Face Pyramid Vision Transformer

Khawar Islam[1]
https://khawar-islam.github.io
Muhammad Zaigham Zaheer[2,4]
https://zaighamz.com
Arif Mahmood[3]
arif.mahmood@itu.edu.pk

[1] FloppyDisk.AI, Pakistan
[2] Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE
[3] Information Technology University, Pakistan
[4] Electronics and Telecommunications Research Institute, South Korea

Abstract

A novel Face Pyramid Vision Transformer (FPVT) is proposed to learn a discriminative multi-scale facial representations for face recognition and verification. In FPVT, Face Spatial Reduction Attention (FSRA) and Dimensionality Reduction (FDR) layers are employed to make the feature maps compact, thus reducing the computations. An Improved Patch Embedding (IPE) algorithm is proposed to exploit the benefits of CNNs in ViTs (e.g., shared weights, local context, and receptive fields) to model lower-level edges to higher-level semantic primitives. Within FPVT framework, a Convolutional Feed-Forward Network (CFFN) is proposed that extracts locality information to learn low level facial information. The proposed FPVT is evaluated on seven benchmark datasets and compared with ten existing state-of-the-art methods, including CNNs, pure ViTs, and Convolutional ViTs. Despite fewer parameters, FPVT has demonstrated excellent performance over the compared methods. Project page is available at https://khawar-islam.github.io/fpvt/

1 Introduction

Transformer models have achieved excellent performance on numerous natural language processing tasks such as machine translation, question answering, and text classification. Later on, these models have also been successfully employed on many computer vision tasks, such as object detection [14], scene recognition [27], segmentation [20], and image super-resolution [16]. Although ViTs are applicable to many computer vision tasks, it is challenging to directly adapt these to pixel-level dense predictions particularly required for object detection and image segmentation tasks. It is because output feature maps of transformers are single-scale and low-resolution. Moreover the computational complexity and memory overhead are quite high even for relatively smaller input image sizes. To handle these issues, Wang et al. [25] have recently proposed Pyramid Vision Transformer (PVT). They introduced a pyramid structure to reduce the sequence length as the network deepens, resulting in significant reduction of the computational complexity. In the current work, we proposed Face Pyramid Vision Transformer (FPVT), which incorporates further complexity

reduction, improved patching strategy, and a loss function more appropriate for face recognition (FR) and verification tasks.

FR task is more challenging than object recognition and image classification tasks due to subtle inter-person discriminative attributes and significant intra-person variations. ViTs have not yet been well explored for the task of FR despite the presence of large scale datasets. Recently, Zhu et al. [35] proposed a dataset with 4M identities and 260M face images. However, training a ViT on a million-scale dataset takes significant time and requires extensive hardware resources. Our work in the current manuscript addresses this problem by employing PVT particularly for the application of FR and face verification.

Learning rich multi-scale features is a useful task for various problems [15, 25]. As for FR, patches may have diverse poses, expressions, and shapes, which make it necessary to learn multi-scale representation. In this work, we hypothesize that our proposed FPVT can capture long-range dependencies and distance-wise pixel relations, by constructing a hierarchical architecture and attention mechanism based on spatial reduction. Our proposed FPVT architecture consists of four stages that generate multi-scale features resulting in less training data requirement, reduced computational resources, reduced number of parameters, and improved FR performance.

Our proposed FPVT uses the advantages of CNNs, such as shared weights, local context, and receptive fields, while maintaining the benefits of ViTs, including attention, global context, and generalization. First, the transformer is divided into four blocks to create a pyramid structure. In the input, we employ an improved-patch approach to tokenize face images and expand the patch window such that it overlaps its surrounding patches. This allows FPVT to capture local facial continuity resulting in improved performance. Second, we introduce a depth-wise convolution in the feed-forward block of transformer to decrease the number of parameters while achieving better performance than PVT. To tackle the memory complexity in the recognition paradigm, we strategically embed the Facial Dimensionalty Reduction (FDR) layer [12] in the training pipeline to minimize the time and hardware cost of our method.

The main contributions of the current work can be summarized as follows:

- We present Face Pyramid Vision Transformer (FPVT) to learn multi-scale discriminative features and reduce the computation of large feature maps while achieving superior accuracy. Face-Spatial Reduction Attention (F-SRA) is designed to reduce the number of parameters.
- We introduce Improved Patch Embedding (IPE) which utilizes all benefits of CNNs to model lower-level edges to higher-level semantic primitives.
- Additionally, FPVT introduces a CFFN that extracts locality information to learn more about local representations as well as consider a long-range relationships.
- Face Dimensionality Reduction (FDR) layer is introduced to make the facial feature map compact using a data dependent algorithm.
- Extensive experiments are performed on seven benchmark datasets, including LFW, CA-LFW, CP-LFW, Age-DB, CFP-FF, CFP-FP, and VGG2-FP. Our FPVT has achieved excellent results compared with existing state of the art methods.

The rest of the paper is arranged as follows. Literature review is summarized in Section 2, the proposed Face Pyramid Vision Transformer (FPVT) is described in Section 3, experiments/results in Section 4, and conclusion and future directions follow in Section 5.

2 Related Work

After the remarkable success of transformers in natural language processing (BERT[5] or GPT [18]), Alexey et al. [6] introduced a transformer for computer vision tasks and obtained superior results compared to the conventional CNN models. In such models, to treat an image as a sentence, it is reshaped into two dimensional flattened patches. Afterwards, similar to the class tokens in BERT, learnable embeddings are added to the embedded patches. Finally, trainable positional embeddings are added on top of patch representations to preserve positional information. Notably, transformer architectures generally rely on self-attention mechanism without utilizing convolutional layers.

ViT [6] models generally require massive amount of training images and, consequently, significantly huge computational cost, which bottlenecks their applicability. To eradicate this issue, a vision transformer architecture was proposed by Hugo et al. [21] that is trained on merely 1.2M images. In this work, the original ViT architecture [6] was modified to adapt teacher-student learning approach while enabling the native distillation process particularly designed for transformers. To this end, the output of the student network is learned from the output of the teacher network. Moreover, a distillation token was added to the transformer, which interacts with classification vectors and image component tokens. In convolution based models, we can generally enhance the performance by adding more convolutional layers. However, transformers are different in this sense and can quickly saturate if the architecture gets deeper. It is because the attention maps become less distinguished as we go deeper into the transformer layers. To eradicate this issue, Daquan et al. [34] proposed a re-attention mechanism, which regenerates the attention map to enhance the diversity between layers at a minute computational burden. This way, the re-attention module was successfully trained on a 32-layer architecture [6] to achieve 1.6% improvement in the top-1 accuracy on Image-Net. Hugo et al. [21] focused on the optimization part in transformers and proposed CaiT which is similar to an encoder-decoder architecture. Hugo et al. [22] also proposed to explicitly split class-attention layers dedicated to extracting the content of the processed patches from transformer layers responsible for self-attention among patches, which further enhances the performance.

Existing transformers [7, 21, 29] have been highly focused on training ViTs from scratch and re-designed the token-to-token process which is helpful in modeling images based on global correlation and local-structure information. This also helps slightly in overcoming the need of deep and hidden layer dimensions. Chen et al. [3] also attempted to reduce the computational complexity of ViTs by introducing dual-branch ViT. The idea was to first extract multi-scale feature representations, then combining them using a cross-attention-based token-fusion mechanism. However, such models are still computationally expensive. Different from these prior works, our FPVT method is resource-efficient, thus works under limited computational resources. FPVT extracts local features while capturing global relationships with fewer parameters than ResNet-18 and recent ViTs.

3 Face Pyramid Vision Transformer

3.1 Overall Architecture

We propose pyramid feature network, which has the capability to extract proportional sized features at high to low levels at different stages. The proposed *Face Pyramid Vision Trans-*

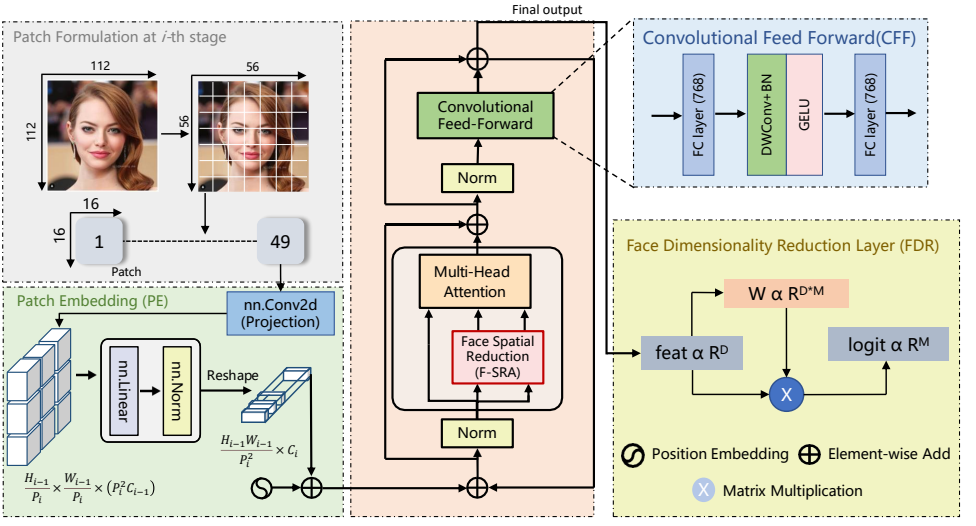

Figure 1: A simplified view of Face Pyramid Vision Transformer (FPVT) capable of training under limited computational resources. Each stage comprises of an improved patch embedding layer and an encoder layer. Following progressive shrinking strategy [25] , the output resolution is diversified at every stage from high to low resolution.

former (FPVT) is a single network for general and age-invariant face recognition (FR). The network diagram of our FPVT is presented in Fig 1. Similar to [15, 25], our FPVT has four different pyramid stages that generate hierarchical feature maps. The construction of our FPVT comprises of improved patch embedding, face spatial reduction attention and convolutional feed-forward network. After that, face dimensionality reduction layer is responsible to compute discriminative compact facial features. At the beginning of our method, given an input face image of size $w \times h \times 3$, we split image into overlapping patches each of size $p_s \times p_s$ with overlap of $q_s \times q_s$ pixels which varies with the variation of stage s. The number of such patches turn out to be $(w/(p_s - q_s) - 1) \times (h/(p_s - q_s) - 1)$ for the stage s. We flatten these patches, feed to an improved patch embedding module, and get embedded patches of size $p_s^2 \times c_s$, where c_s is the number of channels at stage s. Next, positional embedding is attached with embedded patches and fed into encoder module. The output of the encoder is reshaped to be input to the next stage. The reshaped tensor has size $p_s^2 \times c_s$. The features obtained from the stage s are fed into the next stage $s+1$. With pyramidal face features using k stages, features $\{f_1, \cdots f_k\}$ are the output of each stage at a different resolution level. Such an approach is found to be suitable for general as well as age-invariant FR tasks.

3.2 Improved Patch Embedding

Instead of using non-overlapped patches from a face image, we use a simple yet effective technique to increase the performance of ViT's for FR in various scenarios. Motivated by recent work [33], we introduce a token generation strategy in ViTs. We add a token generation scheme in the transformer [25] to generate sliding overlapped patches and use inter-patch information to increase FR performance.

In our Improved Patch Embedding (IPE), we utilize a convolution layer with padding f to

generate these patches. Overlap embedding enables FPVT to extract sequential information from faces while also reducing sequence length and increasing the feature dimension over consecutive stages. It accomplishes spatial down-sampling while simultaneously increasing the number of feature maps. FPVT takes 2d input image from training data with size $h \times w \times c$, and feed into convolution layer with kernel size $2f$+1, stride s, number of kernels p, and padding size f. The final output is $\frac{h}{s} \times \frac{w}{s} \times p$. The IPE layer allows us to adjust the number of visual tokens and feature dimensions at each stage by using a convolution operation.

3.3 Convolutional Feed-Forward Network

When considering global relationships in visual recognition tasks, transformers are the first-priority to create long-range dependencies via self-attention approach. However, transformers require significant computational cost. In order to reduce the computational complexity, we propose light-weight convolutional filters inspired by MobileNet architecture [9]. These filters are helpful in capturing local features from a face image, e.g., forehand lines, nose pattern, nose bridge, and chin. Particularly, we introduce a set of $h = 3 \times w = 3$ filters having cardinality of n_i and the number of input channels and padding of 1. Then, a set of 1×1 depth-wise convolution and cardinalty n_o is applied to conduct across channel convolutions. Our light weight filters require only $hwn_i + n_i n_o$ parameters, which are significantly lesser than $hwn_i n_o$ parameters in the equivalent normal convolution filters.

Our convolutional feed-forward network comprises of one fully connected layer, a light weight convolutional layer, a batch-wise normalization layer, a GELU activation function, followed by another linear layer. Such an architecture brings rich representations and gives low-level information which is not addressed in the previous feed-forward networks. The depth-wise convolution is obtained in two steps. First we convolve each channel $W_q \in \mathbb{R}^{h \times w}$ with a filter $Y_q \in \mathbb{R}^{m \times m}$ to get

$$D_q = W_q \odot Y_q, \text{ where } 1 \leq q \leq n_i, \tag{1}$$

where $W_q \in \mathbb{R}^{h \times w}$ is the result of q-th channel convolution. In the second step, depth-wise convolution is performed by using a set of 1×1 filters $\mathbf{v}_p \in \mathbb{R}^{n_i}$:

$$F_p(a,b) = \sum_{q=1}^{n_i} D(a,b,q)\mathbf{v}_p(q), \text{ where } 1 \leq p \leq n_o \tag{2}$$

F_p is the final output features of our light-weight convolutional feed-forward network. After that, the output features are reshaped to generate a sequence of tokens which are used to feed into the next transformer layer.

3.4 Face Spatial Reduction Attention (F-SRA)

The proposed FPVT encoder at i-th stage has l_i encoder layers and each stage consists of a convolutional feed-forward network and a self-attention [23]. Since FPVT requires to process low-resolution face images for constructing hierarchical feature maps, instead of utilizing standard Multi-Head Attention (MHA) layer, we introduce a simple yet effective Face Spatial-Reduction Attention (F-SRA) layer. Compared to MHA, our F-SRA requires three inputs including query q, key k, and value v while the output consists of refined features. Before attention process, our F-SRA decreases spatial scale of k and v. In this way,

F-SRA has low computational cost and reduced memory overhead. Our F-SRA at the i-th stage having c_i channels and c_i heads is given by the concatenation of all heads:

SRA$(q,k,v) = [h_0, h_1, \cdots, h_j, \cdots h_{c_i}]w_o$, where w_o is a linear projection matrix, and h_j is the output of $j-$th head:

$$h_j = \mathrm{Att}(\mathrm{qw_j^q}, \mathrm{s_r(k)w_j^k}, \mathrm{s_r(v)w_j^v}). \tag{3}$$

Where $\mathrm{Att}(q,k,v) = \mathrm{Softmax}(\frac{qk^\top}{\sqrt{d_\mathrm{h}}})v$, and the parameters for linear projection operation are $w_j^q \in \mathbb{R}^{c_i \times d_h}$, $w_j^k \in \mathbb{R}^{C_i \times d_h}$, $w_j^v \in \mathbb{R}^{c_i \times d_h}$. Dimension of head, d_h is equal to $\frac{c_i}{n_i}$. $s_r(\cdot)$ is the technique for decreasing the spatial dimension of input sequence k and v

$$s_r(x_i, r_i, w_s) = \mathrm{Norm}(\mathrm{Reshape}(x_i, r_i)w_s). \tag{4}$$

where $x_i \in \mathbb{R}^{(h_i w_i) \times c_i}$ denotes an input sequence, and r_i represents the reduction ratio of the attention layers at ith stage. $\mathrm{Reshape(x, r_i)}$ is the method to reshape the input sequence x_i to a sequence of size $\frac{h_i w_i}{r_i^2} \times (r_i^2 c_i)$. $w_s \in \mathbb{R}^{(r_i^2 c_i) \times c_i}$ is a linear projection that decreases the dimension of the input sequence to c_i. $\mathrm{Norm}(\cdot)$ represents layer normalization. Through the above mathematical representations, we can compute the memory-cost of attention operation which is r_i^2 smaller than MHA. Despite the fact that, attention mechanism have strong potential for learning global relationship, the computational overhead of feature maps is still expensive. Thus, we utilize adaptive max pooling layer with output size 7 over an input feature before attention operation. This technique brings a substantial reduction in parameters and handles large feature maps with less computational resources.

$$Y = \text{F-SRA}(\mathrm{AdaptiveMaxPool}(\mathrm{SRA}(\mathbf{q}, \mathbf{k}, \mathbf{v}))), \tag{5}$$

This further reduces the size of input matrix by a factor of n_{mp}^2, where $n_{mp} \times n_{mp}$ is the size of max pooling filter.

3.5 Face Dimensionality Reduction Layer

While FPVT extracts multi-scale features, we also require dimensionality reduction mechanism for training ultra large-scale dataset with limited hardware costs. Inspired by the recent advances in computationally efficient FR methods [13, 24], we introduce a Face Dimensionality Reduction (FDR) layer in the ViT stream which reduces training time while also maintaining superior accuracy.

In the training phase, FDR layer randomly splits k training identities (categories) into m_g groups. The categories from m_g share the l_{th} column in projection matrix w. In this paper, l_{th} column is defined as $anchor$ and w contains anchors shared by m_g. To optimize w, we initialize two anchors, corresponding anchor $anch_{corr}$, and free anchor $anch_{free}$. If mini-batch carries categories from the l_{th} column then $anch_l$ is of type $anch_{corr}$. If not, it is marked as $anch_{free}$.

Corresponding Anchor. If mini-batch carries a category from m_g, $anch_l$ is placed in $anch_{corr}$ in that iteration. With each column in w of the last FC layer representing the centroid of each category, the equation of $anch_{corr}$ can be written as:

$$anchor_{corr,l} = \sum_{i=1}^{K} \alpha_{i,l} f_{i,m} / \sum_{i}^{K} \alpha_{i,l} \tag{6}$$

$f_{i,\,m}$ is feature representation of the l_{th} face belong to m_g. We assume that there is no conflict among anchors. $\{f_{i,m}\}(i = 1, 2, 3, 4 \ldots, k)$ represents an individual identity. $\alpha_{i,l}$ is an estimated attention factor, $f_{i,m} \cdot \{\alpha_{i,l}\}$ is estimated through attention process or set as a constant value. **Free Anchor:** Due to the limited resources, it is not possible to set a larger batch size and $anch_{corr}$ is also restricted by batch size. To overcome this limitation, the concept of free anchors is introduced. If face does not exist in m_g in a given iteration, $anch_l$ will be free and represented as $anch_{free,l}$. This way, it cannot be calculated by Equation 6 due to $f_{i,l} \in \theta$. The concept of free anchors help in ending the restriction of large batch number in the same way as traditional FC layers. Inter-identity representation can also be dispersed among mini-batches. Moreover, the number of m_g could be set independently based on computational resources and accuracy. Free anchors are not restricted by the number of samples or batch size. The FDR layer is comparably better than a traditional FC layer especially with limited hardware resources. The kernel of FDR layer is $w \in \mathbb{R}^{d \times m}$ where m is the hyperparameter which depend on the balance between hardware resources and performance and can be set freely. The number of m must be less than the training categories n. The final output of FDR layer is represented as: $y = w^T f + b$. Here, $y \in \mathbb{R}^m$ is final output, b is bias, and $f \in \mathbb{R}^d$ is feature. In the representation learning case, b can be written as zero.

4 Experiments

We performed extensive experiments to evaluate our proposed FPVT on several benchmark datasets and compared with CNN based methods [4, 7, 10], pure ViT methods [6, 22, 30, 34], and Convolutional ViTs [8, 26, 28].

4.1 Implementation Details

Following previous ViT works [33], we use adamW optimizer with initial LR $3e-4$ for all pure ViTs and ConViTs experiments. For CNN works [10], we use SGD and set the LR to be 0.1 with momentum set to 0.9 and use weight decay 0.05. We train FPVT along with all methods for 60 epochs on the face scrub dataset. We use one Nvidia Tesla V100 supporting a batch size of 496 in all of our experiments. We use ArcFace as a classification head in CNNs and ViTs. We employ standard data augmentation techniques which include resizing, random crop, and random horizontal flip. **Training Dataset:** Face Scrub dataset contains $107,818$ images of 265 male and 265 female celebrities collected from different sources on the internet. For training, a cleaned and aligned version of the dataset is used which includes $91,712$ images of 263 males and 263 females. **Testing Dataset:** Tests are conducted on various databases including Age-DB [17] for age-invariant, LFW for unconstrained FR [11], CFP-FP [19] for frontal-profile, CP-LFW [31] for a crosspose, CA-LFW [32] for cross-age, and VGG2-FP is for frontal pose). **LFW** [11] dataset exhibits natural pose variations, focus, lighting, resolution, make-up, occlusions, background, facial expression, age, gender, race, accessories, and photographic quality variations. It comprises 13,233 images of faces gathered from online websites. **CA-LFW** purely consists of $3,000$ positive face aging pairs. It has been split into 10 distinct folds using the similar identities included in the LFW 10 folds. This database consists of $4,025$ individuals with 2, 3 or 4 face images for each identity. **CP-LFW** [31] is developed to consider pose-related images in FR. It is an extended version of the LFW dataset which consists of $11,652$ face images, $3,968$ unique identities, and two to three images per person. **Age-DB** is manually collected data in wild with noise-free labels. It contains $16,488$ images of various actress/actresses, writers, politicians and etc. It consists

of 568 unique identities, 29 images per person and age ranges start from 1 to 101 years. **CFP** contains 7,000 face images, 500 unique identities and number of images per subject is 14 [19].The database is split into 10 subsets with a pairwise separate set of identities in each split. Every subset comprises 50 individuals and 7,000 pairs of faces for frontal-frontal and frontal-profile (CFP-FP) experiments. **VGG2-FP** is extracted from large-scale VGGFace2 [1] dataset consists of 3.31M images of 9131 categories with large range of ethnicity age and pose. It is specially designed for frontal-profile faces and it consists of 10,000 images of 300 identities with different variation.

4.2 Quantitative and Qualitative Comparison

FPVT is compared with IR-18 and IR-SE-18 which are industrial benchmarks for the FR tasks. To compare FPVT with pure ViTs, we utilize ViT [6], DeepViT [34], and CaiT [22]. Further, we compare FPVT with three convolutional ViTs namely PiT [8], CeiT [28] and CvT [26]. For a fair comparison, we evaluate all models using popular FR metric Face Verification Accuracy (FVF). Table. 2 shows FPVT outperforms existing methods including pure ViTs, ConViTs, and CNNs in terms of face verification accuracy, on three datasets. The higher VA indicates the capability of FPVT to verify general and age-invariant faces. As we can see, ConViTs outperformed the pure ViTs models and are proven to be the robust ConViT models CeiT [28] and CvT [26], which require a large number of parameters to produce superior results. In contrast, FPVT does not need further training data and the number of parameters is less than existing models.

	Methods	LFW ($family$)			Age-DB
		LFW	CA	CP	
CNN	ResNet-18[9]	76.7	60.7	58.1	61.4
	IR-50 [9]	91.7	78.1	68.9	73.4
	IR-SE-50 [10]	90.5	65.8	68.7	65.8
PureViT	DeepViT [34]	75.5	62.6	57.1	59.7
	CaiT [22]	83.4	71.5	57.5	62.2
	ViT [6]	81.9	67.7	58.9	61.4
	ViT [6]+**IPE**	82.5	68.5	61.1	63.1
ConViT	PiT [8]	80.6	66.6	58.7	64.6
	CvT [26]	82.5	69.1	57.1	63.7
	CeiT [28]	84.8	72.6	60.1	65.8
	PVT [25]	78.8	66.8	55.1	59.9
	+IPE	82.9	70.1	59	65.6
	+CFFN	86.7	72.9	62.1	68.9
	+FDR	87.4	73.9	61.6	70.1
	+OA	91.4	77.4	68.9	74.5
	FPVT	**92.0**	77.0	67.8	**75.0**

Table 1: Face verification accuracy on LFW, CA-LFW, CP-LFW and Age-DB: Comparison with FPVT, CNNs, PureViTs and Convolutional ViTs methods.

4.3 Ablation Study

We conduct multiple ablation studies to validate the impact of our proposed work in our FPVT modules. Table. 1 and Table 2 present the results of CNNs, pure ViTs, and Convolutional ViTs on seven datasets. PVT (referred to as baseline) is a standard pure pyramid transformer without convolution. We choose PVT as a baseline due to two main reasons: i) It generates multi-scale features. ii) The number of parameters is larger than ResNet18. The introduction of the IPE block leads to a gain of 1% in terms of performance, highlighting the impact of convolutional tokens. IPE block improves the performance on LFW from 78.8% to 82.9%, CFP-FF from 75.2% to 85.5%, CFP-FP from 52.9% to 65.6%, Age-DB from 59.9% to 65.6%, CA-LFW from 66.8% to 70.1%, and CP-LFW from 55.1% to 59%. On the VGG-FP dataset, IPE increases performance with little margin on VGG2-FP from 57.1% to 62.2%. Overall, IPE improves average performance by 4.5%. The introduction of convolutional FFN in FPVT improves the structural and local relationship between different parts of faces. Interestingly, CFNN adds significant performance gain on all datasets: LFW (3.8%), CFP-FF

(1.1%), CFP-FP (4.6%), Age-DB (3.3%), CA-LFW (2.8%), CP-LFW (3.1%) and VGG2-FP (3.9%). We also evaluate the influence of the FDR layer on FPVT performance (see in Table. 2 and Table. 1). While retaining the same training and implementation details, we replace the previous layer with our "+IPE+CFFN" and it gradually increases the performance on six datasets. The introduction of the FDR layer in FPVT discriminates the features among identities that lead to performance improvement in six datasets. As mentioned in Table. 2 and Table. 1, FDR improves the accuracy on LFW from 86.7% to 87.4%, CFP-FF from 86.6% to 87.4%, CFP-FP from 61.5% to 61.5%, Age-DB from 68.9% to 70.1%, CA-LFW 72.9% to 73.9%, and VGG-FP 66% to 66.1%. However, the accuracy of CP-LFW slightly decreases from 62.1% to 61.1%.

The introduction of data augmentation and F-SRA improves linear computation and reduces the number of parameters. The final number of parameters of FPVT is decreased from 33.3M to 28.8M which is smaller than the recent pure ViTs, ConViTs, and CNNs. Further, we adopt online data augmentation by using some off-the-shelf techniques. As shown in Table. 1, the accuracy improvement is observed as: LFW (87.4% to 91.4%), CFP-FF (87.4% to 90%), CFP-FP (61.5% to 71.8%), Age-DB (70.1% to 74.5%), CA-LFW (73.9% to 77.4%), CP-LFW (61.6% to 68.9%) and VGG2-FP (66.1% to 75.3%). Overall, online augmentation significantly increases performance on all datasets. While F-SRA reduces the overall number of parameters parameters, it increases the accuracy on individual datasets such as on LFW (91.4% to 92.0%), CFP-FF (90% to 90.3%), CFP-FP (71.8% to 73.3%), Age-DB (74.5% to 77%).

	Methods	Dim	Depth	Param	CFP (*family*) FF	CFP (*family*) FP	VGG2-FP
CNN	ResNet-18[9]	-	-	30.7M	76.7	52.2	61.4
	IR-50 [9]	-	-	65.1M	91.7	74.2	73.4
	IR-SE-50 [10]	-	-	65.5M	90.5	71.6	65.8
PureViT	DeepViT [30]	512	6	11.6M	75.5	56.1	59.7
	CaiT [29]	512	3	7.8M	83.4	56.6	62.2
	ViT [8]	512	6	17.8M	81.9	58.9	61.4
	ViT [8]+IPE	512	6	17.9M	82.5	60.6	63.1
ConViT	PiT [8]	64	20	12.5M	80.6	57.2	64.6
	CvT [27]	64	10	19.8M	82.5	56.4	63.7
	CeiT [28]	64	20	21.5M	84.8	59.1	65.8
	PVT [26]	512	18	32.2M	78.8	52.9	59.9
	+IPE	512	6	33.3M	82.9	56.4	65.6
	+CFFN	512	6	33.3M	86.7	61	68.9
	+FDR	512	6	33.3M	87.4	61.5	70.1
	+OA	512	6	33.3M	91.4	71.8	74.5
	FPVT	**512**	**6**	**28.2M**	**92.0**	**73.3**	**75.0**

Table 2: Face verification accuracy of models with different dimensions, depths, and parameters on CFP-FF, CFP-FP and VGG2-FP.

5 Conclusion

A Face Pyramid Vision Transformer (FPVT) is proposed for FR and verification tasks. Within the FPVT framework, a convolutional feed-forward network is used to encode local structural relations among different facial parts and to maintain long range relations. To ensure parameters reduction, a Face-Spatial Reduction Attention layer is introduced in the encoder that efficiently decreases the number of parameters. Additionally, a Face Dimensionality Reduction (FDR) layer is used to ensure facial feature map compactness. The proposed FPVT is evaluated on seven datasets and compared with ten SOTA methods. The experiments have exhibited the robustness of the proposed algorithm.

References

[1] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74. IEEE, 2018.

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, 2020.

[3] Chun-Fu Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. *International Conference on Computer Vision*, 2021.

[4] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *IEEE/CVF Computer Vision and Pattern Recognition Conference*, 2019.

[5] 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.

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations*, 2021.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE/CVF Computer Vision and Pattern Recognition Conference*, 2016.

[8] Byeongho Heo, Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Junsuk Choe, and Seong Joon Oh. Rethinking spatial dimensions of vision transformers. *International Conference on Computer Vision*, 2021.

[9] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[10] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *IEEE/CVF Computer Vision and Pattern Recognition Conference*, 2018.

[11] 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*, 2008.

[12] Pengyu Li, Biao Wang, and Lei Zhang. Virtual fully-connected layer: Training a large-scale face recognition dataset with limited computational resources. In *IEEE/CVF Computer Vision and Pattern Recognition Conference*, 2021.

[13] Wenyu Li, Tianchu Guo, Pengyu Li, Binghui Chen, Biao Wang, Wangmeng Zuo, and Lei Zhang. Virface: Enhancing face recognition via unlabeled shallow data. In *IEEE/CVF Computer Vision and Pattern Recognition Conference*, 2021.

[14] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. *arXiv preprint arXiv:2203.16527*, 2022.

[15] 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. *International Conference on Computer Vision*, 2021.

[16] Zhisheng Lu, Hong Liu, Juncheng Li, and Linlin Zhang. Efficient transformer for single image super-resolution. *arXiv preprint arXiv:2108.11084*, 2021.

[17] Stylianos Moschoglou, Athanasios Papaioannou, Christos Sagonas, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. Agedb: the first manually collected, in-the-wild age database. In *IEEE/CVF Computer Vision and Pattern Recognition ConferenceW*, 2017.

[18] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[19] 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 *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2016.

[20] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7262–7272, 2021.

[21] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. *International Conference on Machine Learning*, 2020.

[22] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. *International Conference on Computer Vision*, 2021.

[23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Neural Information Processing Systems*, 2017.

[24] Kai Wang, Shuo Wang, Zhipeng Zhou, Xiaobo Wang, Xiaojiang Peng, Baigui Sun, Hao Li, and Yang You. An efficient training approach for very large scale face recognition. *arXiv preprint arXiv:2105.10375*, 2021.

[25] 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. *International Conference on Computer Vision*, 2021.

[26] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. *International Conference on Computer Vision*, 2021.

[27] Yu-Huan Wu, Yun Liu, Xin Zhan, and Ming-Ming Cheng. P2t: Pyramid pooling transformer for scene understanding. *arXiv preprint arXiv:2106.12011*, 2021.

[28] Kun Yuan, Shaopeng Guo, Ziwei Liu, Aojun Zhou, Fengwei Yu, and Wei Wu. Incorporating convolution designs into visual transformers. *International Conference on Computer Vision*, 2021.

[29] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. *International Conference on Computer Vision*, 2021.

[30] Zizhao Zhang, Han Zhang, Long Zhao, Ting Chen, and Tomas Pfister. Aggregating nested transformers. *arXiv*, 2021.

[31] Tianyue Zheng and Weihong Deng. Cross-pose lfw: A database for studying cross-pose face recognition in unconstrained environments. *BUPT*, 2018.

[32] Tianyue Zheng, Weihong Deng, and Jiani Hu. Cross-age lfw: A database for studying cross-age face recognition in unconstrained environments. *arXiv*, 2017.

[33] Yaoyao Zhong and Weihong Deng. Face transformer for recognition. *arXiv*, 2021.

[34] Daquan Zhou, Bingyi Kang, Xiaojie Jin, Linjie Yang, Xiaochen Lian, Qibin Hou, and Jiashi Feng. Deepvit: Towards deeper vision transformer. *International Conference on Computer Vision*, 2021.

[35] Zheng Zhu, Guan Huang, Jiankang Deng, Yun Ye, Junjie Huang, Xinze Chen, Jiagang Zhu, Tian Yang, Jiwen Lu, Dalong Du, et al. Webface260m: A benchmark unveiling the power of million-scale deep face recognition. In *IEEE/CVF Computer Vision and Pattern Recognition Conference*, 2021.