SWIN FACE

A Multi-task Transformer for Face Recognition

Matheus Levi, 26 de dezembro de 2024

SOBRE O ARTIGO

SwinFace: A Multi-task Transformer for Face Recognition, Expression Recognition, Age Estimation and Attribute Estimation

Publisher: IEEE | Cite This | PDF

Lixiong Qin ; Mei Wang ; Chao Deng ; Ke Wang ; Xi Chen ; Jiani Hu ; Weihong Deng All Authors

Abstract:
In recent years, vision transformers have been introduced into face recognition and analysis and have achieved performance breakthroughs. However, most previous methods generally train a single model or an ensemble of models to perform the desired task, which ignores the synergy among different tasks and fails to achieve improved prediction accuracy, increased data efficiency, and reduced training time. This paper presents a multi-purpose algorithm for simultaneous face recognition, facial expression recognition, age estimation, and face attribute estimation (40 attributes including gender) based on a single Swin Transformer. Our design, the SwinFace, consists of a single shared backbone together with a subnet for each set of related tasks. To address the conflicts among multiple tasks and meet the different demands of tasks, a Multi-Level Channel Attention (MLCA) module is integrated into each task-specific analysis subnet, which can adaptively select the features from optimal levels and channels to perform the desired tasks. Extensive experiments show that the proposed model has a better understanding of the face and achieves excellent performance for all tasks. Especially, it achieves 90.97% accuracy on RAF-DB and 0.22 ε-error on CLAP2015, which are state-of-the-art results on facial expression recognition and age estimation respectively. The code and models will be made publicly available at https://github.com/lxq1000/SwinFace.

Published in: IEEE Transactions on Circuits and Systems for Video Technology (Early Access)

INTRODUÇÃO

OBJETIVOS

- Aproveitamento de Features Faciais para diversos problemas
- Aumento da eficiência do modelo por ser multi-tarefa
- Compreender o funcionamento de um transformer em redes multi-tarefa
- Estudar as relações e **conflitos** das features faciais

TAREFAS

- **RECONHECIMENTO FACIAL:** Reconhecer a face de um indivíduo
- **RECONHECIMENTO DE EXPRESSÃO:** Reconhecer expressões faciais de qualquer indivíduo
- **ESTIMAÇÃO DE IDADE:** Estimar a idade de um indivíduo baseado na face
- **ESTIMAÇÃO DE ATRIBUTOS:** Avaliar atributos específicos presentes na face de um indivíduo

ATRIBUTOS

Grupo	Tarefas	N°
Expression	Expression, Smiling	2
Age	Age, Young	2
Gender	Male	1
Whole	Attractive, Blurry, Chubby, Pale Skin...	6
Hair	Bald, Bangs, Hat...	10
Eyes	Arched eyebrows, bags, eyeglasses...	5

Grupo	Tarefas	N°
Nose	Big Nose, Pointy Nose	2
Cheek	High Cheeks, Rosy Cheeks, Earrings...	4
Mouth	Mustache, Lipstick, Open...	6
Chin	Double Chin, Goatee	2
Neck	Necklace, Necktie	2
Total	----------	42

PROBLEMA

RECONHECIMENTO FACIAL

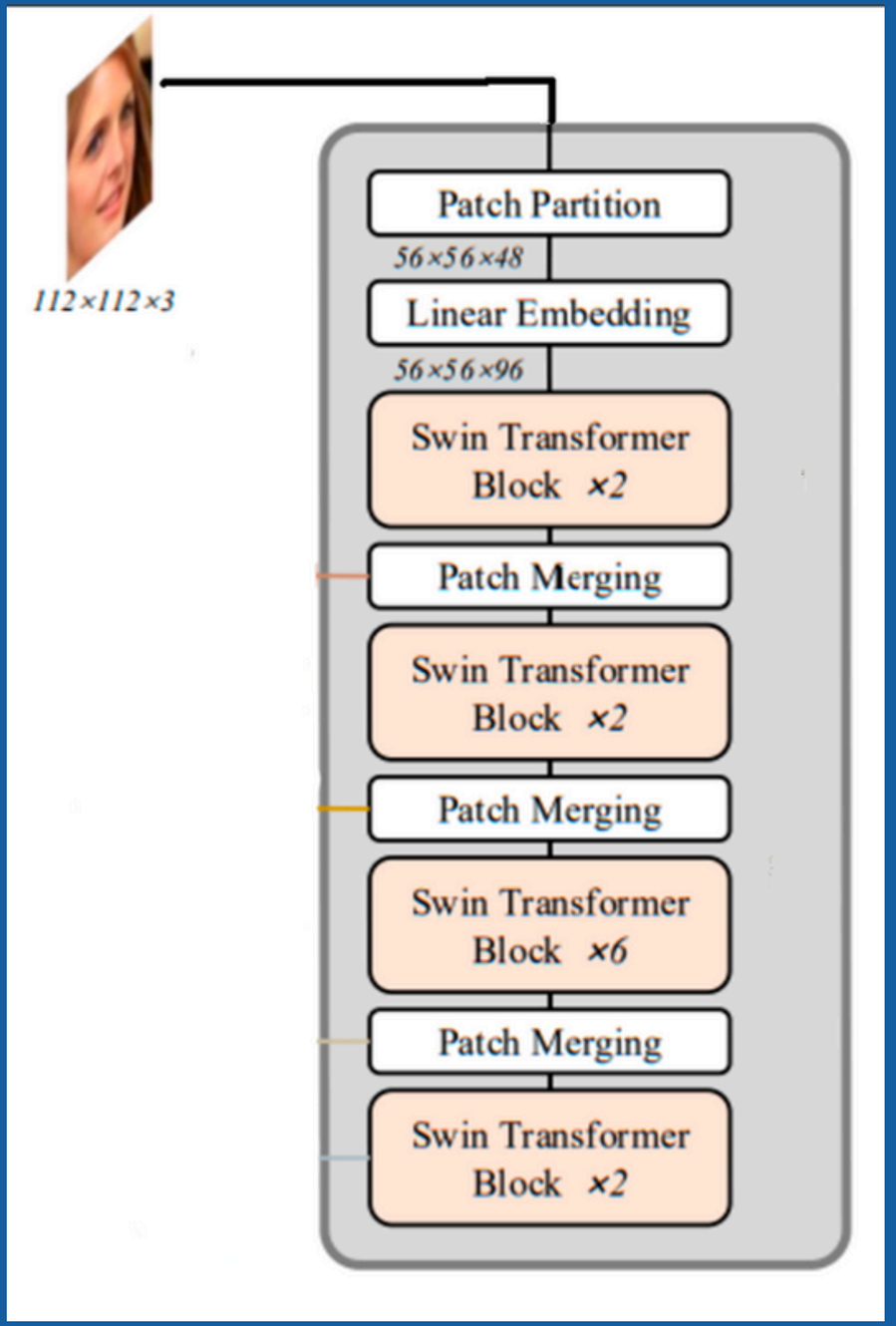
- Identificação de um indivíduo
- Expressão-independente
- Baixa variação intra-classe

RECONHECIMENTO DE EXPRESSÕES

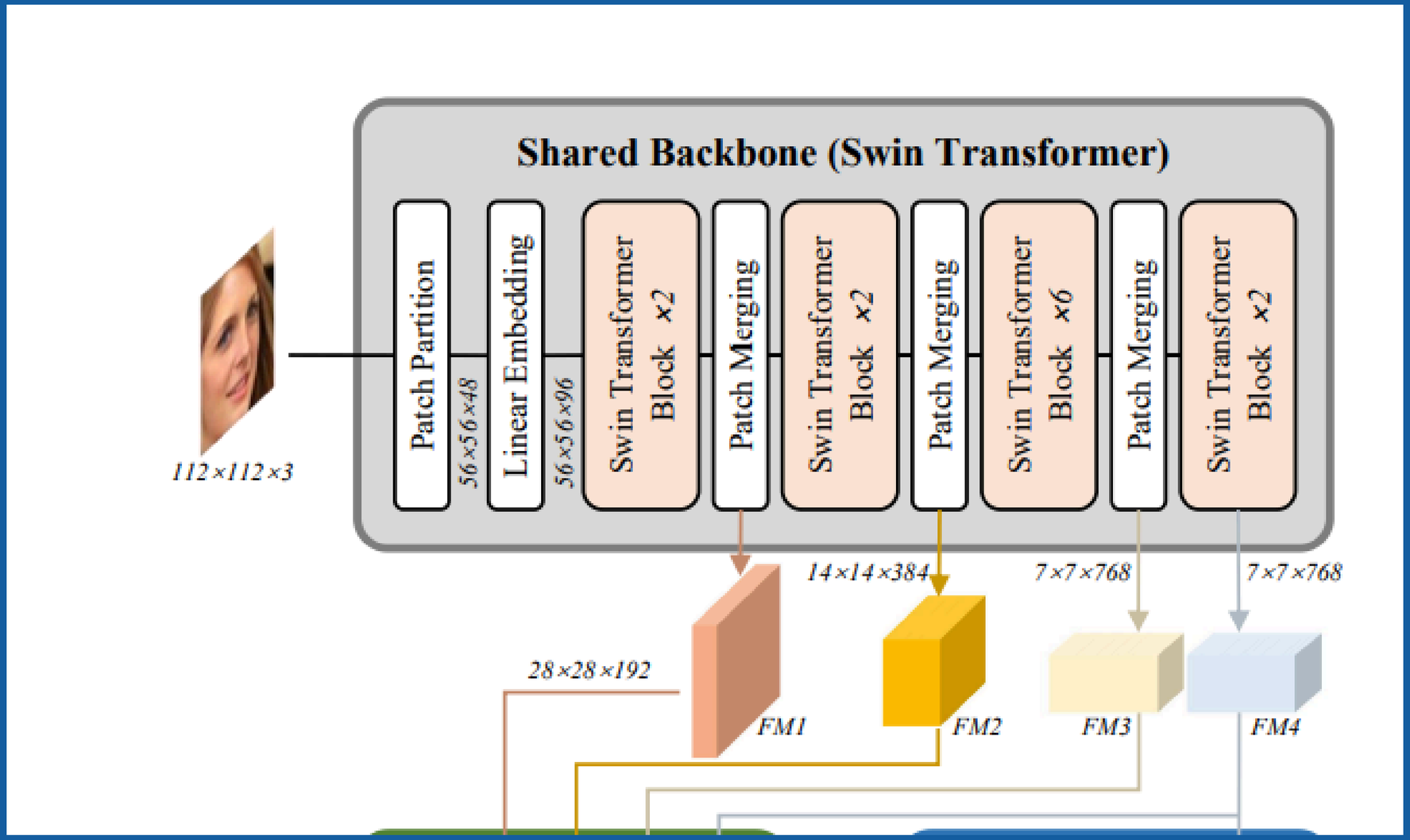
- Identificação de uma expressão
- Indivíduo-independente
- Alta variação intra-classe

ARQUITETURA - BACKBONE

- Patch Partition - Divide a imagem em patches 2x2
- Linear Embedding - Projeta os patches
- Swin Transformer Block - Transformer de visão
- Patch Merging - Reduz os patches em 4 e dobra a dimensão

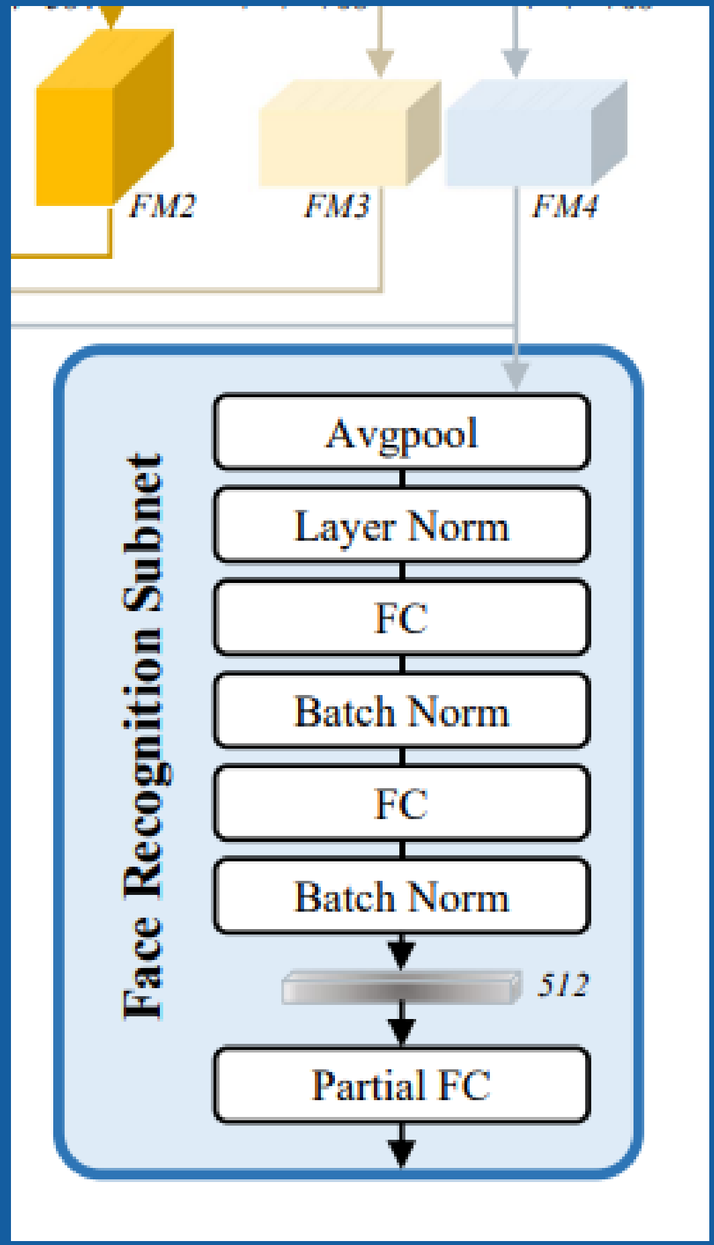


MAPAS DE FEATURES



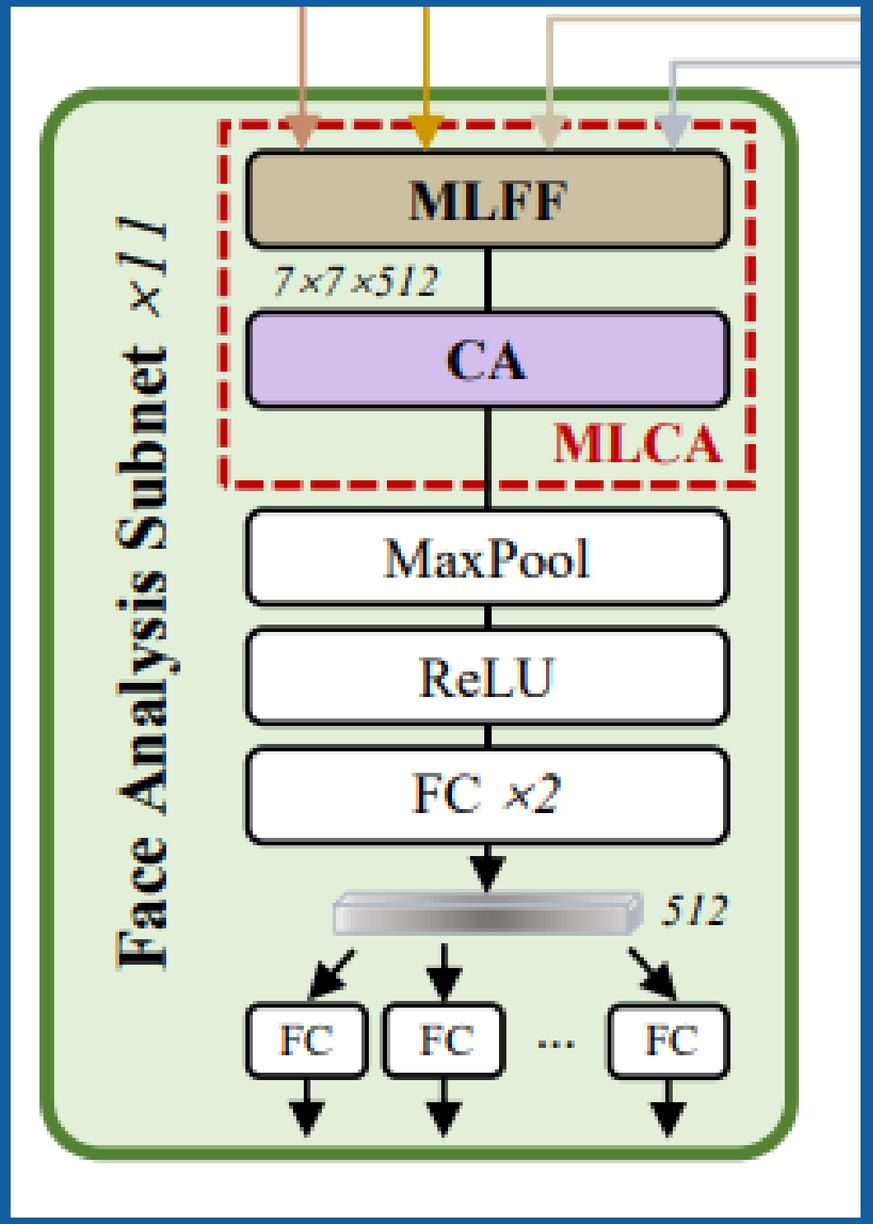
SUB-REDE DE RECONHECIMENTO

- Estrutura similar a ArcFace
- Recebe apenas o FM4

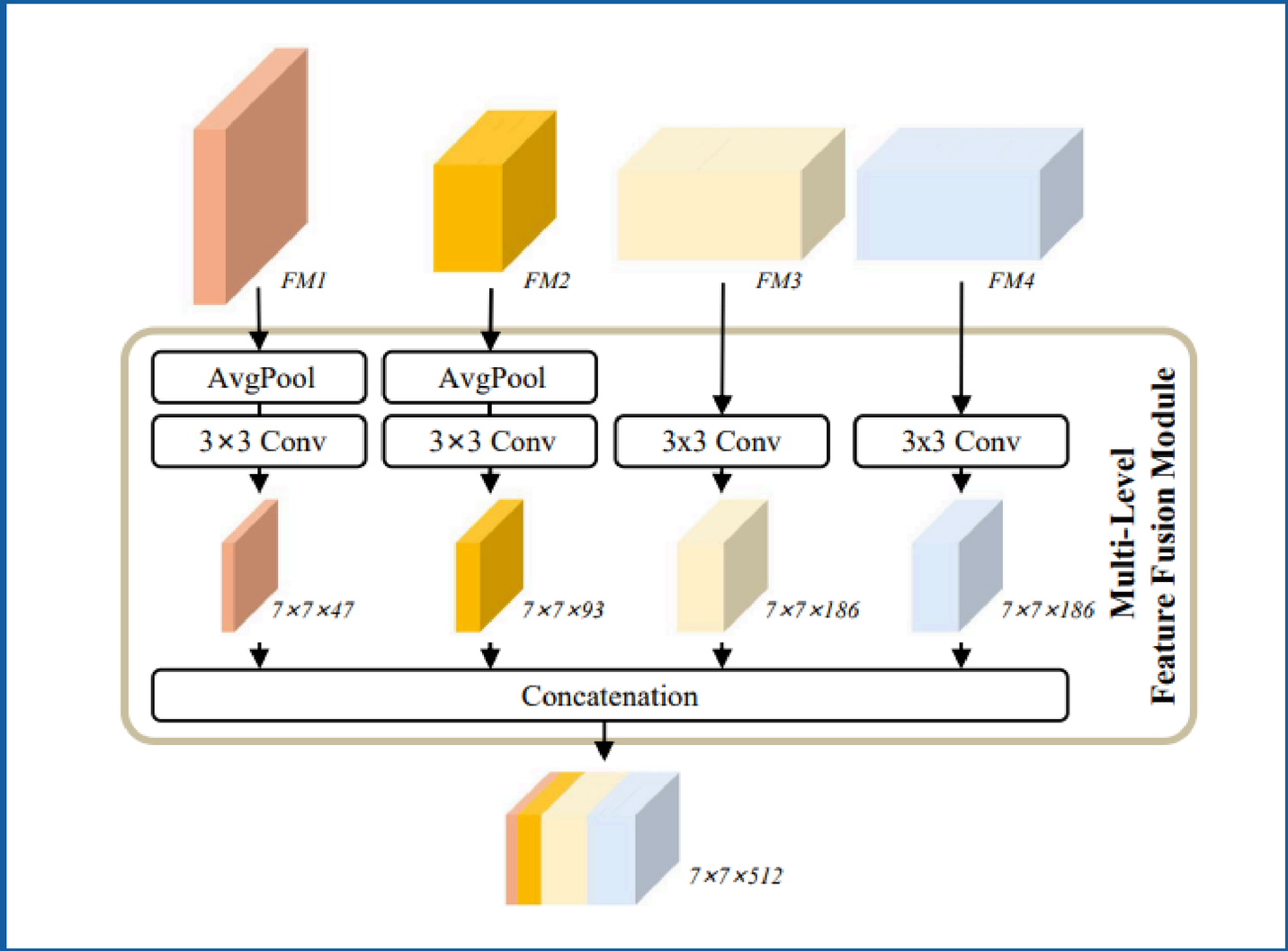


SUB-REDES DE ANÁLISE

- Multi-Level Channel Attention (MLCA)
- Cada grupo possui sua sub-rede

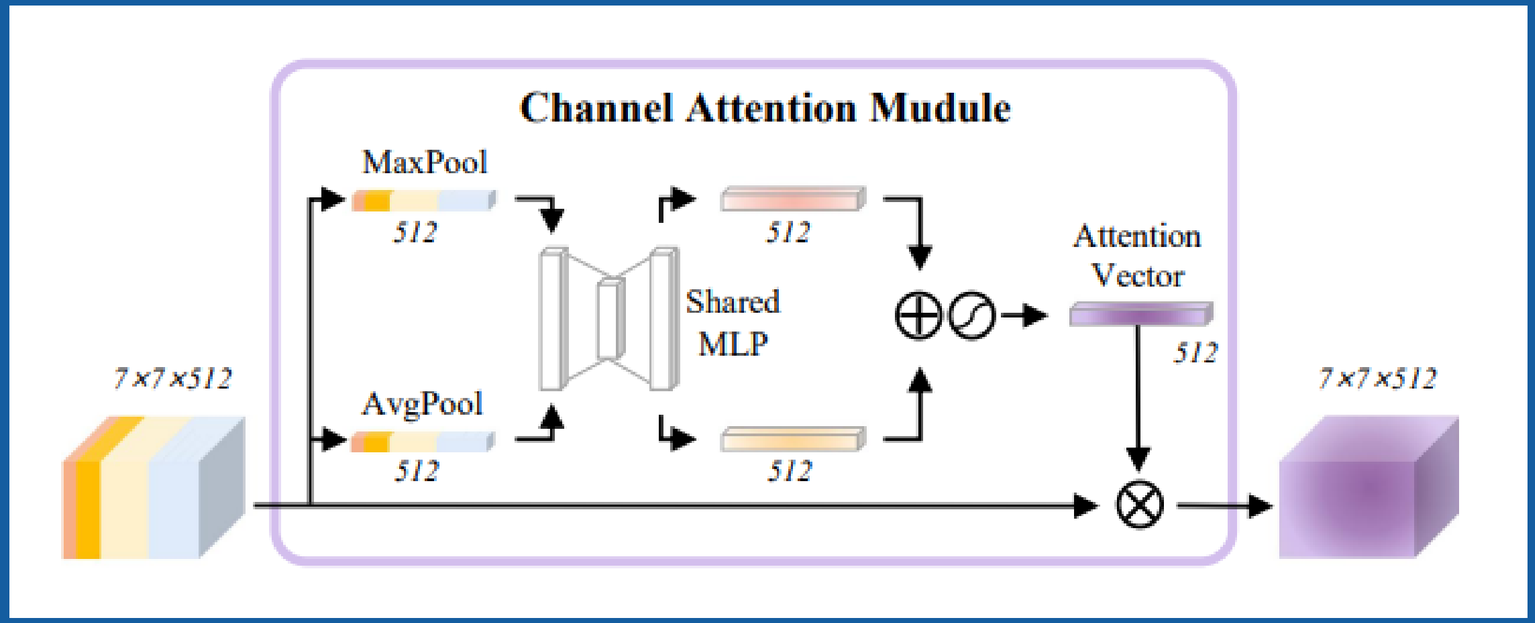


MULTI-LEVEL FEATURE FUSION

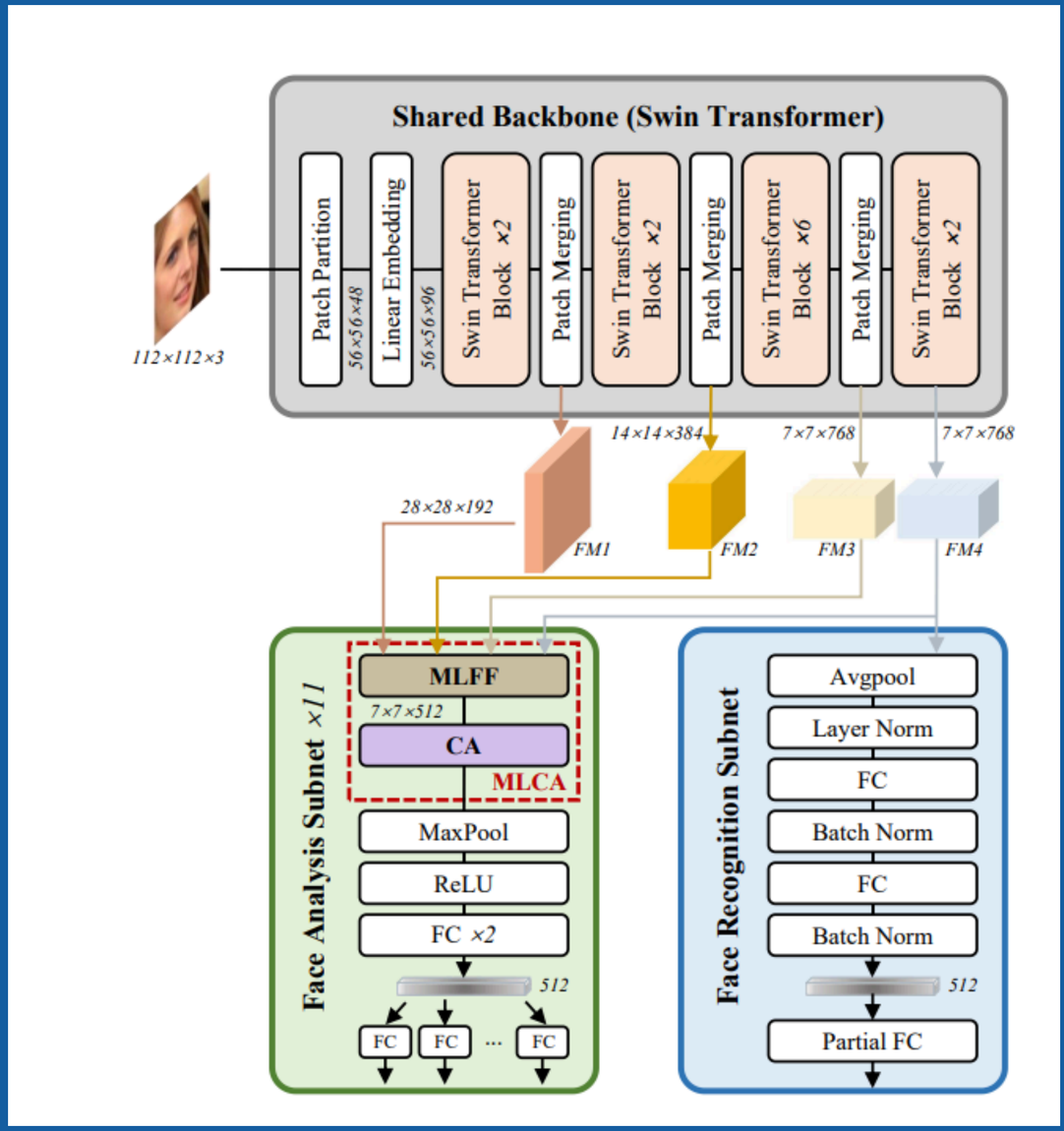


CHANNEL ATTENTION

Convolutional Block Attention Module (CBAM)



VISÃO GERAL


Shared Backbone (Swin Transformer)
Patch Partition
56×56×48
Linear Embedding
56×56×96
Swin Transformer Block ×2
Patch Merging
Swin Transformer Block ×2
Patch Merging
Swin Transformer Block ×6
Patch Merging
Swin Transformer Block ×2
112×112×3
28×28×192
14×14×384
7×7×768
7×7×768
FM1
FM2
FM3
FM4
Face Analysis Subnet ×11
MLFF
7×7×512
CA
MLCA
MaxPool
ReLU
FC ×2
512
FC
FC
...
FC
Face Recognition Subnet
Avgpool
Layer Norm
FC
Batch Norm
FC
Batch Norm
512
Partial FC

TREINAMENTO

PASSO-A-PASSO

- Pré-treino em reconhecimento facial do backbone e da subrede
- Congelamento do backbone
- Treinamento das sub-redes de análise com diversas bases

RESULTADOS

Reconhecimento Facial: Supera os outros modelos em praticamente todos os testes

Method	Params (M)	Verification Accuray					IJB-C TAR@FAR					
		LFW	CFP-FP	AgeDB-30	CALFW	CPLFW	1e-6	1e-5	1e-4	1e-3	1e-2	1e-1
ResNet-50 [3]	43.6	99.69	98.14	97.53	95.87	92.45	81.43	90.98	94.32	96.38	97.82	98.75
ViT [2]	63.2	99.83	96.19	97.82	95.92	92.55	-	-	95.96	97.28	98.22	98.99
V2T-ViT [67]	63.5	99.82	96.59	98.07	95.85	93.00	-	-	95.67	97.10	98.14	98.90
ViT-P10S8 [20]	63.5	99.77	96.43	97.83	95.95	92.93	-	-	96.06	97.45	98.23	98.96
ViT-P12S12 [20]	63.5	99.80	96.77	98.05	**96.18**	93.08	-	-	96.31	97.49	98.38	99.04
Swin-T [57]	28.5	99.80	97.91	97.85	95.98	92.60	88.54	93.71	95.75	97.13	98.01	98.86
SwinFace	28.5	**99.87**	**98.60**	**98.15**	96.10	**93.42**	**90.82**	**94.93**	**96.73**	**97.79**	**98.43**	**99.08**

RESULTADOS

Reconhecimento de Expressões: Supera os métodos do estado da arte utilizando um método mais simples, e com menos parâmetros.

Method	Accuray
DLP-CNN [9]	80.89
gACNN [26]	85.07
IPA2LT [25]	86.77
RAN [21]	86.90
CovPool [71]	87.00
SCN [22]	87.03
DACL [23]	87.78
KTN [24]	88.07
Zhang et al. [27]	89.01
AMP-Net [28]	89.25
TransFER [4]	90.91
SwinFace	**90.97**

RESULTADOS

Estimativa de Idade: Também supera os métodos do estado da arte utilizando um método mais simples, e com menos parâmetros.

Method	Validatation		Test	
	MAE	ϵ-error	MAE	ϵ-error
AIO [15]	-	0.29	-	-
AgeNet [72]	3.33	0.29	-	0.26
DEX [29]	3.25	0.28	-	0.26
AGEn [38]	3.21	0.28	2.94	0.26
AL-RoR [30]	3.14	0.27	-	0.25
BridgeNet [31]	2.98	0.26	2.87	0.26
MWR [39]	2.95	0.26	2.77	0.25
SwinFace	**2.50**	**0.20**	**2.47**	**0.22**

RESULTADOS

Estimativas de Atributos:
Performance comparável aos outros modelos

	5 o'clock Shadow	Arched Eyebrows	Attractive	Bags Under Eyes	Bald	Bangs	Big Lips	Big Nose	Black Hair	Blond Hair	Blurry	Brown Hair	Bushy Eyebrows	Chubby
PANDA-1 [40]	88.00	78.00	81.00	79.00	96.00	92.00	67.00	75.00	85.00	93.00	86.00	77.00	86.00	86.00
LNets+ANet [13]	91.00	79.00	81.00	79.00	98.00	95.00	68.00	78.00	88.00	95.00	84.00	80.00	90.00	91.00
MOON [41]	94.03	82.26	81.67	84.92	98.77	95.80	71.48	84.00	89.40	95.86	95.67	89.38	92.62	95.44
NSA [73]	93.13	82.56	82.76	84.86	98.03	95.71	69.28	83.81	89.03	95.76	95.96	88.25	92.66	94.94
MCNN-AUX [42]	94.51	83.42	83.06	84.92	98.90	96.05	71.47	84.53	89.78	96.01	96.17	89.15	92.84	95.67
MCFA [43]	94.00	83.00	83.00	85.00	99.00	96.00	72.00	84.00	89.00	96.00	96.00	88.00	92.00	96.00
DMM-CNN [44]	94.84	84.57	83.37	85.81	99.03	96.22	72.93	84.78	90.50	96.13	96.40	89.46	93.01	95.86
SwinFace	94.60	83.91	82.61	84.24	98.99	96.09	71.26	83.98	90.17	95.94	96.04	89.11	92.62	95.69

	Double Chin	Eyeglasses	Goatee	GrayHair	Heavy Makeup	High Cheekbones	Male	Mouth Slightly Open	Mustache	Narrow Eyes	No Beard	Oval Face	Pale Skin	Pointy Nose
PANDA-1 [40]	88.00	98.00	93.00	94.00	90.00	86.00	97.00	93.00	93.00	84.00	93.00	65.00	91.00	71.00
LNets+ANet [13]	92.00	99.00	95.00	97.00	90.00	88.00	98.00	92.00	95.00	81.00	95.00	66.00	91.00	72.00
MOON [41]	96.32	99.47	97.04	98.10	90.99	87.01	98.10	93.54	96.82	86.52	95.58	75.73	97.00	76.46
NSA [73]	95.80	99.51	96.68	97.45	91.59	87.61	97.95	93.78	95.86	86.88	96.17	74.93	97.00	76.47
MCNN-AUX [42]	96.32	99.63	97.24	98.20	91.55	87.58	98.17	93.74	96.88	87.23	96.05	75.84	97.05	77.47
MCFA [43]	96.00	100.00	97.00	98.00	92.00	87.00	98.00	93.00	97.00	87.00	96.00	75.00	97.00	77.00
DMM-CNN [44]	96.39	99.69	97.63	98.27	91.85	87.73	98.29	94.16	97.03	87.73	96.41	75.89	97.00	77.19
SwinFace	96.09	99.67	97.21	98.27	91.41	87.24	98.96	93.78	96.91	87.30	96.14	74.72	96.85	77.08

	Receding Hairline	Rosy Cheeks	Sideburns	Smiling	Straight Hair	Wavy Hair	Wearing Earrings	Wearing Hat	Wearing Lipstick	Wearing Necklace	Wearing Necktie	Young		**Average**
PANDA-1 [40]	85.00	87.00	93.00	92.00	69.00	77.00	78.00	96.00	93.00	67.00	91.00	84.00		85.43
LNets+ANet [13]	89.00	90.00	96.00	92.00	73.00	80.00	82.00	99.00	93.00	71.00	93.00	87.00		87.33
MOON [41]	93.56	94.82	97.59	92.60	82.26	82.47	89.60	98.95	93.93	87.04	96.63	88.08		90.94
NSA [73]	92.25	94.79	97.17	92.70	80.41	81.70	89.44	98.74	93.21	85.61	96.05	88.01		90.61
MCNN-AUX [42]	93.81	95.16	97.85	92.73	83.58	83.91	90.43	99.05	94.11	86.63	96.51	88.48		91.29
MCFA [43]	94.00	95.00	98.00	93.00	85.00	85.00	90.00	99.00	94.00	88.00	97.00	88.00		91.23
DMM-CNN [44]	94.12	95.32	97.91	93.22	84.72	86.01	90.78	99.12	94.49	88.03	97.15	88.98		**91.70**
SwinFace	93.92	94.96	97.75	93.18	84.73	85.57	89.87	99.19	94.07	86.72	96.97	89.05		91.32

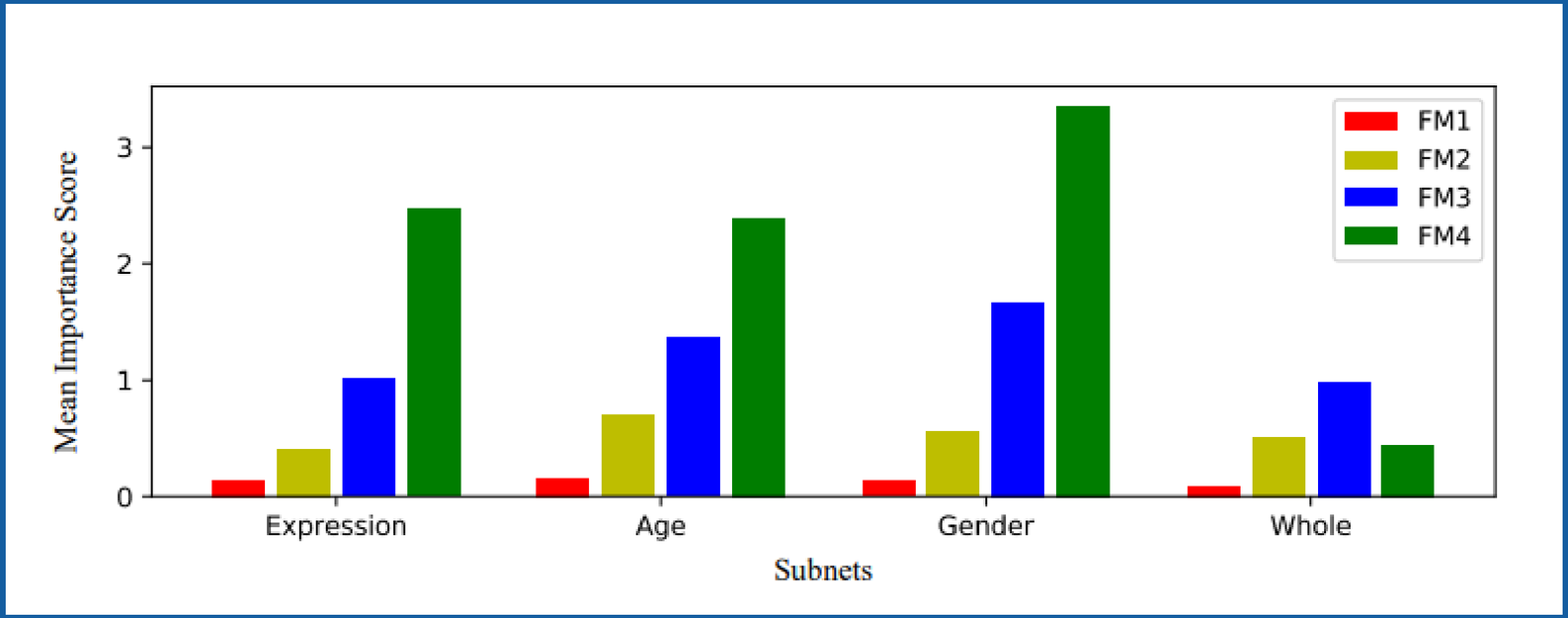
ANÁLISE DOS RESULTADOS

Comparação entre aprendizado de uma tarefa e multi-tarefa

Setting	Age ϵ-error on CLAP2015 [12] val	Smiling Acc. on CelebA [13]	Young Acc. on CelebA [13]
Single-task	0.357	92.40	88.38
Multi-task	**0.318**	**93.18**	**89.05**

ANÁLISE DOS RESULTADOS

Importância dos mapas de features em vários tipos de tarefas



CONCLUSÃO

- Os resultados obtidos demonstram o potencial do aprendizado multi-tarefas
- o MLCA é capaz de conciliar features diferentes e potencialmente conflitantes
- A possibilidade de aproveitar o backbone para resolver tarefas ajuda a contornar a falta de dados de algumas tarefas

Obrigado

Referência:

https://arxiv.org/pdf/2103.14803v2.pdf