Face Pyramid Vision Transformer

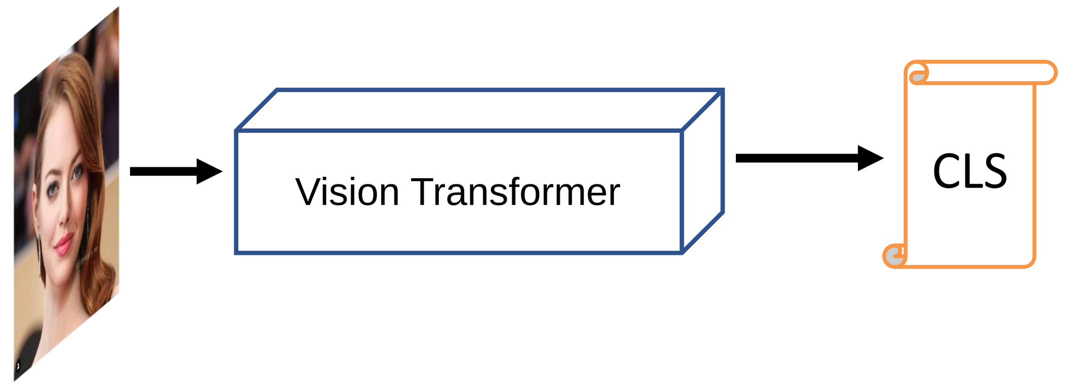
Khawar Islam, Muhammad Zaigham Zaheer and Arif Mahmood

INFORMATION TECHNOLOGY UNIVERSITY · ETRI 한국전자통신연구원 Electronics and Telecommunications Research Institute · UST · MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE

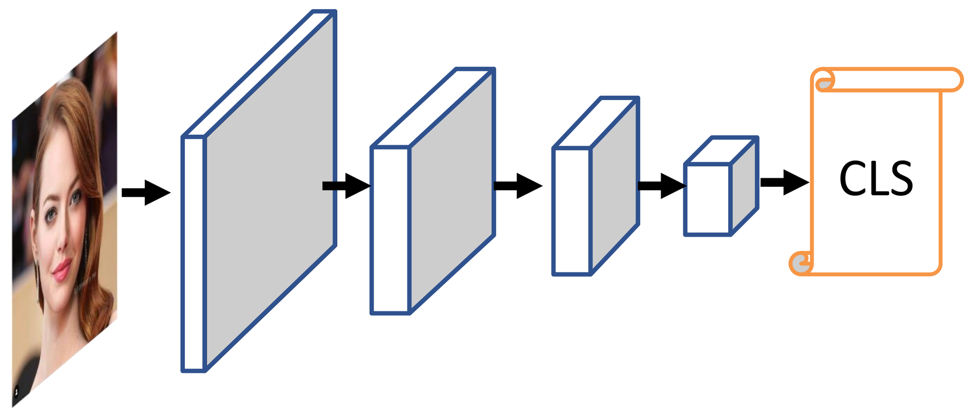
BMVC 2022

Motivation

- ❑ Vision transformers for general face recognition (FR) and age-invariant FR is not well-studied.
- ❑ Working with limited computational resources and medium-scale datasets are critical challenges for FR.
- ❑ Are vision transformers better for recognizing general and age-invariant face recognition?
- ❑ How much training data would ViT require to obtain state-of-the-art results on such tasks?
- ❑ How can we develop specific designed for FR?



In traditional ViT, we have single scale features containing MHA and a feed forward network.



In pyramid networks, high- and low-level features are merging for better face recognition.

Related Work

- ❑ Transformer for Face Recognition
  - ViT-P (arxiv'21), investigated first ViT for face recognition

Goal

- ❑ Learning multi-scale features with global and local context.
- ❑ Simplified single architecture for general and age-invariant face recognition

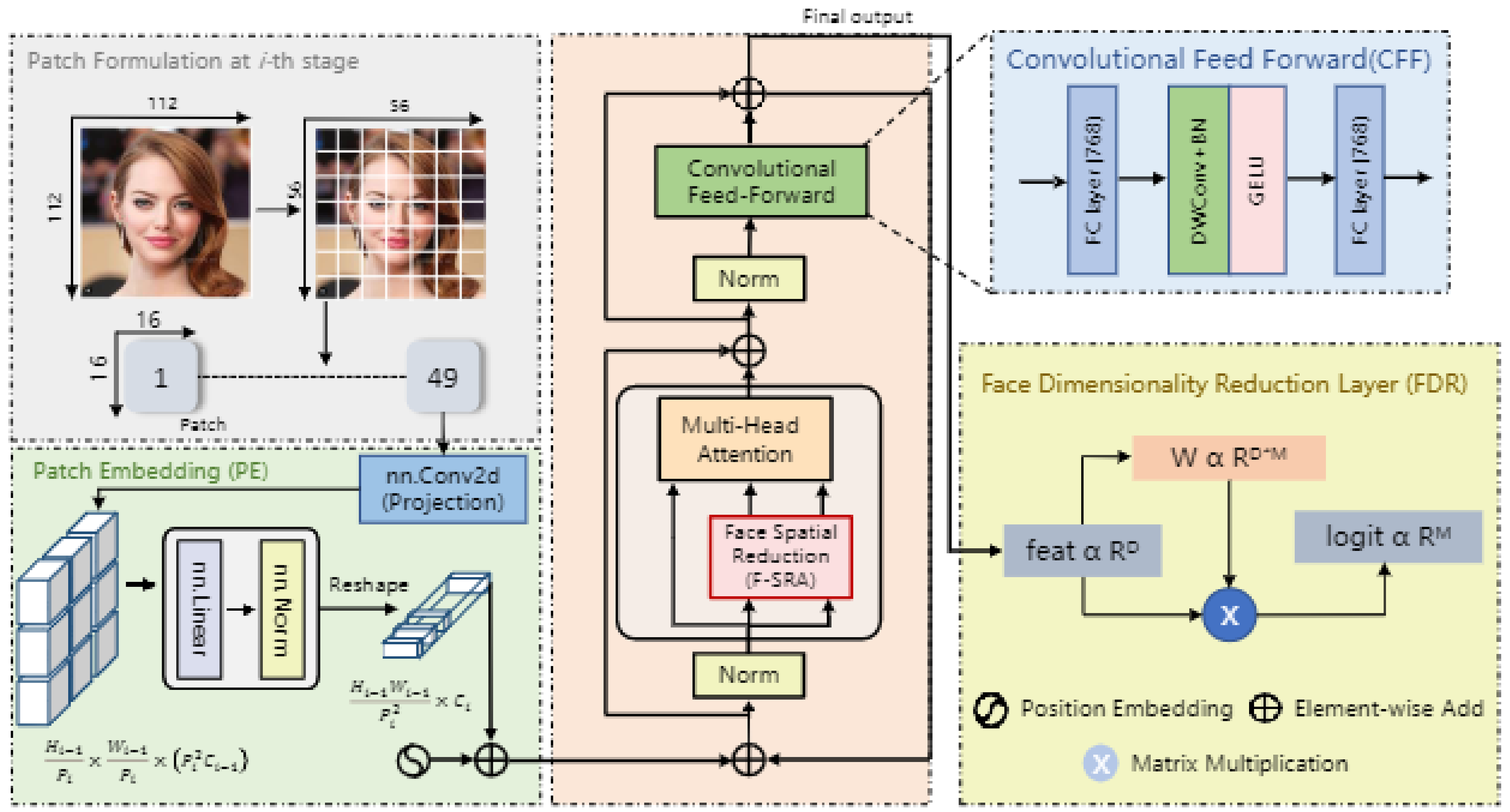
Challenges

- ❑ Conventional ViTs fail when train on limited data under limited computational resources.
- ❑ Extracting global context while ignoring local features and information.
- ❑ Pure ViTs do not improve performance against CNNs for FR.

Contributions

- ❑ First attempt to learn multi-scale discriminative features.
- ❑ Considering benefits of CNNs to model lower-level edges to higher-level semantic primitives.
- ❑ Capturing local representations while considering long-range relationships.
- ❑ Reduce the computations of large feature maps via simplified MHA.
- ❑ Make the facial feature map compact using a data dependent algorithm.
- ❑ Extensive experiments on LFW, CA-LFW, CP-LFW, Age-DB, CFP-FF, CFP-FP, and VGG2-FP datasets.

Our Architecture



- ❑ Simplified view of our 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, the output resolution is diversified at every stage from high to low resolution.
- ❑ FPVT is capable of computing discriminative compact facial features.

Results

Methods	LFW (family)			Age-DB
	LFW	CA	CP	
ResNet-18	76.7	60.7	58.1	61.4
IR-50	91.7	78.1	68.9	73.4
IR-SE-50	90.5	65.8	68.7	65.8
DeepViT	75.5	62.6	57.1	59.7
CaiT	83.4	71.5	57.5	62.2
ViT	81.9	67.7	58.9	61.4
ViT+IPE	82.5	68.5	61.1	63.1
PiT	80.6	66.6	58.7	64.6
CvT	82.5	69.1	57.1	63.7
CeiT	84.8	72.6	60.1	65.8
PVT	**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**

Methods	Dim	Depth	Param	CFP (family)		VGG2-FP
				FF	FP	
ResNet-18	-	-	30.7M	76.7	52.2	61.4
IR-50	-	-	65.1M	91.7	74.2	73.4
IR-SE-50	-	-	65.5M	90.5	71.6	65.8
DeepViT	512	6	11.6M	75.5	56.1	59.7
CaiT	512	3	7.8M	83.4	56.6	62.2
ViT	512	6	17.8M	81.9	58.9	61.4
ViT+IPE	512	6	17.9M	82.5	60.6	63.1
PiT	64	20	12.5M	80.6	57.2	64.6
CvT	64	10	19.8M	82.5	56.4	63.7
CeiT	64	20	21.5M	84.8	59.1	65.8
PVT	**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**

- ❑ Plugged module by module
- ❑ CFNN and OA add significant accuracy gains on all dataset
- ❑ FSRA decreased parameters from 33.3M to 28.8M.
- ❑ IPE increases performance on six datasets.