Part-based Face Recognition with Vision Transformers

Zhonglin Sun
zhonglin.sun@qmul.ac.uk
Georgios Tzimiropoulos
g.tzimiropoulos@qmul.ac.uk

School of Electronic Engineering and Computer Science
Queen Mary university of London
London, UK

Abstract

Holistic methods using CNNs and margin-based losses have dominated research on face recognition. In this work, we depart from this setting in two ways: (a) we employ the Vision Transformer as an architecture for training a very strong baseline for face recognition, simply called *fViT*, which already surpasses most state-of-the-art face recognition methods. (b) Secondly, we capitalize on the Transformer's inherent property to process information (visual tokens) extracted from irregular grids to devise a pipeline for face recognition which is reminiscent of part-based face recognition methods. Our pipeline, called *part fViT*, simply comprises a lightweight network to predict the coordinates of facial landmarks followed by the Vision Transformer operating on patches extracted from the predicted landmarks, and it is trained end-to-end with no landmark supervision. By learning to extract discriminative patches, our part-based Transformer further boosts the accuracy of our Vision Transformer baseline achieving state-of-the-art accuracy on several face recognition benchmarks.

1 Introduction

Face recognition(FR) is an important problem in computer vision with many applications such as border control and surveillance. With the advent of Deep Learning, the de-facto pipeline for FR over the last years comprises (a) a CNN(Convolutional Neural Network) backbone, which processes the face image holistically to compute a facial feature embedding which is used to calculate a similarity score, and (b) an appropriate loss function for discriminative embedding learning. While the bulk of recent work on FR has focused on (b), i.e., designing more effective loss functions [8, 13, 32, 35, 41, 49, 52], this work mostly focuses on (a) i.e. devising new architectures for facial feature extraction.

The first motivation of our work is the recently introduced Vision Transformer [16], which is gaining increasing popularity in Computer Vision with recent results reported being very competitive to the ones produced by CNN backbones [36, 56]. Hence, our first contribution is to explore how far one can go with a vanilla ViT for face recognition using the vanilla loss of [49]. We show that such a backbone with appropriate hyper-parameter optimization already achieves state-of-the-art results for face recognition. The second motivation for our work is that the ViT, contrary to CNNs, can actually operate on patches extracted

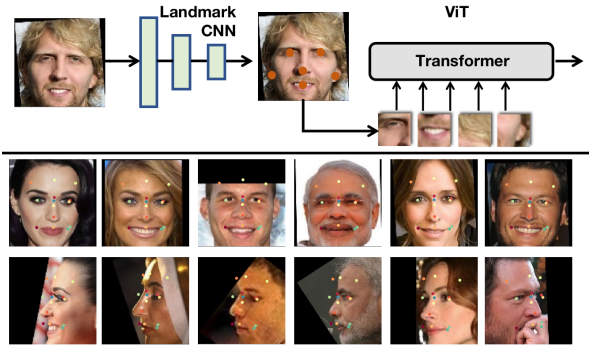

Figure 1: Illustration of our part-based ViT for face recognition. A facial image is processed by a lightweight landmark CNN which produces a set of facial landmarks. The landmarks are used to sample facial parts from the input image which are then used as input to a ViT for feature extraction and recognition. The whole system is trained end-to-end without landmark supervision. Examples of landmarks detected by the landmark CNN are shown.

from irregular grids and does not require the uniformly spaced sampling grid used for convolutions. As the human face is a structured object composed of parts (e.g., eyes, nose, lips), and inspired by seminal work on part-based face recognition before deep learning [5], in this paper, we propose to apply ViT on patches representing facial parts. Specifically, our second contribution is a newly proposed parts-based pipeline for deep face recognition where discriminatively learned landmarks are firstly predicted through a lightweight landmark CNN, patches are extracted around them and then fed to a ViT. Notably, the whole system, called part fViT, can be trained end-to-end without landmark supervision. Fig. 1 shows an overview of the proposed pipeline.

In summary, **our contributions** are:

- We appropriately train a vanilla ViT for face recognition using a vanilla loss, which we coin *fViT*, and show that fViT produces state-of-the-art results on several popular face recognition benchmarks.

- We capitalize on the Transformer architecture to propose a new pipeline for face recognition, coined *part fViT*, where discriminatively learned patches are firstly extracted and then fed to the ViT for recognition, essentially building a part-based ViT for face recognition. Notably, the landmark CNN used for predicting the landmarks is trained end-to-end with the ViT without landmark supervision.

- We show that our part fViT surpasses our strong baseline fViT setting a new state-of-the-art on several face recognition datasets. Moreover we ablate several components of our pipeline illustrating their impact on face recognition accuracy.

- We show that the landmark CNN which is part of our pipeline, is effective for the side task of unsupervised landmark discovery.

2 Related Work

A detailed review of face recognition papers is out of scope, herein we focus on losses, Region-aware methods and Vision Transformers which are more related to our work.

Loss functions: Several papers [8, 13, 32, 35, 41, 49, 52] have focused on learning features which are both separable and discriminative through using an appropriate loss function. While separability can be achieved with the softmax loss, learning discriminative features is more difficult as, within the mini-batch, training cannot see the global feature distribution [52]. To this end, FaceNet [41] uses triplets to directly learn a mapping to a compact Euclidean space such that facial features from the same identity are as close as possible while features from different identities are as far as possible.

To avoid the problem of triple selection, Center loss [52] minimizes the distance between the learned deep features for each face and their corresponding class centres in order to achieve intra-class concentration. Observing that the inter-class boundaries are not well separated in Softmax Loss, L-softmax [34] considers the joint formulation of softmax cross-entropy loss and linear layer, penalizing the distance of the class boundary, resulting in more discriminative features. Following that, CosFace[49] applied normalization not only on the weights, but also on the feature embedding, and proposed to add the margin on $cos(\theta)$ where θ is the angle between linear weight and embedding. ArcFace[9] further defined the margin on the angle θ rather than $cos(\theta)$. VPL[13] pays attention to learning the prototype of each class by regarding the distribution of classes on the feature space, and proposed to change the static prototype by injecting memorized features for approximating the prototype variation. Recently, Sphereface2 [53] proposes to conduct binary classification for recognition, and a number of general principles are also summarized in the work on how to design a good loss.

Region-aware methods: Although CNNs provide standard backbones for face recognition relying on global information, they ignore the fact that the face is a structured object with parts which can be used for more effective learning of facial features. For example, the seminal work of [5], which was the state-of-the-art before the advent of deep learning, shows that extracting a very large number of multi-scale features around 5 pre-defined landmarks (e.g. eye, nose, mouth) can be very effective for face recognition. To address local features via deep learning-based solutions, TUA[33] proposed to integrate local and global face features from different disjoint CNN via different GPUs, to aggregate the feature concatenation operation is used. FAN-Face [58] explored how features from a pre-trained facial landmark localization network can be used to enhance face recognition accuracy, however the landmark localization and recognition networks were not jointly trained. Moreover, [15, 24, 25] have all come up with methods to extract landmark-related features during CNN training, however, they still require pre-defined landmarks. To avoid explicit landmark supervision, Comparator Networks [57] propose a pipeline that performs attention to multiple discriminative local regions (landmarks), and uses them to compare local descriptors between pairs of faces. Finally, HPD [50] takes full use of the attention mechanism to predict attention masks for local features.

Our part fViT is inspired by [5, 57] but works *in a completely different manner*. Firstly, landmarks are learned by directly predicting their x,y coordinates using a very lightweight network (i.e. mobilenetV3 [19]). Then patches centred at the predicted landmarks are sampled and fed to a Transformer [16, 48] for face recognition. Notably we take advantage of the Transformer architecture to provide as input a set of patches sampled at irregular spatial locations which departs from standard face recognition methods based on CNNs which use a regular image grid (necessary to define convolutions) but also from ViT [16] which also

uses a regular grid for processing an input image. Moreover, our system is trained in an end-to-end manner without landmark supervision.

Vision Transformer: The Transformer was firstly introduced in Natural Language Processing for machine translation and other NLP tasks [43]. It comprises Self-attention and Feed-Forward layers. Vision Transformer (ViT) was introduced in [16], and since then it has been shown to provide competitive accuracy to CNNs [56]. Training ViT is more difficult compared to CNNs [46, 47]. A number of approaches have been proposed to facilitate ViT's training [6, 17, 36, 44, 47, 51, 55, 56, 60, 61, 61]. In this work, we discard the previous approach using ViT for face recognition [65] where regular overlapped patches are extracted from faces, instead we adopted the standard ViT backbone [16] with the training improvements of [44]. This already gives us a very strong baseline which surpasses most existing state-of-the-art methods for face recognition on MS1M [18] dataset. Next, we go beyond [16] and follow-up works [6, 17, 36, 44, 47, 51, 55, 56, 60, 61, 61] by applying the transformer, for the first time to the best of our knowledge on a set of patches extracted from *non-regular grids* provided by a lightweight network which is trained end-to-end to provide discriminative landmarks without explicit supervision.

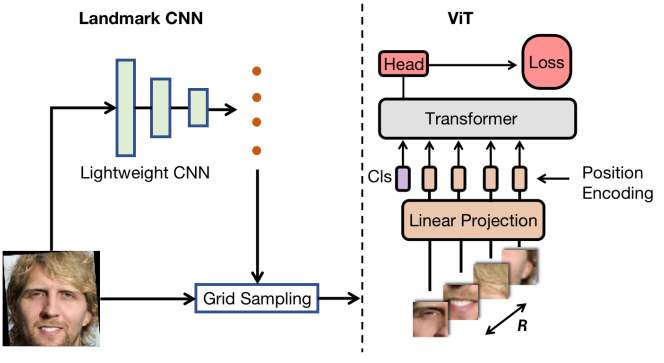

Figure 2: The overall structure of our proposed part fViT: A lightweight CNN is used to predict a set of facial landmarks. Then, differentiable grid sampling is applied to extract the discriminative facial parts which are then used as input to a ViT for feature extraction and recognition. Yellow nodes represent the regressed facial landmark coordinates extracted

3 Methodology

In Section 3.1, we firstly describe our strong baseline, called fViT, obtained by training ViT with CosFace loss. Then in Section 3.2, we introduce our proposed part-based ViT for face recognition, called part fViT.

3.1 fViT: ViT for Face Recognition

We are given a facial image $\mathbf{X} \in \mathbb{R}^{H \times W \times C}$ ($C = 3$). Following ViT [16], the image is divided into $R = P \times P$ non-overlapping patches which are then mapped into visual tokens using a linear embedding layer $\mathbf{E} \in \mathbb{R}^{P^2 \times d}$. To preserve spatial information a positional embedding

$\mathbf{p}_s \in \mathbb{R}^{P^2 \times d}$ is also learned which is added to the initial visual tokens. Then, the token sequence is processed by L Transformer layers.

The visual token at layer l and spatial location s is $\mathbf{z}_s^l \in \mathbb{R}^d$, $l = 0, \ldots, L-1$, $s = 0, \ldots, P^2-1$. In addition to the R visual tokens, a classification token $\mathbf{z}_{cls}^l \in \mathbb{R}^d$ is prepended to the token sequence [14]. The $l-$th Transformer layer processes the visual tokens $\mathbf{Z}^l \in \mathbb{R}^{(P^2+1)\times d}$ of the previous layer using a series of Multi-head Self-Attention (MSA), Layer Normalization (LN), and MLP ($\mathbb{R}^d \rightarrow \mathbb{R}^{4d} \rightarrow \mathbb{R}^d$) layers as follows:

$$\mathbf{Y}^l = \mathrm{MSA}(\mathrm{LN}(\mathbf{Z}^{l-1})) + \mathbf{Z}^{l-1}, \quad (1)$$

$$\mathbf{Z}^l = \mathrm{MLP}(\mathrm{LN}(\mathbf{Y}^l)) + \mathbf{Y}^l. \quad (2)$$

A single Self-Attention (SA) head is given by:

$$\mathbf{y}_s^l = \sum_{s'=0}^{P^2-1} \sigma\{(\mathbf{q}_s^l \cdot \mathbf{k}_{s'}^l)/\sqrt{d_h}\}\mathbf{v}_{s'}^l, s = 0, \ldots, P^2-1, \quad (3)$$

where $\sigma(.) = \mathrm{Softmax}(.)$, $\mathbf{q}_s^l, \mathbf{k}_s^l, \mathbf{v}_s^l \in \mathbb{R}^{d_h}$ are the query, key, and value vectors computed from $\mathbf{z}_s^l$ using embedding matrices $\mathbf{W_q}, \mathbf{W_k}, \mathbf{W_v} \in \mathbb{R}^{d \times d_h}$, d_h is the scale factor in self-attention. Finally, the outputs of the h heads are concatenated and projected using embedding matrix $\mathbf{W_h} \in \mathbb{R}^{hd_h \times d}$.

The classification token $\mathbf{z}_{cls}^L$ is trained for face recognition using the CosFace loss [29]:

$$\mathbf{Loss} = \frac{1}{N}\sum_i -\log \frac{\mathbf{e}^{\mathbf{b}(\cos(\theta_{\mathbf{y_i,i}})-\mathbf{m})}}{\mathbf{e}^{\mathbf{b}(\cos(\theta_{\mathbf{y_i,i}})-\mathbf{m})} + \sum_{j \neq \mathbf{y_i}} \mathbf{e}^{\mathbf{b}\cos(\theta_{\mathbf{j,i}})}}, \quad (4)$$

where N is the number of samples in a batch, $\mathbf{z} = \frac{\mathbf{z}_{cls}^L}{||\mathbf{z}_{cls}^L||}$, $\mathbf{z}_i$ is the $i-$th sample and $\mathbf{y}_i$ the corresponding ground-truth, $\mathbf{W} = \frac{\mathbf{W}^*}{||\mathbf{W}^*||}$ is the weight matrix of the last linear layer, W_j is the normalized $j-$th column (class) of the weight matrix, $\cos(\theta_{\mathbf{y_i,i}}) = W_{\mathbf{y_i}}^T \mathbf{z_i}$, $\mathbf{m}$ is the margin and $\mathbf{b}$ is fixed to be $||\mathbf{z}_{cls}^L||$.

We found that fViT, similarly to ViT is prone to overfitting. Hence, to obtain high accuracy, we used a combination of approaches for training including stochastic depth regularization [29], random resize & crop, RandAugment [7], Cutout, and finally Mixup [62]. The details of the choice of these are given in supplementary material 2.1.

3.2 Part fViT

The ViT as described by Eqs. 1 & 2 operates on a sequence of visual token which do not need to be computed on uniform grid. Inspired by work on part-based FR [5], in this section we describe how to apply ViT on patches representing facial parts.

Specifically, we use a lightweight weight CNN to predict a set of $R = P \times P$ landmarks:

$$\mathbf{r} = \mathrm{CNN}(\mathbf{X}),\ r_i = [x_i, y_i]^T,\ i = 1, \ldots, P^2, \quad (5)$$

where for our CNN we used a MobilenetV3 [19].

Then, we sample a patch centered at each landmark coordinate r_i. To accommodate for fractional coordinates, we used the differentiable grid sampling method of STN [22] for extracting each patch. Following this, each patch is tokenized by the embedding layer

E, giving rise to R part tokens which together with the class token are processed by the Transformer of Eqs. 1 & 2. We explore a number of options for the positional encodings added to the part tokens in an ablation study in Section 4.2.

The whole pipeline, called part fViT is very simple, and is shown in Fig. 2. It is trained end-to-end with no landmark supervision using simply the CosFace loss of Eq. 4. Notably, the landmark regression network forms an information bottleneck which was previously found useful in methods for unsupervised landmark discovery [23]. We also confirm this finding in an ablation study in Section 4.2. Finally, although heatmap regression methods with softmax could be used, we opted for direct coordinate regression which is simpler.

4 Experiments

In this section, we evaluate accuracy of the proposed face transformers on several well-known datasets and compare them with that of recently proposed state-of-the-art methods.

4.1 Implementation details

For training, and for a fair comparison with other methods, we used the refined version [10] of MS1M [18] (MS1MV3) containing 93,431 identities unless specificed. We also provide result training on VGGFace2 [3] with 3.1M images and 8.6K identities. Face images are of resolution 112×112 and aligned (provided by [3]) We tested our models on LFW [20], CFP-FP [43], AgeDB-30 [39], IJB-B[54], IJB-C[37] and MegaFace[26] for conducting recognition performance evaluation. For LFW, CFP-FP and AgeDB-30, we use 1:1 verification accuracy(%). We report TAR@FAR=1e-4 results on IJB-B and IJB-C. For Megaface, Megaface/id refers to the rank-1 identification accuracy (%) on 1M distractors, and Megaface/ver refers to TAR@FAR=1e-6 verification accuracy. For training the Transformer, we opted to use a large amount of data augmentation compared to the original FR setting used in ResNets, please refers to supplementary material Section 2.1.1 and 2.1.2 for more details regarding hyper-parameters, augmentations, model structure and training details.

4.2 Ablation Studies

We conducted a number of studies to highlight the impact of different design choices for our face Transformers. Our ablation studies are mainly carried out on the patch number $R = 49$ for its efficient training speed. We also attached the **improvement of data augmentation**, **degree of overlap** and **Effect of different landmark CNNs** in the supplementary material Section 2.2.

Effect of patch number and different fViT models: Our first experiment focuses on how the number of patches (or equivalently the number of landmarks R for the part fViT) impacts the accuracy of the proposed face Transformers. The number of patches chosen are 16, 49 and 196 with the FLOPs 1.17G, 3.3G and 12.64G respectively, and both fViT-B and fViT-S models are tested, as illustrated in Table 1. Note that when the number of patches increases, the patch size K is reduced; specifically for 196 landmarks the corresponding patch size is 8 and for 16 landmarks, the patch size is 28, ensuring that for the case of small number of landmarks the whole facial image is still analyzed. fViT-B has feature dim with 768 and

MLP dim with 2048 while fViT-S has 512 and MLP dim with 2560. In both cases the number of heads is 11. The results are shown in Table 1.

A number of interesting conclusions can be drawn by this experiment: **(1)** More patches (landmarks) result in more accurate prediction, as expected. **(2)** When the number of patches (landmarks) is very large (i.e. 196) then the part fViT outperforms fViT by small margin. **(3)** As the number of patches/landmarks decreases this gap increases specifically for CFP-FP and AgeDB. This is important as models processing fewer tokens are significantly more lightweight. For example the 49 landmark model is 4× faster than the 196 landmark model.

Backbone	Patch No.	Model	LFW	CFP-FP	AgeDB	IJB-C
fViT-B	196	part fViT	99.83	**99.21**	**98.29**	**97.29**
	196	fViT	99.85	99.01	98.13	97.21
	49	part fViT	99.80	98.78	97.85	96.37
	49	fViT	99.78	98.00	97.56	96.30
	16	part fViT	99.80	97.30	97.22	94.90
	16	fViT	99.78	96.87	96.46	94.85
fViT-S	196	part fViT	99.83	99.09	98.18	96.58
	196	fViT	99.83	98.90	97.90	96.50
	49	part fViT	99.80	98.7	97.81	96.33
	49	fViT	99.80	98.0	97.31	96.05
	16	part fViT	99.71	97.25	97.06	94.21
	16	fViT	99.71	96.95	96.25	94.19

Table 1: Impact of number of patches and different fViT models on FR accuracy.

Effect of different positional encodings Herein, we explore the function of positional encoding in our part fViT-B $R = 49$ landmarks. We test 3 types of positional encodings: (a) trainable ones as in the original fViT [16], (b) cosine [48] and (c) coordinate-based. For coordinate-based, we used a linear layer to embed each landmark r_i into $\mathcal{R}^d$ and then added this vector to the corresponding visual token. Results are shown in Table 2 (top section). As it can be observed the trainable one and the coordinate-based achieve the best accuracy.

Experiment	Content	LFW	CFP-FP	AgeDB	IJB-C
Positional encoding	Trainable	99.80	**98.78**	97.85	**96.37**
	Cosine	99.80	98.65	**98.03**	96.08
	Coordinate	99.80	98.71	97.66	96.29
Information bottleneck	w/ IB	99.80	**98.78**	**97.85**	**96.37**
	w/o IB	99.76	97.73	97.31	96.05
Unsupervised landmark	Vanilla fViT	99.78	98.00	97.56	96.30
	part fViT (MobilenetV3)	99.80	**98.78**	**97.85**	**96.37**
	part fViT (FAN (Frozen))	99.36	95.31	96.11	93.96
	part fViT (MobilenetV3 (Frozen))	99.81	98.72	97.66	96.35

Table 2: Results of various ablation studies: (a) Top section: impact of different positional encodings. (b) Middle section: impact of information bottleneck. (c) Last section: impact of unsupervised landmark discovery. All experiments are with part fViT-B with $R = 49$.

Effect of information bottleneck: We experimented with providing to the part fViT as input the penultimate layer's feature from the landmark CNN, essentially injecting features from the CNN to the fViT and violating the information bottleneck of our pipeline in Section 3.2. Specifically, the CNN penultimate layer's feature was concatenated with the (trainable) positional encoding and then projected to $\mathcal{R}^d$. Results are shown in Table 2 (middle section). As observed, violating the information bottleneck leads to decreased accuracy.

Effect of unsupervised landmark discovery: Since supervised facial landmark localization methods are widely used in literature, we compare our part fViT with a model that uses the landmarks provided by a state-of-the-art facial landmark localization, namely FAN [2]. We freeze the landmark CNN part from the well-trained part fViT to train a new ViT, coined as part fViT(mobilenet (Frozen)). Results are shown in Table 2 (bottom section). As it can be observed, using FAN (Pretrained and frozen parameters) to provide the input landmarks to fViT reduces to suboptimal performance. This way of directly using patches of landmarks provided by an accurate supervised landmark network leads to worse results than training a vanilla fViT. With a pretrained R=49 landmark network and only training the fViT part, we achieved a significant improvement than FAN network. We can conclude that for directly using patches of landmarks on the FR task, FAN is unable to provide the proper landmarks.

Method	LFW	CFP-FP	AgeDB	IJB-B	IJB-C	MegaFace/id	MegaFace/ver
CosFace[49]	99.81	98.12	98.11	94.80	96.37	97.91	97.91
ArcFace[9]	99.83	92.27	92.28	94.25	96.03	98.35	98.48
GroupFace[27]	98.85	98.63	96.20	94.93	96.26	98.74	98.79
CircleLoss[45]	99.73	96.02	-	-	93.95	98.50	98.73
DUL[4]	99.83	98.78	-	-	94.61	98.60	-
CurricularFace[24]	99.80	98.37	98.32	94.8	96.1	98.71	98.64
Sub-center ArcFace[10]	99.80	98.80	98.31	94.94	96.28	98.16	98.36
FAN-Face[63]	99.85	98.63	98.38	94.97	96.38	98.70	98.95
BroadFace[28]	99.85	98.63	98.38	94.97	96.38	98.70	98.95
ArcFace-challenge[10]	99.85	99.06	98.48	-	96.81	-	-
VPL[13]	99.83	99.11	**98.60**	95.56	96.76	98.80	**98.97**
ALN[64]	-	96.53	97.25	93.13	95.27	-	-
VirFace[31]	99.56	97.15	-	88.90	90.54	-	-
MagFace[38]	99.83	98.46	96.15	94.51	95.97	-	-
SCL[53]	99.80	98.59	98.26	94.74	96.09	81.40	97.15
Face Transformer [65]	99.83	96.19	97.82	-	95.96	-	-
fViT-B, ours	**99.85**	99.01	98.13	95.97	97.21	98.69	98.91
Part fViT-B, ours	99.83	**99.21**	98.29	**96.11**	**97.29**	**98.96**	98.78

Table 3: Comparison with the state-of-the-art on multiple datasets. Our baseline fViT and part fViT achieve state-of-the-art results on most datasets.

4.3 Comparison with the State-of-the-Art

We chose our part fViT-B and fViT with patch size 8 and R=196 to compare with recently proposed state-of-the-art FR methods. The landmark CNN used was MobilenetV3.

Quantitative results: We report the results of the models trained on MS1MV3, and tested on various benchmarks. The results are shown in Table 3. As observed, on LFW which is saturated, our proposed methods achieved top accuracy along with a few other methods. On the pose-sensitive dataset CFP-FP, our part-fViT has obtained the accuracy of 99.21%, surpassing the other state-of-the-art methods of VPL [13] and Arcface-challenge[10]. Similar results are observed for IJB-B and IJB-C benchmarks: not only does our part fViT outperform the other state-of-the-art methods by significant margin (97.29 TAR on IJB-C, 96.11 TAR on IJB-B), but even our baseline fViT is the second best method (97.21 TAR on IJB-C and 95.97 TAR on IJB-B). Similar results are obtained on MegaFace evaluation, where our part fViT is the top performing along with a few other methods. The only exception is on AgeDB-30, where our part fViT obtains 98.29%. We need to mention that the loss function used is CosFace [49] which was chosen for its simplicity and stability. It is possible

that using more advanced loss functions for training, including VPL [13], ArcFace [9] and Sphereface2 [34]. We also conducted experiments on the VGGFace2 dataset using similar

	LFW	AgeDB-30	IJB-B	IJB-C	MegaFace/Id	MegaFace/Ver
Comparator Networks [31]	-	-	85.0	88.5		
FAN-Face [33]	-	-	91.1	**93.5**	-	-
SphereFace [35]	99.55	92.88	89.41	91.96	71.53	85.02
CosFace [49]	99.51	92.98	88.61	90.98	71.65	85.45
ArcFace [9]	99.47	91.97	89.11	91.60	73.65	87.77
Circle Loss [43]	99.48	92.90	88.56	90.83	71.32	84.34
SphereFace2 [34]	99.50	93.68	**91.31**	93.25	**74.38**	**89.19**
fViT, Ours	99.44	93.52	88.13	90.26	71.11	85.04
part fViT, Ours	**99.56**	**93.92**	88.98	91.03	71.63	85.91

Table 4: Comparison with the state-of-the-art results on VGGFace2.

parameters with Resnet64 in SphereFace [35] to show the results of our part fViT in Table 4. Despite adding a large amount of data augmentation, our baseline fViT perform worse than the results provided by Resnet64 which is similar to the Face Transformer when training on a small scale dataset such as CASIA-webface [59]. Our part fViT also achieves a better result than the baseline fViT when training on MS1M, while it is still a little worse than the Resnet64 with advanced losses(e.g. ArcFace [9]). Our future work will investigate how our method works on other large scale benchmarks like Glink360 [1].

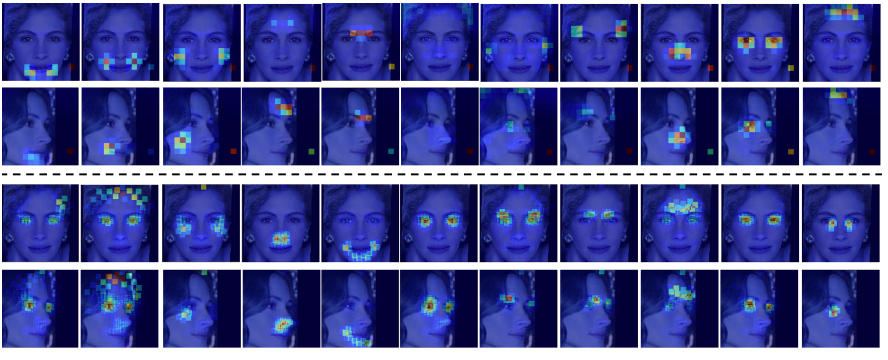

Figure 3: Visualization of attention maps. The first and second rows show the 11 attention maps produced by the 11 heads of the baseline fViT-B; The third and fourth rows show the 11 attention maps produced by the 11 heads of the part fViT-B with $R = 196$ landmarks.

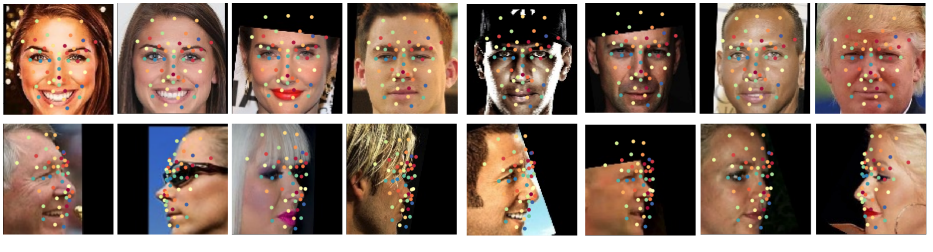

Figure 4: Visualization of the learned landmarks from our part fViT-B with $R = 49$. landmarks of same colour in different images across pose was learned to some good degree.

Qualitative results: We first compare the attention maps produced by the 11 heads of the baseline fViT and the part fViT in Fig. 3. We observe that for both methods, the heads achieve good correspondence across pose as each head fires at corresponding areas in both the frontal and the profile images. Then, a closer look reveals that the 6-th and 7-th attention heads (6-th and 7-th columns of Fig. 3) of the baseline fViT (1-st and 2-nd rows) do not focus on specific facial parts. Moreover, for the baseline fViT there's only one head that focuses on the eyes. This is in stark contrast with the part fViT where there are multiple heads focusing on the eyes region which are well-known to be the most discriminative facial parts for FR [30, 40, 42, 50, 57, 63]. Fig. 4 shows the 49 landmarks learned by our part fViT. As shown landmark correspondence across pose was learned to some good degree. Besides FR results, our landmark CNN can be useful for providing facial landmarks learned without landmark supervision. The detailed explanation can be observed in the supplementary material Section 2.3

5 Conclusions

We proposed face Transformers as architectures for highly accurate face recognition. We described two models: (a) fViT, our strong baseline trained appropriately on MS1M. (b) part fViT, we capitalized on the Transformer's property to process visual tokens extracted from irregular grids to propose a part-based face Transformer which is trained end-to-end to perform landmark localization and face recognition without explicit landmark supervision. Our pipeline is extremely simple comprising a lightweight CNN for direct coordinate regression followed by a ViT operating on the patches extracted from the predicted landmarks. Both models, and especially our part fViT, achieve state-of-the-art or near state-of-the-art accuracy on several face recognition benchmarks.

Acknowledgement

Zhonglin Sun is supported by China Scholarship Council(CSC).

References

[1] Xiang An, Xuhan Zhu, Yuan Gao, Yang Xiao, Yongle Zhao, Ziyong Feng, Lan Wu, Bin Qin, Ming Zhang, Debing Zhang, et al. Partial fc: Training 10 million identities on a single machine. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1445–1449, 2021.

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

[3] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *FG*, 2018.

[4] Jie Chang, Zhonghao Lan, Changmao Cheng, and Yichen Wei. Data uncertainty learning in face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5710–5719, 2020.

[5] Dong Chen, Xudong Cao, Fang Wen, and Jian Sun. Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3025–3032, 2013.

[6] Zhengsu Chen, Lingxi Xie, Jianwei Niu, Xuefeng Liu, Longhui Wei, and Qi Tian. Visformer: The vision-friendly transformer. *arXiv preprint arXiv:2104.12533*, 2021.

[7] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 702–703, 2020.

[8] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, 2019.

[9] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019.

[10] Jiankang Deng, Jia Guo, Debing Zhang, Yafeng Deng, Xiangju Lu, and Song Shi. Lightweight face recognition challenge. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[11] Jiankang Deng, Jia Guo, Tongliang Liu, Mingming Gong, and Stefanos Zafeiriou. Sub-center arcface: Boosting face recognition by large-scale noisy web faces. In *European Conference on Computer Vision*, pages 741–757. Springer, 2020.

[12] Jiankang Deng, Jia Guo, Xiang An, Zheng Zhu, and Stefanos Zafeiriou. Masked face recognition challenge: The insightface track report. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1437–1444, 2021.

[13] Jiankang Deng, Jia Guo, Jing Yang, Alexandros Lattas, and Stefanos Zafeiriou. Variational prototype learning for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11906–11915, 2021.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[15] Changxing Ding and Dacheng Tao. Trunk-branch ensemble convolutional neural networks for video-based face recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):1002–1014, 2017.

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

[17] Ben Graham, Alaaeldin El-Nouby, Hugo Touvron, Pierre Stock, Armand Joulin, Hervé Jégou, and Matthijs Douze. Levit: a vision transformer in convnet's clothing for faster inference. *arXiv preprint arXiv:2104.01136*, 2021.

[18] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European conference on computer vision*, pages 87–102. Springer, 2016.

[19] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1314–1324, 2019.

[20] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric 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*, 2008.

[21] Yuge Huang, Yuhan Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li, and Feiyue Huang. Curricularface: adaptive curriculum learning loss for deep face recognition. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5901–5910, 2020.

[22] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *Advances in neural information processing systems*, 28:2017–2025, 2015.

[23] Tomas Jakab, Ankush Gupta, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object landmarks through conditional image generation. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 4020–4031, 2018.

[24] Bong-Nam Kang, Yonghyun Kim, and Daijin Kim. Pairwise relational networks for face recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 628–645, 2018.

[25] Bong-Nam Kang, Yonghyun Kim, Bongjin Jun, and Daijin Kim. Hierarchical feature-pair relation networks for face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[26] Ira Kemelmacher-Shlizerman, Steven M Seitz, Daniel Miller, and Evan Brossard. The megaface benchmark: 1 million faces for recognition at scale. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4873–4882, 2016.

[27] Yonghyun Kim, Wonpyo Park, Myung-Cheol Roh, and Jongju Shin. Groupface: Learning latent groups and constructing group-based representations for face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5621–5630, 2020.

[28] Yonghyun Kim, Wonpyo Park, and Jongju Shin. Broadface: Looking at tens of thousands of people at once for face recognition. In *European Conference on Computer Vision*, pages 536–552. Springer, 2020.

[29] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Fractalnet: Ultra-deep neural networks without residuals. *arXiv preprint arXiv:1605.07648*, 2016.

[30] Susan J Lederman, Roberta L Klatzky, and Ryo Kitada. Haptic face processing and its relation to vision. In *Multisensory object perception in the primate brain*, pages 273–300. Springer, 2010.

[31] Pengyu Li, Biao Wang, and Lei Zhang. Virtual fully-connected layer: Training a large-scale face recognition dataset with limited computational resources. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13315–13324, 2021.

[32] Shen Li, Jianqing Xu, Xiaqing Xu, Pengcheng Shen, Shaoxin Li, and Bryan Hooi. Spherical confidence learning for face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15629–15637, 2021.

[33] Jingtuo Liu, Yafeng Deng, Tao Bai, Zhengping Wei, and Chang Huang. Targeting ultimate accuracy: Face recognition via deep embedding. *arXiv preprint arXiv:1506.07310*, 2015.

[34] Weiyang Liu, Yandong Wen, Zhiding Yu, and Meng Yang. Large-margin softmax loss for convolutional neural networks. In *ICML*, volume 2, page 7, 2016.

[35] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 212–220, 2017.

[36] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022, October 2021.

[37] Brianna Maze, Jocelyn Adams, James A Duncan, Nathan Kalka, Tim Miller, Charles Otto, Anil K Jain, W Tyler Niggel, Janet Anderson, Jordan Cheney, et al. Iarpa janus benchmark-c: Face dataset and protocol. In *2018 International Conference on Biometrics (ICB)*, pages 158–165. IEEE, 2018.

[38] Qiang Meng, Shichao Zhao, Zhida Huang, and Feng Zhou. Magface: A universal representation for face recognition and quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14225–14234, 2021.

[39] Stylianos Moschoglou, Athanasios Papaioannou, Christos Sagonas, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. Agedb: the first manually collected, in-the-wild age database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 51–59, 2017.

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

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

[42] Philippe G Schyns, Lizann Bonnar, and Frédéric Gosselin. Show me the features! understanding recognition from the use of visual information. *Psychological science*, 13(5):402–409, 2002.

[43] Soumyadip Sengupta, Jun-Cheng Chen, Carlos Castillo, Vishal M Patel, Rama Chellappa, and David W Jacobs. Frontal to profile face verification in the wild. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9. IEEE, 2016.

[44] Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your vit? data, augmentation, and regularization in vision transformers. *arXiv preprint arXiv:2106.10270*, 2021.

[45] Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. Circle loss: A unified perspective of pair similarity optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6398–6407, 2020.

[46] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021.

[47] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. *arXiv preprint arXiv:2103.17239*, 2021.

[48] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[49] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5265–5274, 2018.

[50] Qiangchang Wang, Tianyi Wu, He Zheng, and Guodong Guo. Hierarchical pyramid diverse attention networks for face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8326–8335, 2020.

[51] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. *arXiv preprint arXiv:2102.12122*, 2021.

[52] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *European conference on computer vision*, pages 499–515. Springer, 2016.

[53] Yandong Wen, Weiyang Liu, Adrian Weller, Bhiksha Raj, and Rita Singh. Sphereface2: Binary classification is all you need for deep face recognition. *arXiv preprint arXiv:2108.01513*, 2021.

[54] Cameron Whitelam, Emma Taborsky, Austin Blanton, Brianna Maze, Jocelyn Adams, Tim Miller, Nathan Kalka, Anil K Jain, James A Duncan, Kristen Allen, et al. Iarpa janus benchmark-b face dataset. In *proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 90–98, 2017.

[55] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. *arXiv preprint arXiv:2103.15808*, 2021.

[56] Tete Xiao, Piotr Dollar, Mannat Singh, Eric Mintun, Trevor Darrell, and Ross Girshick. Early convolutions help transformers see better. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=Lpfh1Bpqfk.

[57] Weidi Xie, Li Shen, and Andrew Zisserman. Comparator networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 782–797, 2018.

[58] Jing Yang, Adrian Bulat, and Georgios Tzimiropoulos. Fan-face: a simple orthogonal improvement to deep face recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12621–12628, 2020.

[59] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014.

[60] Kun Yuan, Shaopeng Guo, Ziwei Liu, Aojun Zhou, Fengwei Yu, and Wei Wu. Incorporating convolution designs into visual transformers. *arXiv preprint arXiv:2103.11816*, 2021.

[61] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zihang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. *arXiv preprint arXiv:2101.11986*, 2021.

[62] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[63] Lei Zhang, Meng Yang, Xiangchu Feng, Yi Ma, and David Zhang. Collaborative representation based classification for face recognition. *arXiv preprint arXiv:1204.2358*, 2012.

[64] Yaobin Zhang, Weihong Deng, Yaoyao Zhong, Jiani Hu, Xian Li, Dongyue Zhao, and Dongchao Wen. Adaptive label noise cleaning with meta-supervision for deep face recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15065–15075, 2021.

[65] Yaoyao Zhong and Weihong Deng. Face transformer for recognition. *arXiv preprint arXiv:2103.14803*, 2021.