)	MS1MV2	99.82	98.49	98.05	96.08	93.53	97.19	95.67	96.89
CosFace [12,26] (CVPR 2018)	WebFace4M	99.80	99.25	97.45	95.95	94.40	97.37	–	96.86
ArcFace [11,26] (CVPR2019)	WebFace4M	99.85	99.04	97.82	95.93	94.31	97.39	–	96.77
CurricularFace [26,29] (CVPR2020)	WebFace4M	99.83	99.11	97.83	96.03	94.21	97.40	–	97.02
AdaFace [30] (CVPR2022)	WebFace4M	99.80	99.17	97.90	96.05	94.63	97.51	96.03	97.39

Table 5
The verification accuracies obtained by the baseline (trained from scratch) models, along with the foundation models DINOv2 and CLIP, are presented after training/fine-tuning on various synthetic datasets, including IDiff-Face and SFace2.

Method	Backbone	Train data	LFW	CFP-FP	AgeDB30	CALFW	CPLFW	Avg.	IJB-B			IJB-C		
									10^{-3}	10^{-4}	10^{-5}	10^{-3}	10^{-4}	10^{-5}
Baseline	ViT-S	SFace2	92.70	69.84	69.42	78.77	65.87	75.32	13.51	4.30	1.79	11.71	3.72	1.47
		IDiff-Face	96.35	77.50	79.48	85.78	75.57	82.93	60.30	33.77	12.55	58.33	30.08	11.61
	ViT-L	SFace2	91.37	71.56	68.05	75.83	68.82	75.13	40.94	18.65	6.62	36.31	14.53	5.34
		IDiff-Face	96.13	76.56	77.43	85.25	75.83	82.24	56.93	26.14	6.07	54.30	21.20	4.20
DINOv2	ViT-S	–	78.73	71.15	54.70	59.47	59.48	64.70	14.66	5.90	2.33	18.10	7.44	2.87
		SFace2	93.57	76.97	73.70	78.95	71.85	79.01	58.14	39.84	22.77	60.12	43.23	28.00
		IDiff-Face	96.13	80.54	80.13	86.12	74.32	83.45	64.86	42.01	15.99	67.23	47.80	23.58
	ViT-L	–	80.37	71.97	55.25	60.52	64.33	66.49	17.18	6.44	2.52	20.36	7.84	2.77
		SFace2	95.55	84.58	77.73	82.48	78.10	83.69	68.33	48.92	28.30	71.37	53.32	34.94
		IDiff-Face	97.77	86.94	86.67	90.28	81.73	88.68	75.11	55.75	32.80	78.75	62.17	40.94
CLIP	ViT-B	–	93.33	88.86	74.67	77.13	79.23	82.64	49.19	27.79	11.93	52.21	32.40	16.34
		SFace2	93.98	86.56	76.77	81.95	78.45	83.54	66.07	39.93	14.12	69.44	47.24	23.13
		IDiff-Face	97.58	87.91	80.78	88.17	80.48	86.98	74.47	54.15	26.27	77.54	61.01	37.43
	ViT-L	–	95.90	90.66	79.82	83.10	82.73	86.44	62.72	40.90	20.52	64.74	44.69	25.23
		SFace2	97.88	89.87	79.03	86.50	83.88	87.43	75.31	54.86	25.14	78.64	60.81	37.73
		IDiff-Face	98.75	91.40	86.50	90.78	84.87	90.46	82.92	64.56	34.82	86.07	71.77	48.18

- The recognition performance of CLIP and DINOv2 improves when fine-tuning with any of the considered synthetic datasets compared to using the pre-trained models alone, highlighting the significant potential of synthetic datasets in enhancing model performance.
- The models trained/fine-tuned on the IDiff-Face dataset generally deliver higher recognition performance across various benchmarks compared to those trained on the SFace2 dataset. For example, the ViT-L model of CLIP achieves higher verification accuracy on all evaluation benchmarks when fine-tuned on IDiff-Face,

Table 6
Evaluation results on RFW reported as average recognition performance in (%), standard deviation (STD) and skewed error ratio (SER) across four different demographic groups. The higher STD indicates a more biased model and the higher average (avg) indicates, in general, better recognition performance. The rightmost columns represent the average verification performance on small benchmarks (SB) and on IJB-C as TAR at FAR 1e−4, previously presented in Tables 1 and 3.

Method	Backbone	Train data	African	Asian	Caucasian	Indian	Avg.	STD	SER	Avg. SB	IJB-C 1e−4
Baseline	ViT-S	CASIA-WebFace	75.25	75.68	84.75	78.58	78.57	4.38	1.62	88.56	36.77
		MS1MV2	96.65	96.32	98.63	96.68	97.07	1.05	2.68	96.18	94.51
		WebFace4M	94.40	94.12	97.73	95.02	95.32	1.65	2.59	96.17	95.62
	ViT-L	CASIA-WebFace	71.87	73.47	81.08	74.23	75.16	4.06	1.48	87.39	62.51
		MS1MV2	97.32	96.48	98.45	97.25	97.38	0.81	2.27	95.88	92.49
		WebFace4M	93.37	92.25	96.33	93.27	93.80	1.75	2.11	95.65	94.79
DINOv2	ViT-S	–	54.77	61.13	65.00	60.42	60.33	4.21	1.29	64.70	7.44
		CASIA-WebFace	76.15	76.90	85.98	80.65	79.92	4.49	1.70	90.94	69.50
		MS1MV2	83.43	83.77	91.18	87.05	86.36	3.60	1.87	91.98	85.25
		WebFace4M	80.83	82.90	88.50	84.77	84.25	3.25	1.66	91.46	85.13
	ViT-L	–	58.46	64.20	67.47	60.93	62.77	3.91	1.27	66.49	7.84
		CASIA-WebFace	85.97	84.00	93.15	86.65	87.44	3.96	2.33	94.28	84.45
		MS1MV2	93.75	93.40	97.08	93.92	94.54	1.70	2.26	96.01	93.94
		WebFace4M	92.85	91.75	95.97	93.40	93.49	1.78	2.04	96.03	94.79
CLIP	ViT-B	–	70.75	69.73	79.32	68.98	72.19	4.80	1.49	82.64	32.40
		CASIA-WebFace	80.13	80.53	89.18	80.30	82.54	4.43	1.83	92.13	67.45
		MS1MV2	85.60	86.30	91.82	87.40	87.78	2.79	1.76	93.75	86.31
		WebFace4M	84.43	84.62	90.80	85.97	86.45	2.97	1.69	93.11	85.63
	ViT-L	–	74.03	72.15	82.60	73.15	75.48	4.80	1.60	86.44	44.69
		CASIA-WebFace	84.65	84.47	92.60	85.02	86.69	3.94	2.09	94.26	83.12
		MS1MV2	90.63	90.77	95.03	91.92	92.09	2.04	1.88	95.59	93.66
		WebFace4M	90.40	90.28	94.73	90.90	91.57	2.11	1.84	95.59	93.40

compared to when it is fine-tuned on SFace2. On the small benchmarks, the model achieved an average accuracy of 90.46% when fine-tuned on IDiff-Face, compared to 87.43% when fine-tuned on SFace2. A similar trend is observed on the large benchmarks, where the model fine-tuned on IDiff-Face achieved accuracies of 64.56% and 71.77% on IJB-B and IJB-C at a FAR of 1e−4, respectively, while the model fine-tuned on SFace2 achieved accuracies of 54.86% and 60.81% on the same benchmarks.
- Fine-tuning the larger ViT models of pre-trained DINOv2 and CLIP consistently led to higher verification accuracies on most benchmarks compared to their smaller counterparts. This is not the case for the baseline models, where both the ViT-S and ViT-L demonstrate competitive performance across various benchmarks. This can be attributed to the fact that larger models are more prone to overfitting when trained on smaller-scale datasets [10].

This investigation highlights the potential of synthetic data in enhancing the performance of foundation models for FR, demonstrating superior results compared to both pre-trained and baseline models. Furthermore, it encourages further exploration of additional synthetic datasets for FR, such as DCFace [38] and the more recent ID^3 [37].

5.5. Bias evaluations foundation models FR

We evaluated the baseline model, CLIP, and DINOv2 on the Racial Face in the Wild (RFW) dataset [23] to assess the models' bias and their performance across different demographic groups. The RFW dataset contains four testing subsets corresponding to:Caucasian, Asian, Indian, and African demographic groups. Following [20,23], we reported the results as verification accuracies in (%) on each subset and as average accuracies to evaluate general recognition performance. To evaluate the bias, we reported the standard deviation (STD) between all subsets and the skewed error ratio (SER), which is given by

$$\frac{max_g Error_g}{min_g Error_g}, \tag{4}$$

where g represents the demographic group, as reported in [20,21]. A higher STD value indicates more bias across demographic groups and vice versa. For SER, the model that achieved a value closer to 1 is less biased. Table 6 presents the evaluation results on RFW. As baseline models, we report the results for ViT-S and ViT-L (noted as baseline) trained from scratch on CASIA-WebFace, MS1MV2, and WebFace4M. We also reported the results of DINOv2 and CLIP without fine-tuning and for the case where the models are fine-tuned on CASIA-WebFace, MS1MV2, or WebFace4M. One can observe the following from the reported results in Table 6:

- Fine-tuning DINOv2 and CLIP improved the general recognition performance on all demographics, in comparison to the case where pre-trained DINOv2 and CLIP are used without fine-tuning. These results are complementary to the ones reported in the previous section.
- In general, the presented models achieved unequal verification performances on different subsets of RFW, where all models achieved their best verification performances on the Caucasian subset. This can be observed from the high STD values (far from optimal zero) and SER values that are far from 1. These observations are aligned with the previous works [18,20–23], reporting bias in FR when trained/finetuned on unbalanced datasets. Potential causes of this bias [73] link to the unbalanced training datasets, model's sensitivity to skin color [74] and/or hairstyles [75]. The racial distribution of the current FR datasets [16,25,26] used in the literature is not balanced with a majority of identities being Caucasians [20,21,23]. For example, as reported in [23], CASIA-WebFace [16] and MS-Celeb-1M [25], contain 84.5% and 76.3% Caucasians, 2.6% and 6.6 Asian, 1.6% and 2.6% Indian and 11.3% and 14.5% African, respectively. This is also true for the widely used WebFace260M [26] and its subsets, such as WebFace42M and WebFace4M, where the authors reported that the majority of identities are Caucasian. The previous works [18,20–23] reported that bias in these datasets would reflect in the FR algorithms, leading to higher verification accuracies on Caucasians compared to other ethnicities.
- Using CASIA-WebFace (0.5M images) for training/fine-tuning, the fine-tuned DINOv2 and CLIP achieved higher recognition performances than the baseline model trained from scratch. However, DINOv2 and CLIP are slightly more biased than baseline models. For example, the STD achieved by the baseline (ViT-S) was 4.38, which is slightly lower than the 4.49 and 4.43 STD achieved by fine-tuned DINOv2 (ViT-S) and CLIP (ViT-B).

- The models trained from scratch on large-scale datasets, MS1MV2 (5.8M images) and WebFace4M (4M images) achieved higher average recognition performances and lower standard deviation (in most of the settings) than fine-tuned CLIP and DINOv2. For example, using the ViT-L architecture trained/fine-tuned on WebFace4M, the baseline achieved an average accuracy of 93.80%, in comparison to 93.43% and 91.57% achieved by DINOv2 and CLIP, respectively. Also, in this case, the baseline model achieved a lower STD (1.74) than DINOv2 (1.78) and CLIP (2.11).
- In general, fine-tuned DINOv2 achieved higher average accuracies and lower STD than fine-tuned CLIP.
- Fine-tuning larger model architectures of pre-trained DINOv2 (ViT-L) or CLIP (ViT-L) achieved higher recognition accuracies and lower STD than fine-tuning smaller architectures: DINOv2 (ViT-S) or CLIP (ViT-B).
- A higher STD and an SER further away from the optimal value of 1 indicate that there is more room for improvement in achieving uniform performance across demographics. SER values are more sensitive to slight changes in the accuracy when the error margins are small (see Eq. (4)). This causes models trained or fine-tuned on MS1MV2 and WebFace4M to generally have higher SER values compared to those trained or fine-tuned on CASIA-WebFace, while typically maintaining a lower STD value. This is primarily due to the higher accuracy range of the models trained on the MS1MV2 and WebFace4M datasets (resulting in lower errors), which makes the SER more sensitive to small performance differences across demographic groups.
- When comparing the average verification performance on small benchmarks and IJB-C between the same model trained on WebFace4M and MS1MV2, we observe that both have comparable results in most settings. However, this observation does not hold for the performance on RFW, as the models trained on MS1MV2 outperformed those trained on WebFace4M in all cases and sometimes by a significant margin. For example, the baseline ViT large shows a performance gap of approximately 4%. This can be attributed [23] to the identity overlap between RFW and MS1MV2, a cleaned version derived from MS-Celeb-1M [25].

6. Conclusion

This paper was the first to propose and investigate the use of foundation models for the task of FR. We additionally propose the adaptation of the foundation models to this specific task under various levels of data availability. Our experiments on multiple foundation models, training datasets, and a wide range of evaluation benchmarks led to interesting conclusions, which are summarized as follows. The studied pre-trained foundation models used as feature extractors without fine-tuning, demonstrated relatively acceptable (far from random) accuracy on less challenging face verification benchmarks like LFW. However, they underperformed on more challenging benchmarks such as AgeDB30, CALFW, IJB-B, and IJB-C. Our adaptation of foundation models for FR showed that fine-tuning these models on even small subsets of CASIA-WebFace, e.g. 1K identities, significantly boosts their verification accuracy across benchmarks, outperforming models trained from scratch on similar data subsets. Additionally, increasing the dataset size (dataset width) or the number of images per identity (dataset depth) enhances performance, demonstrating the scalability of foundation models with diverse data availability. The results additionally showed that, with a limited training dataset like CASIA-WebFace, fine-tuning pre-trained foundation models outperforms training models from scratch in recognition accuracy. However, when large datasets (MS1MV2 or WebFace4M) are available, training from scratch yields competitive performance, though it incurs a significantly higher training computational cost compared to fine-tuning. This highlights the importance of selecting an appropriate training strategy according to the size of the dataset. Our bias evaluation results indicate that fine-tuning foundation models enhance recognition performance across demographics but introduces slightly more bias compared to baseline models trained from scratch, especially on smaller fine-tuning datasets. Models trained from scratch on larger datasets (MS1MV2 and WebFace4M) achieved superior recognition accuracy and exhibited lower bias. All models achieved their best verification performances on the Caucasian subset, which may be attributed to factors such as unbalanced training datasets, sensitivity to skin color, and hairstyles that are overrepresented or underrepresented in the training datasets. Finally, the effectiveness of synthetic data in improving the FR performance of foundation models is demonstrated, as it outperforms both pre-trained and baseline models. This also encourages further exploration of additional synthetic datasets. The outcomes of this work set the stage for broader adoption of foundation models as a basis for biometric recognition, particularly in scenarios with limited data availability, being aware of the technical and legal constraints on biometric data collection and management.

CRediT authorship contribution statement

Tahar Chettaoui: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Conceptualization. **Naser Damer:** Writing – review & editing, Writing – original draft, Validation, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization. **Fadi Boutros:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research work has been funded by the German Federal Ministry of Education and Research and the Hessen State Ministry for Higher Education, Research and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE.

Data availability

The authors do not have permission to share data.

References

[1] M. Oquab, T. Darcet, T. Moutakanni, H.V. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P. Huang, S. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jégou, J. Mairal, P. Labatut, A. Joulin, P. Bojanowski, DINOv2: Learning robust visual features without supervision, Trans. Mach. Learn. Res. 2024 (2024).

[2] H. Bao, L. Dong, S. Piao, F. Wei, Beit: BERT pre-training of image transformers, in: ICLR, OpenReview.net, 2022.

[3] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, R.B. Girshick, Masked autoencoders are scalable vision learners, in: CVPR, IEEE, 2022, pp. 15979–15988.

[4] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: NAACL-HLT (1), Association for Computational Linguistics, 2019, pp. 4171–4186.

[5] T.B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D.M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, in: NeurIPS, 2020.

[6] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, LLaMA: Open and efficient foundation language models, 2023, CoRR abs/2302.13971.

[7] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H.W. Chung, C. Sutton, S. Gehrmann, P. Schuh, K. Shi, S. Tsvyashchenko, J. Maynez, A. Rao, P. Barnes, Y. Tay, N. Shazeer, V. Prabhakaran, E. Reif, N. Du, B. Hutchinson, R. Pope, J. Bradbury, J. Austin, M. Isard, G. Gur-Ari, P. Yin, T. Duke, A. Levskaya, S. Ghemawat, S. Dev, H. Michalewski, X. Garcia, V. Misra, K. Robinson, L. Fedus, D. Zhou, D. Ippolito, D. Luan, H. Lim, B. Zoph, A. Spiridonov, R. Sepassi, D. Dohan, S. Agrawal, M. Omernick, A.M. Dai, T.S. Pillai, M. Pellat, A. Lewkowycz, E. Moreira, R. Child, O. Polozov, K. Lee, Z. Zhou, X. Wang, B. Saeta, M. Diaz, O. Firat, M. Catasta, J. Wei, K. Meier-Hellstern, D. Eck, J. Dean, S. Petrov, N. Fiedel, PaLM: Scaling language modeling with pathways, J. Mach. Learn. Res. 24 (2023) 240:1–240:113.

[8] R. Anil, S. Borgeaud, Y. Wu, J. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A.M. Dai, A. Hauth, K. Millican, D. Silver, S. Petrov, M. Johnson, I. Antonoglou, J. Schrittwieser, A. Glaese, J. Chen, E. Pitler, T.P. Lillicrap, A. Lazaridou, O. Firat, J. Molloy, M. Isard, P.R. Barham, T. Hennigan, B. Lee, F. Viola, M. Reynolds, Y. Xu, R. Doherty, E. Collins, C. Meyer, E. Rutherford, E. Moreira, K. Ayoub, M. Goel, G. Tucker, E. Piqueras, M. Krikun, I. Barr, N. Savinov, I. Danihelka, B. Roelofs, A. White, A. Andreassen, T. von Glehn, L. Yagati, M. Kazemi, L. Gonzalez, M. Khalman, J. Sygnowski, et al., Gemini: A family of highly capable multimodal models, 2023, CoRR abs/2312.11805.

[9] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision, in: ICML, in: Proceedings of Machine Learning Research, Vol. 139, PMLR, 2021, pp. 8748–8763.

[10] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, in: ICLR, OpenReview.net, 2021.

[11] J. Deng, J. Guo, J. Yang, N. Xue, I. Kotsia, S. Zafeiriou, ArcFace: Additive angular margin loss for deep face recognition, IEEE Trans. Pattern Anal. Mach. Intell. 44 (10) (2022) 5962–5979, http://dx.doi.org/10.1109/tpami.2021.3087709.

[12] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, W. Liu, CosFace: Large margin cosine loss for deep face recognition, in: CVPR, Computer Vision Foundation / IEEE Computer Society, 2018, pp. 5265–5274.

[13] F. Boutros, N. Damer, F. Kirchbuchner, A. Kuijper, ElasticFace: Elastic margin loss for deep face recognition, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2022, New Orleans, la, USA, June 19-20, 2022, IEEE, 2022, pp. 1577–1586, http://dx.doi.org/10.1109/CVPRW56347.2022.00164.

[14] J. Dan, Y. Liu, H. Xie, J. Deng, H. Xie, X. Xie, B. Sun, TransFace: Calibrating transformer training for face recognition from a data-centric perspective, in: IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023, IEEE, 2023, pp. 20585–20596, http://dx.doi.org/10.1109/ICCV51070.2023.01887.

[15] E.J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, Lora: Low-rank adaptation of large language models, in: ICLR, OpenReview.net, 2022.

[16] D. Yi, Z. Lei, S. Liao, S.Z. Li, Learning face representation from scratch, 2014, CoRR abs/1411.7923.

[17] F. Boutros, V. Struc, J. Fiérrez, N. Damer, Synthetic data for face recognition: Current state and future prospects, Image Vis. Comput. 135 (2023) 104688, http://dx.doi.org/10.1016/J.IMAVIS.2023.104688.

[18] P. Melzi, R. Tolosana, R. Vera-Rodríguez, M. Kim, C. Rathgeb, X. Liu, I. DeAndres-Tame, A. Morales, J. Fiérrez, J. Ortega-Garcia, W. Zhao, X. Zhu, Z. Yan, X. Zhang, J. Wu, Z. Lei, S. Tripathi, M. Kothari, M.H. Zama, D. Deb, B. Biesseck, P. Vidal, R. Granada, G.P. Fickel, G. Führ, D. Menotti, A. Unnervik, A. George, C. Ecabert, H. Otroshi-Shahreza, P. Rahimi, S. Marcel, I. Sarridis, C. Koutlis, G. Baltsou, S. Papadopoulos, C. Diou, N. Di Domenico, G. Borghi, L. Pellegrini, E. Mas-Candela, Á. Sánchez-Pérez, A. Atzori, F. Boutros, N. Damer, G. Fenu, M. Marras, FRCSyn-onGoing: Benchmarking and comprehensive evaluation of real and synthetic data to improve face recognition systems, Inf. Fusion 107 (2024) 102322, http://dx.doi.org/10.1016/J.INFFUS.2024.102322.

[19] H. Otroshi-Shahreza, C. Ecabert, A. George, A. Unnervik, S. Marcel, N. Di Domenico, G. Borghi, D. Maltoni, F. Boutros, J. Vogel, N. Damer, Á. Sánchez-Pérez, E. Mas-Candela, J. Calvo-Zaragoza, B. Biesseck, P. Vidal, R. Granada, D. Menotti, I. DeAndres-Tame, S.M.L. Cava, S. Concas, P. Melzi, R. Tolosana, R. Vera-Rodríguez, G. Perelli, G. Orrù, G.L. Marcialis, J. Fiérrez, SDFR: synthetic data for face recognition competition, in: 18th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2024, Istanbul, Turkey, May 27-31, 2024, IEEE, 2024, pp. 1–9, http://dx.doi.org/10.1109/FG59268.2024.10581946.

[20] M. Wang, Y. Zhang, W. Deng, Meta balanced network for fair face recognition, IEEE Trans. Pattern Anal. Mach. Intell. 44 (11) (2022) 8433–8448, http://dx.doi.org/10.1109/TPAMI.2021.3103191.

[21] M. Wang, W. Deng, Mitigating bias in face recognition using skewness-aware reinforcement learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 9322–9331.

[22] S. Gong, X. Liu, A.K. Jain, Mitigating face recognition bias via group adaptive classifier, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, Virtual, June 19-25, 2021, Computer Vision Foundation / IEEE, 2021, pp. 3414–3424, http://dx.doi.org/10.1109/CVPR46437.2021.00342, URL: https://openaccess.thecvf.com/content/CVPR2021/html/Gong_Mitigating_Face_Recognition_Bias_via_Group_Adaptive_Classifier_CVPR_2021_paper.html.

[23] M. Wang, W. Deng, J. Hu, X. Tao, Y. Huang, Racial faces in the wild: Reducing racial bias by information maximization adaptation network, in: 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019, IEEE, 2019, pp. 692–702, http://dx.doi.org/10.1109/ICCV.2019.00078.

[24] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2015, pp. 770–778, URL: https://api.semanticscholar.org/CorpusID:206594692.

[25] Y. Guo, L. Zhang, Y. Hu, X. He, J. Gao, MS-celeb-1M: A dataset and benchmark for large-scale face recognition, in: B. Leibe, J. Matas, N. Sebe, M. Welling (Eds.), Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, the Netherlands, October 11-14, 2016, Proceedings, Part III, in: Lecture Notes in Computer Science, vol. 9907, Springer, 2016, pp. 87–102, http://dx.doi.org/10.1007/978-3-319-46487-9_6.

[26] Z. Zhu, G. Huang, J. Deng, Y. Ye, J. Huang, X. Chen, J. Zhu, T. Yang, J. Lu, D. Du, J. Zhou, WebFace260M: A benchmark unveiling the power of million-scale deep face recognition, in: CVPR, Computer Vision Foundation / IEEE, 2021, pp. 10492–10502.

[27] Q. Cao, L. Shen, W. Xie, O.M. Parkhi, A. Zisserman, VGGFace2: A dataset for recognising faces across pose and age, in: FG, IEEE Computer Society, 2018, pp. 67–74.

[28] Q. Meng, S. Zhao, Z. Huang, F. Zhou, MagFace: A universal representation for face recognition and quality assessment, in: CVPR, Computer Vision Foundation / IEEE, 2021, pp. 14225–14234.

[29] Y. Huang, Y. Wang, Y. Tai, X. Liu, P. Shen, S. Li, J. Li, F. Huang, CurricularFace: Adaptive curriculum learning loss for deep face recognition, in: CVPR, Computer Vision Foundation / IEEE, 2020, pp. 5900–5909.

[30] M. Kim, A.K. Jain, X. Liu, AdaFace: Quality adaptive margin for face recognition, in: CVPR, IEEE, 2022, pp. 18729–18738.

[31] Z. Sun, G. Tzimiropoulos, Part-based face recognition with vision transformers, in: BMVC, BMVA Press, 2022, p. 611.

[32] M. Kim, Y. Su, F. Liu, A. Jain, X. Liu, KeyPoint relative position encoding for face recognition, in: CVPR, IEEE, 2024, pp. 244–255.

[33] F. Schroff, D. Kalenichenko, J. Philbin, FaceNet: A unified embedding for face recognition and clustering, in: CVPR, IEEE Computer Society, 2015, pp. 815–823.

[34] W. Liu, Y. Wen, Z. Yu, M. Yang, Large-margin softmax loss for convolutional neural networks, in: ICML, in: JMLR Workshop and Conference Proceedings, vol. 48, JMLR.org, 2016, pp. 507–516.

[35] K. Sohn, Improved deep metric learning with multi-class N-pair loss objective, in: D.D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, R. Garnett (Eds.), Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain, 2016, pp. 1849–1857, URL: https://proceedings.neurips.cc/paper/2016/hash/6b180037abbebea991d8b1232f8a8ca9-Abstract.html.

[36] J. Deng, J. Guo, D. Zhang, Y. Deng, X. Lu, S. Shi, Lightweight face recognition challenge, in: 2019 IEEE/CVF International Conference on Computer Vision Workshops, ICCV Workshops 2019, Seoul, Korea (South), October 27-28, 2019, IEEE, 2019, pp. 2638–2646, http://dx.doi.org/10.1109/ICCVW.2019.00322.

[37] J. Xu, S. Li, J. Wu, M. Xiong, A. Deng, J. Ji, Y. Huang, G. Mu, W. Feng, S. Ding, B. Hooi, ID^3: Identity-preserving-yet-diversified diffusion models for synthetic face recognition, in: The Thirty-Eighth Annual Conference on Neural Information Processing Systems, 2024, URL: https://openreview.net/forum?id=x4HMnqs6IE.

[38] M. Kim, F. Liu, A.K. Jain, X. Liu, Dcface: Synthetic face generation with dual condition diffusion model, in: CVPR, IEEE, 2023, pp. 12715–12725.

[39] F. Boutros, M. Huber, A.T. Luu, P. Siebke, N. Damer, SFace2: Synthetic-based face recognition with w-space identity-driven sampling, IEEE Trans. Biom. Behav. Identity Sci. 6 (3) (2024) 290–303, http://dx.doi.org/10.1109/TBIOM.2024.3371502.

[40] F. Boutros, J.H. Grebe, A. Kuijper, N. Damer, Idiff-face: Synthetic-based face recognition through fizzy identity-conditioned diffusion models, in: ICCV, IEEE, 2023, pp. 19593–19604.

[41] R. Bommasani, D.A. Hudson, E. Adeli, R.B. Altman, S. Arora, S. von Arx, M.S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, E. Brynjolfsson, S. Buch, D. Card, R. Castellon, N.S. Chatterji, A.S. Chen, K. Creel, J.Q. Davis, D. Demszky, C. Donahue, M. Doumbouya, E. Durmus, S. Ermon, J. Etchemendy, K. Ethayarajh, L. Fei-Fei, C. Finn, T. Gale, L.E. Gillespie, K. Goel, N.D. Goodman, S. Grossman, N. Guha, T. Hashimoto, P. Henderson, J. Hewitt, D.E. Ho, J. Hong, K. Hsu, J. Huang, T. Icard, S. Jain, D. Jurafsky, P. Kalluri, S. Karamcheti, G. Keeling, F. Khani, O. Khattab, P.W. Koh, M.S. Krass, R. Krishna, R. Kuditipudi, et al., On the opportunities and risks of foundation models, 2021, CoRR abs/2108.07258.

[42] J. Gui, T. Chen, J. Zhang, Q. Cao, Z. Sun, H. Luo, D. Tao, A survey on self-supervised learning: Algorithms, applications, and future trends, IEEE Trans. Pattern Anal. Mach. Intell. 46 (12) (2024) 9052–9071.

[43] H. Zhu, B. Chen, C. Yang, Understanding why ViT trains badly on small datasets: An intuitive perspective, 2023, CoRR abs/2302.03751.
[44] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A.C. Berg, W. Lo, P. Dollár, R.B. Girshick, Segment anything, in: ICCV, IEEE, 2023, pp. 3992–4003.
[45] N. Ravi, V. Gabeur, Y. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, E. Mintun, J. Pan, K.V. Alwala, N. Carion, C. Wu, R.B. Girshick, P. Dollár, C. Feichtenhofer, SAM 2: Segment anything in images and videos, 2024, CoRR abs/2408.00714.
[46] A. Wang, M. Islam, M. Xu, Y. Zhang, H. Ren, SAM meets robotic surgery: An empirical study on generalization, robustness and adaptation, in: ISIC/Care-AI/MedAGI/DeCaF@MICCAI, in: Lecture Notes in Computer Science, Vol. 14393, Springer, 2023, pp. 234–244.
[47] S. Chen, C. Ge, Z. Tong, J. Wang, Y. Song, J. Wang, P. Luo, AdaptFormer: Adapting vision transformers for scalable visual recognition, in: NeurIPS, 2022.
[48] Z. Chen, Y. Duan, W. Wang, J. He, T. Lu, J. Dai, Y. Qiao, Vision transformer adapter for dense predictions, in: ICLR, OpenReview.net, 2023.
[49] B. Zhang, Y. Chen, L. Bai, Y. Zhao, Y. Sun, Y. Yuan, J. Zhang, H. Ren, Learning to adapt foundation model DINOv2 for capsule endoscopy diagnosis, 2024, CoRR abs/2406.10508.
[50] B. Cui, M. Islam, L. Bai, H. Ren, Surgical-DINO: adapter learning of foundation models for depth estimation in endoscopic surgery, Int. J. Comput. Assist. Radiol. Surg. 19 (6) (2024) 1013–1020.
[51] F. Chen, M.V. Giuffrida, S.A. Tsaftaris, Adapting vision foundation models for plant phenotyping, in: ICCV (Workshops), IEEE, 2023, pp. 604–613.
[52] M. Zanella, I.B. Ayed, Low-rank few-shot adaptation of vision-language models, in: CVPR Workshops, IEEE, 2024, pp. 1593–1603.
[53] P. Farmanifard, A. Ross, Iris-SAM: Iris segmentation using a foundational model, 2024, CoRR abs/2402.06497.
[54] F.P. Papantoniou, A. Lattas, S. Moschoglou, J. Deng, B. Kainz, S. Zafeiriou, Arc2Face: A foundation model for ID-consistent human faces, in: ECCV (37), in: Lecture Notes in Computer Science, vol. 15095, Springer, 2024, pp. 241–261.
[55] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, A. Joulin, Emerging properties in self-supervised vision transformers, in: ICCV, IEEE, 2021, pp. 9630–9640.
[56] J. Zhou, C. Wei, H. Wang, W. Shen, C. Xie, A.L. Yuille, T. Kong, iBOT: Image BERT pre-training with online tokenizer, 2021, CoRR abs/2111.07832.
[57] A. Aghajanyan, S. Gupta, L. Zettlemoyer, Intrinsic dimensionality explains the effectiveness of language model fine-tuning, in: ACL/IJCNLP (1), Association for Computational Linguistics, 2021, pp. 7319–7328.
[58] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, in: ICLR (Poster), 2015.
[59] D. Kalajdzievski, A rank stabilization scaling factor for fine-tuning with lora, 2023, CoRR abs/2312.03732.
[60] G.B. Huang, M. Mattar, T. Berg, E. Learned-Miller, Labeled Faces in the Wild: A Database forStudying Face Recognition in Unconstrained Environments, in: Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition, Erik Learned-Miller and Andras Ferencz and Frédéric Jurie, Marseille, France, 2008, URL: https://inria.hal.science/inria-00321923.
[61] S. Sengupta, J. Chen, C. Castillo, V. Patel, R. Chellappa, D. Jacobs, Frontal to profile face verification in the wild, in: 2016 IEEE Winter Conference on Applications of Computer Vision, WACV 2016, in: 2016 IEEE Winter Conference on Applications of Computer Vision, WACV 2016, Institute of Electrical and Electronics Engineers Inc., 2016, http://dx.doi.org/10.1109/WACV.2016.7477558, Publisher Copyright: © 2016 IEEE.; IEEE Winter Conference on Applications of Computer Vision, WACV 2016 ; Conference date: 07-03-2016 Through 10-03-2016.
[62] S. Moschoglou, A. Papaioannou, C. Sagonas, J. Deng, I. Kotsia, S. Zafeiriou, Agedb: the first manually collected, in-the-wild age database, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop, Vol. 2, 2017, p. 5.
[63] T. Zheng, W. Deng, J. Hu, Cross-age LFW: a database for studying cross-age face recognition in unconstrained environments, 2017, CoRR abs/1708.08197.
[64] T. Zheng, W. Deng, Cross-Pose LFW: A Database for Studying Cross-Pose Face Recognition in Unconstrained Environments, Technical Report 18–01, Beijing University of Posts and Telecommunications, 2018.
[65] C. Whitelam, E. Taborsky, A. Blanton, B. Maze, J.C. Adams, T. Miller, N.D. Kalka, A.K. Jain, J.A. Duncan, K. Allen, J. Cheney, P. Grother, IARPA janus benchmark-b face dataset, in: CVPR Workshops, IEEE Computer Society, 2017, pp. 592–600.
[66] B. Maze, J.C. Adams, J.A. Duncan, N.D. Kalka, T. Miller, C. Otto, A.K. Jain, W.T. Niggel, J. Anderson, J. Cheney, P. Grother, IARPA janus benchmark - C: face dataset and protocol, in: ICB, IEEE, 2018, pp. 158–165.
[67] K. Zhang, Z. Zhang, Z. Li, Y. Qiao, Joint face detection and alignment using multitask cascaded convolutional networks, IEEE Signal Process. Lett. 23 (10) (2016) 1499–1503.
[68] H. Touvron, A. Vedaldi, M. Douze, H. Jégou, Fixing the train-test resolution discrepancy, in: NeurIPS, 2019, pp. 8250–8260.
[69] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, in: ICLR (Poster), OpenReview.net, 2019.
[70] I. Loshchilov, F. Hutter, SGDR: stochastic gradient descent with warm restarts, in: ICLR (Poster), OpenReview.net, 2017.
[71] E.D. Cubuk, B. Zoph, J. Shlens, Q.V. Le, Randaugment: Practical automated data augmentation with a reduced search space, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2020, Seattle, WA, USA, June 14-19, 2020, Computer Vision Foundation / IEEE, 2020, pp. 3008–3017, http://dx.doi.org/10.1109/CVPRW50498.2020.00359, URL: https://openaccess.thecvf.com/content_CVPRW_2020/html/w40/Cubuk_Randaugment_Practical_Automated_Data_Augmentation_With_a_Reduced_Search_Space_CVPRW_2020_paper.html.
[72] F. Boutros, M. Klemt, M. Fang, A. Kuijper, N. Damer, Exfacegan: Exploring identity directions in gan's learned latent space for synthetic identity generation, in: IEEE International Joint Conference on Biometrics, IJCB 2023, Ljubljana, Slovenia, September 25-28, 2023, IEEE, 2023, pp. 1–10, http://dx.doi.org/10.1109/IJCB57857.2023.10449036.
[73] S. Yucer, F. Tektas, N.A. Moubayed, T.P. Breckon, Racial bias within face recognition: A survey, 2023, CoRR abs/2305.00817.
[74] K.S. Krishnapriya, V. Albiero, K. Vangara, M.C. King, K.W. Bowyer, Issues related to face recognition accuracy varying based on race and skin tone, IEEE Trans. Technol. Soc. 1 (1) (2020) 8–20, http://dx.doi.org/10.1109/TTS.2020.2974996.
[75] K. Öztürk, H. Wu, K.W. Bowyer, Can the accuracy bias by facial hairstyle be reduced through balancing the training data? in: CVPR Workshops, IEEE, 2024, pp. 1519–1528.