LAFS: Landmark-based Facial Self-supervised Learning for Face Recognition

Supplementary Material

6. Introduction

This is the supplementary material for the paper **LAFS: Landmark-based Facial Self-supervised Learning for Face Recognition**. We first introduce the baseline Part fViT in Section 7.1. Then we describe the dataset collection details, model details, and hyper-parameters details for DINO, LAFS and fine-tuning in Section 8.1.1, Section 8.1.2 and Section 8.1.3 respectively. Finally, additional ablation studies including the number of shots, finetuning options for landmark supervision and comparison of full landmark view and global view are given in Section 8.2.

7. Additions to section 3: Method

7.1. Part fViT

Face saliency area has shown its capability of improving the recognition accuracy [7]. In this work, we adopt Part fViT [55] as our default backbone for recognition and investigate the property of the learned landmark CNN. Part fViT consists of two sequential components:(a) A landmark CNN which is responsible for predicting patch (landmark) centres and extracting corresponding patches. (b) A patch-based backbone, namely ViT, to predict the final identity embedding.

An image X is processed by a light-weight CNN to compute R landmark center r,

$$\mathbf{r} = \text{CNN}(\mathbf{X}),\ r_i = [x_i, y_i]^T,\ i = 1, \ldots, R \qquad (8)$$

Where r is the coordinates representing the landmark centre, normalized by the min-max scaler. Then, a patch whose centre is given by the landmark coordinate r_i with a fixed size of 8 in $R = 196$ is sampled using the differentiable grid sampling method of STN [31]. Following this, each patch is projected by the linear layer $\mathbf{E}$, then positioned by the positional encoding, appended with the class tokens before feeding to Transformer.

The common pipeline for ViT is non-overlapped patch split, patch to embedding projection with positional embedding, and Transformer block consists of self-attention and feed-forward networks. For more comprehensive information, please refer to [18].

The derivative of the Part fViT, the landmark CNN, is of providing stable landmarks with good correspondence [55], giving us a hint that it can be useful in developing a new perspective for self-supervised learning for face recognition.

8. Additions to section 4: Experiments

8.1. Implementation details

8.1.1 Dataset Details

Images are aligned by [11] and resized to 112× 112. Our models are evaluated on LFW [30], CFP-FP [51], AgeDB-30 [47], IJB-B[63], IJB-C[45] and MegaFace[35]. We report 1:1 verification accuracy on LFW, CFP-FP and AgeDB-30 datasets. Performance of Tar@Far=1e-4 is reported for IJB-B and IJB-C datasets. For Megaface, we report rank-1 identification accuracy (%) on 1M distractors and TAR@FAR=1e-6 verification accuracy, noted as Megaface/id and Megaface/ver respectively. We adopt a series of data augmentation while they are not typically used for Resnet training settings [11, 58]. We also provide results without data augmentation.

Flickr Dataset The Webface4M dataset are curated dataset containing high-quality and clean faces, we follow facial research [4] to collect Flickr dataset in which images are collected from in-the-wild. Differes from facial research [4], we downloaded images with following tags: *40s, 50s, 60s, 70s, 80s, 90s, baby, boss, celebrity, face, human*. All images are normalized and aligned by Retina-face [14]. In total, we collect 1.2M images which provide a similar data volume to our pretraining Webface 1-shot dataset.

8.1.2 Model details

We use fViT as proposed in Part fViT [55] as the default backbone with comparable parameters and Flops to ResNet-100 [25]. The landmark CNN is pretrained from Part fViT [55] which is MobilenetV3 [29].

8.1.3 Details for self-supervised pretraining and fine-tuning

DINO pretraining We follow the bulk of the default setting in DINO [6], specifically includes 2 global views and 8 local views, the augmentations contain Gaussian Blur, ColorJitter, Random Grayscale, and Solarization. The learning rate is 5e-4 and the optimizer is AdamW [43]. The exception is that the number of epochs is 40 while the warmup epoch is 10, the resolution of the global view is set to be 112 and for the local view is 48, and the dimensionality of the DINO head is 100K.

Data amount	Pretrain Method	Backbone	IJB-B					IJB-C				
			1 shot	2 shot	4 shot	10 shot	full	1 shot	2 shot	4 shot	10 shot	full
1%	Scratch	Resnet	14.13	19.69	30.58	22.19	58.66	15.39	22.15	33.56	25.01	63.16
	DINO	fViT	33.48	41.67	56.10	66.20	74.42	37.49	46.93	60.13	70.44	78.95
	LAFS	Part fViT	**33.97**	**42.80**	**57.32**	**68.76**	**75.67**	**37.89**	**47.03**	**61.09**	**72.57**	**79.66**
	DINO(flickr-ours)	fViT	17.23	20.65	28.34	47.67	69.38	21.43	23.59	32.98	47.32	73.11
	LAFS(flickr-ours)	Part fViT	17.55	21.72	30.79	47.76	70.11	22.10	25.95	34.89	50.63	73.85
10%	Scratch	Resnet	27.23	41.96	59.27	68.25	89.33	28.84	45.25	63.26	72.40	92.28
	DINO	fViT	46.99	66.61	81.56	88.35	91.35	51.35	70.73	85.02	91.42	93.85
	LAFS	Part fViT	**48.56**	**66.71**	**81.67**	**88.70**	**92.16**	**51.97**	**71.10**	**85.09**	**91.56**	**94.32**
	DINO(Flickr-ours)	fViT	28.87	44.39	71.51	76.11	89.15	31.55	57.01	80.17	79.86	85.76
	LAFS(Flickr-ours)	Part fViT	29.07	45.67	71.93	76.56	90.90	31.98	57.65	80.31	80.71	86.74

Table 6. Ablation study for different pretraining dataset. The default pretraining dataset is Webface-1shot, and Flickr-our denotes the Flickr data we collected.

Landmark-based pretraining Our LAFS pretraining starts by taking a fixed landmark CNN pre-trained from supervised learning. Since the augmentation for supervised training and self-supervised pretraining is different, the landmark CNN is not capable of producing accurate localization for self-supervised (DINO) augmented images. To address this, we disentangle the augmentations for the landmark CNN, while maintaining the flip and random resize & crop operations, in order to generate landmarks for the augmented images. Next, we change the local crop scale to be the same as the global views, i.e. from $[0.08 - 0.4]$ to $[0.4, 1.0]$. After converting the image into patches which are set to be R=196, we randomly sample 36 out of the 196 landmarks before feeding them to the student branch.

Finetuning We follow Part fViT [55] to use the same regularization and data augmentations. Then We adopt layer-wise learning rate decay [53] of 0.58 inspired by SimMIM [65]. The optimizer opted for is AdamW [43]. We conduct finetuning for 34-80 epochs based on the amount of available data, where we use 34 epochs for 100% of the data and 80 epochs for 1% of the data. The weight decay for all networks is 1e-1, the learning rate is 1e-4 with 5 warm-up epochs and cosine learning rate decay.

8.2. Ablation Study

8.3. Impact of In-the-wild dataset

We ablate the choice of pretraining data in this part, where we use Flickr and Webface-1shot for pretraining. Results are presented in Table 6. As it is shown, regardless of the amount of data, adopting Flickr as a pretraining dataset consistently has a negative impact. It even performs worse than Resnet pretrained by DINO on the 1shot setting using Webface-1shot. However, even with this negative impact, using Flickr as a pretraining dataset still yields better results compared to training the model from scratch.

8.3.1 Effect of number of shots

We conduct experiments to show how the number of shots in pretraining affects the fine-tuning results. To this end, we construct a training set consisting of 1 million images with 250k identities, which we refer to as the 4-shot setting. We compare pretraining on this 4-shot dataset to pretraining on the 1-shot setting, and find that pretraining on the 1-shot dataset yields better results, as is shown at the top of Table 7.

8.3.2 Effect of Landmark Finetuning choices

In this study, we aim to investigate the tolerance of fViT to varying the patch coordinates. To this end, we design four fine-tuning methods: (1) fixing the landmark CNN, (2) training the entire landmark CNN and fViT, (3) using an additional pre-trained and fixed landmark CNN to provide pseudo labels, and then training the entire network, We also exam (4) self-supervised training under the landmark pattern, then finetuning with a standard grid. Moreover, under the setting of (3), we control the strength of the pseudo-label to supervise the new model, that is, $\beta = 0.1, 1, 100$, results are available in the middle part of Table 7.

Our experimental results show that the strong supervision of $\beta = 100$ gives similar results to (2), while (4) results in worse training results than training fViT from scratch. We can conclude that when the gap of the input grid is very large, self-supervised pre-training and fine-tuning may lead to invalid pre-training. On the other hand, weak supervision of $\beta = 0.1$ achieves the best recognition accuracy. This indicates that adhering strictly to the supervised pattern of landmark coordinates will not obtain the optimal result. These findings suggest that Vision Transformers (Part fViT and fViT) are sensitive to coordinates (grids), and exploring the surrounding area of the landmark (i.e., using coordinate perturbation) during the pre-training stage may provide a more general landmark CNN for face recognition in the

Experiment	Content	1% data				100% data		
		LFW	CFP-FP	AgeDB	IJB-B	CFP-FP	AgeDB	IJB-B
Number of shot	4-shot pretraining	87.96	71.7	66.08	40.75	95.70	95.28	91.64
	1-shot pretraining	88.5	72	66.9	**41.3**	95.82	95.57	**91.90**
Finetune options	(1) Fixed landmark	87.34	70.93	64.67	38.31	95.14	95.06	87.82
	(2) Trainable landmark	88.53	72.19	66.8	41.25	95.94	95.51	91.85
	(3) β =100	87.37	70.88	64.53	38.24	95.10	95.10	87.87
	(3) β =1	88.01	71.39	65.37	39.47	95.34	95.27	90.14
	(3) β =0.1	88.5	72	66.9	**41.3**	95.82	95.57	**91.90**
	(4) landmark to standard grid	90.06	72.57	63.56	21.53	93.44	91.68	86.60
Global view vs Landmark View	Global views	86.7	70.3	64.9	39.5	95.32	95.22	91.54
	Mixed views	88.1	71.2	65.6	40.7	95.68	95.29	91.71
	Landamrk View	88.5	72	66.9	**41.3**	95.82	95.57	**91.90**

Table 7. Ablation studies for the number of shots, finetuning options and impact of different views.

fine-tuning stage.

8.3.3 Full Landmark View vs Global View

Our idea for this work is to minimise the presentation of all landmarks (full landmark views) with subsets of landmarks. However, we also explore the possibility of requiring the global image to have a similar representation of a subset of landmarks. To investigate this, we design an experiment where we compared the performance of our method with varying numbers of landmark views on the teacher branch as is shown in Table 7 (bottom part). Mixing views means the teacher processes global view(standard grid) and landmark view at the same time. The results indicate that with more landmark views on the teacher branch, finetuning performance behaves better. It can be drawn that minimising the representation with the global view with subset landmarks is much more challenging, and training with landmark view only can produce better self-supervised pretraining representation.

Pretraining	IJB-B	IJB-C
WebFace4m 1-shot	48.56	51.97
MS1M Random	48.47	51.76

Table 8. Difference between 1-shot and unlabeled pretraining

8.4. Discussion

8.4.1 1-shot Simulate unlabeled problem

We are aware that 1-shot pretraining may not accurately replicate the real-world conditions. To tackle this issue we randomly select 1M facial images from the MS1MV3 dataset to compare the difference between 1-shot pretraining and unlabeled pretraining. Finetune is carried on 1-shot with 10% of the available data as outlined in Table 8. One can conclude that unlabeled pretraining brings slightly worse but comparable performance than that of 1-shot pretraining.

8.4.2 Applicable Architectures

LAFS is not limited to ViT, but can be extended to any other patch-based backbones, e.g. MLP-Mixer

8.4.3 Training Cost

The training time on WebFace4M using 2 A100 GPUs is: 2.5 days for LAFS pre-training and 2.1 days for fine-tuning. Note that landmark learning is not included as we used a pre-trained facial landmark detector. Observe that the proposed LAFS is efficient; it only requires 40 epochs of pre-training to achieve SOTA performance. In comparison, DINO pre-training on ImageNet takes 300 epochs, i.e. about 18 days.

8.4.4 Use landmark CNN Fair?

The comparison is fair, many recognition methods [7, 17, 66] deploy some landmark CNN. Moreover, the network we adopted is very lightweight (MobileNet V3 containing parameters of 2.8M and 0.06 GMAC). Consequently, when we finetune the model, the supervision is fair as the landmarks CNN is pre-trained from identities only.