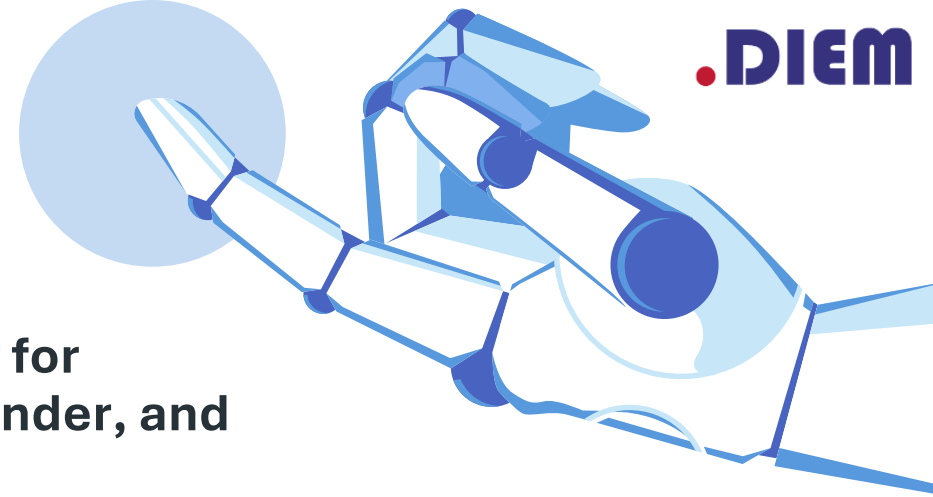


Adapting Vision Language Models via parameter-efficient fine-tuning for Multitask Classification of Age, Gender, and Emotion

Relatori:
Prof. Mario Vento
Prof. Antonio Greco

Canditato: *Antonio Sessa*
Matricola: *0622702311*

Soft Biometric Recogntion, What and Why?

Definition:

Soft biometrics are **non-unique** human attributes, that can be collected from images

Applications:

•**Social Robotics:** a robot estimates a user is a child and automatically switches to a simpler speech and more playful voice.

•**Marketing & Commerce:** A digital sign detects a shopper's likely age and gender to show a targeted ad, like for a video game or a new perfume.

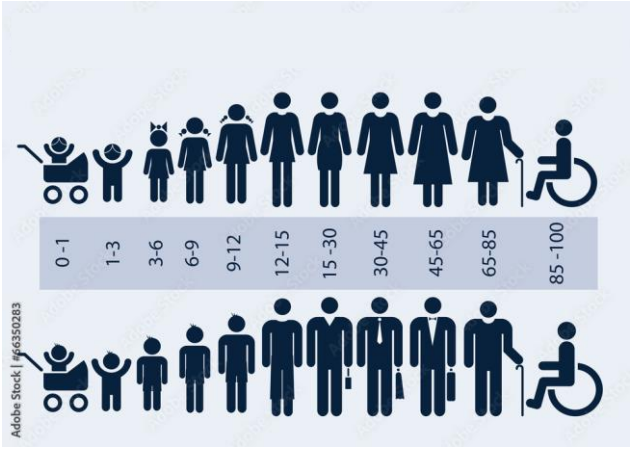
•**Security & Access:** A website uses facial age estimation to automatically block a user who appears underage from accessing mature content.

•**Healthcare & Wellness:** A wellness app monitors a user's vocal tone or facial expression through their phone to detect signs of stress or fatigue.

Challenges

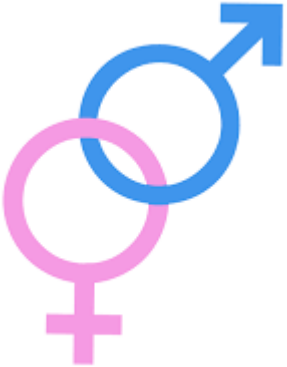
. Sadness, Surprise, Happiness, Disgust, Angry, Fear

. **Small dataset, class imbalances, low annotator agreements**



. 0-2,3-9,10-19,20-29,30-39,40-49,50-59,60-69,70+

. **High class intra variance, class imbalances**

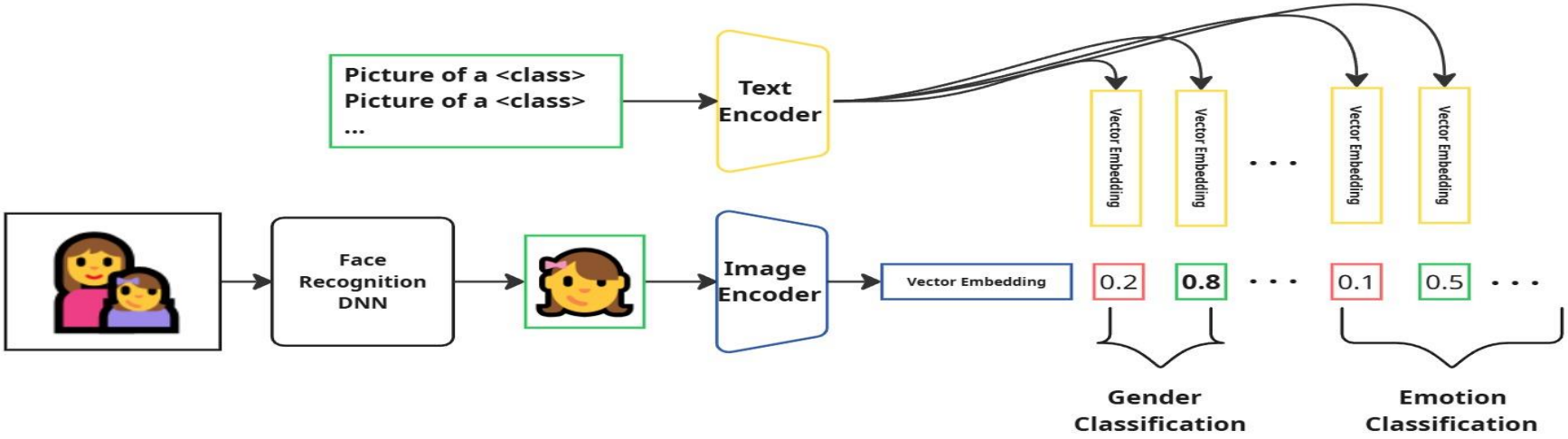


. Male and Female

. **Different age-groups and ethnicities**

Our Approach, let's use Vision Language Models

Vision language model (VLM) are large neural networks, trained on billion of image-text pairs, that can be use in a **zero-shot manner**.
We can create a soft-biometric recognition system using a VLM.



UTK-Age	UTK-Gender	FairFace-Age	FairFace-Gender	RAF-DB	VGG-Age	VGG-Gender
48.62%	96.63%	46.11%	97.60%	66.57%	42.01%	95.78%

Table 4.1: Baseline zero-shot accuracy results across testing datasets for age, gender, and emotion recognition tasks.

Age*	Gender	Emotion	Global
47.36%	96.67%	66.57%	69.61%

*Average calculation excludes the VggFace2 dataset as its age-labels data are synthetically obtained.

Table 4.2: Mean baseline accuracy for age, gender, and emotion recognition tasks.

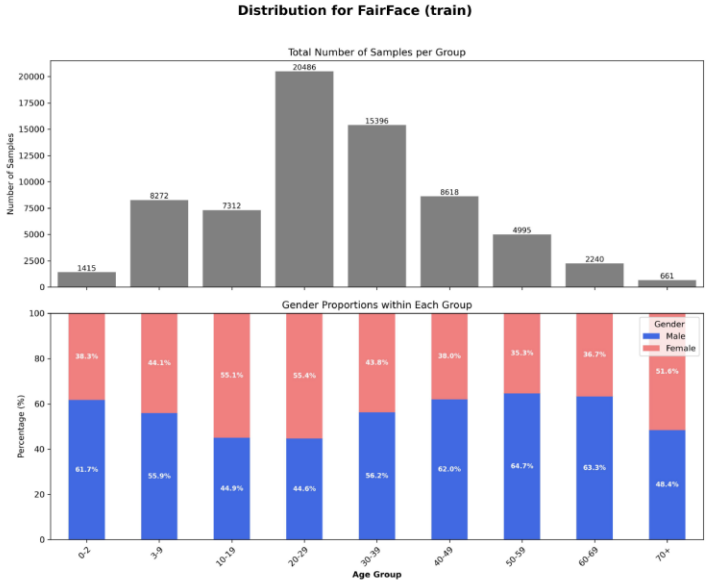
Component	Parameters
Text Encoder	353,986,561
Visual Encoder	318,212,106
Total Parameters	**671,137,793**
GFLOPs	**699.76**

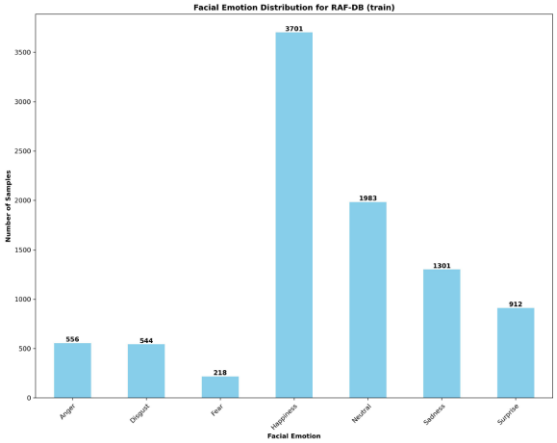
Table 4.3: Number of parameters used by the zero-shot baseline during inference

Age Group	Samples	%	Weighted %
0-2	2935	2.15%	6.63%
3-9	15 699	11.48%	11.31%
10-19	16 032	11.72%	11.36%
20-29	29 784	21.78%	16.37%
30-39	24 462	17.89%	14.56%
40-49	17 976	13.15%	12.14%
50-59	14 413	10.54%	10.89%
60-69	10 618	7.76%	9.44%
70+	4828	3.53%	7.31%

Gender	Samples	%	Weighted %
Male	82 147	48.79%	49.26%
Female	86 215	51.21%	50.74%

Emotion	Samples	%	Weighted %
Surprise	912	9.90%	12.14%
Fear	218	2.37%	8.44%
Disgust	544	5.90%	9.99%
Happy	3701	40.16%	26.85%
Sad	1301	14.12%	14.20%
Angry	556	6.03%	10.25%
Neutral	1983	21.52%	18.13%





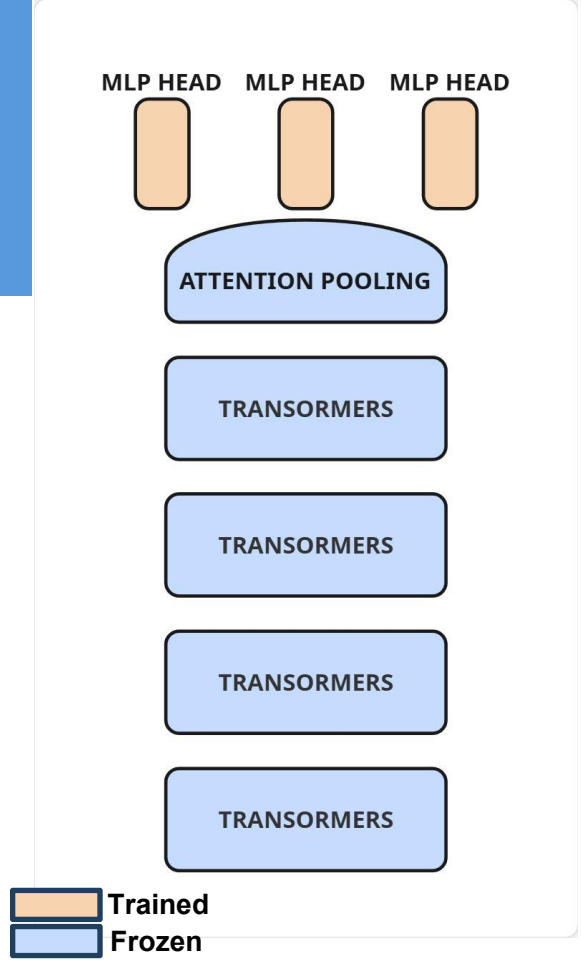
Task	Samples	%
Gender	168362	100.00%
Age	136747	81.22%
Emotion	9215	5.47%
Total	168362	–

Training Datasets:

- FairFace
- Lagenda
- RAF-DB
- CelebA-HQ

Testing Datasets:

- UTKFace
- VggFace2



Linear Probing:

Table 4.11: Summary of Model Average Performance (%)

Model	Emotion Acc	Avg. Age Acc	Avg. Gender Acc	**Overall Avg**
ZS	66.57	47.36	96.67	69.61
LP	84.91	61.28	97.54	78.76
FT_4	88.43 ↓	63.00 ↑	97.41 ↓	82.94
LoRA	91.21 ↑	63.53 ↑	97.49 ↓	**84.07**
MTLoRA	90.06 ↓	64.03 ↑	97.49 ↓	83.88

↓ Denotes a decrease compared to single-task equivalent.
↑ Denotes an increase compared to single-task equivalent.

Multi-task learning, challenges:

Loss Balancing:
Missing Labels:
Task Imbalance:

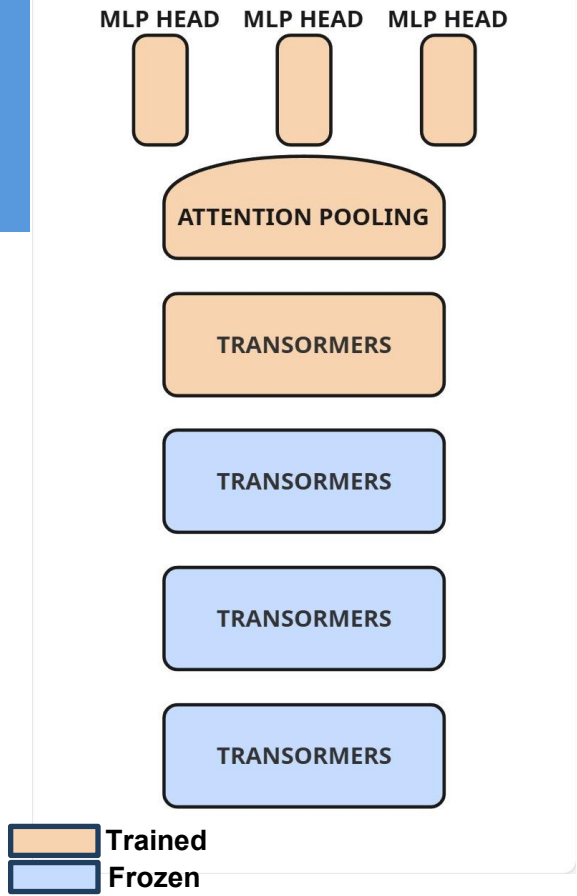


Table 4.11: Summary of Model Average Performance (%)

Model	Emotion Acc	Avg. Age Acc	Avg. Gender Acc	**Overall Avg**
ZS	66.57	47.36	96.67	69.61
LP	84.91	61.28	97.54	78.76
FT_4	88.43 ↓	63.00 ↑	97.41 ↓	82.94
LoRA	91.21 ↑	63.53 ↑	97.49 ↓	**84.07**
MTLoRA	90.06 ↓	64.03 ↑	97.49 ↓	83.88

↓ Denotes a decrease compared to single-task equivalent.
↑ Denotes an increase compared to single-task equivalent.

$$Softmax(f^{\theta}(x))$$

$$Softmax(\frac{1}{\sigma^2}f^{\theta}(x))$$

$$\mathcal{L}_{mtl}(W, \sigma_a, \sigma_g, \sigma_e) = \frac{1}{\sigma_a^2}\mathcal{L}_a(W) + \frac{1}{\sigma_g^2}\mathcal{L}_g(W) + \frac{1}{\sigma_e^2}\mathcal{L}_e(W) + \log\sigma_a + \log\sigma_g + \log\sigma_e$$

Parameter Efficient Fine-tune LoRA

Enhancaments for LoRA: LoRA+ & DoRA

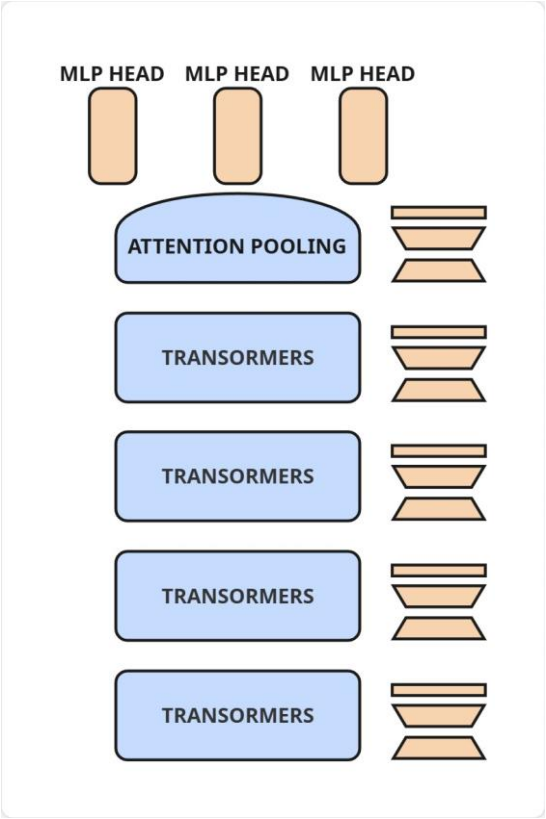


Table 4.11: Summary of Model Average Performance (%)

Model	Emotion Acc	Avg. Age Acc	Avg. Gender Acc	**Overall Avg**
ZS	66.57	47.36	96.67	69.61
LP	84.91	61.28	97.54	78.76
FT_4	88.43 ↓	63.00 ↑	97.41 ↓	82.94
LoRA	91.21 ↑	63.53 ↑	97.49 ↓	**84.07**
MTLoRA	90.06 ↓	64.03 ↑	97.49 ↓	83.88

↓ Denotes a decrease compared to single-task equivalent.
↑ Denotes an increase compared to single-task equivalent.

PEFT & Multi-task: MTLoRA

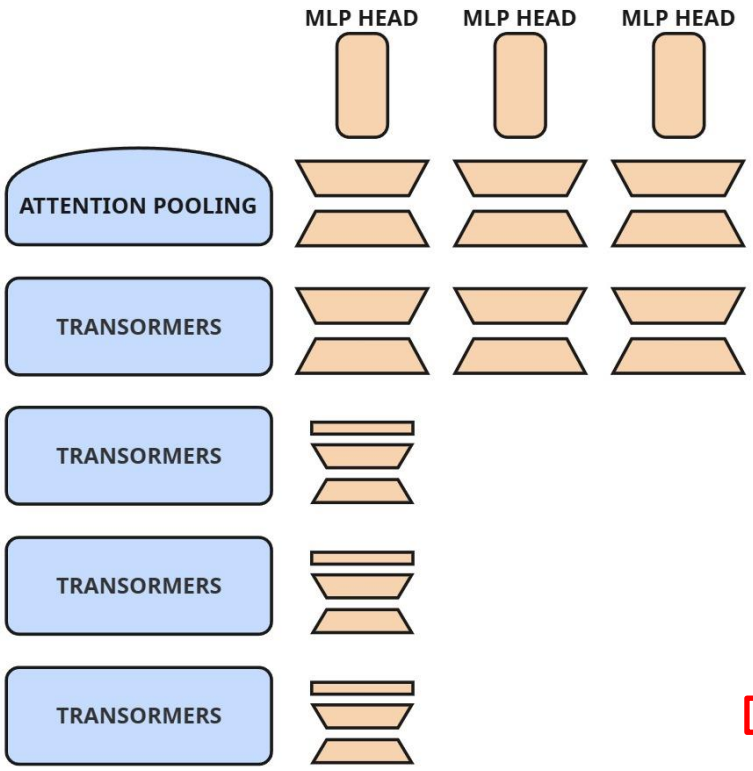


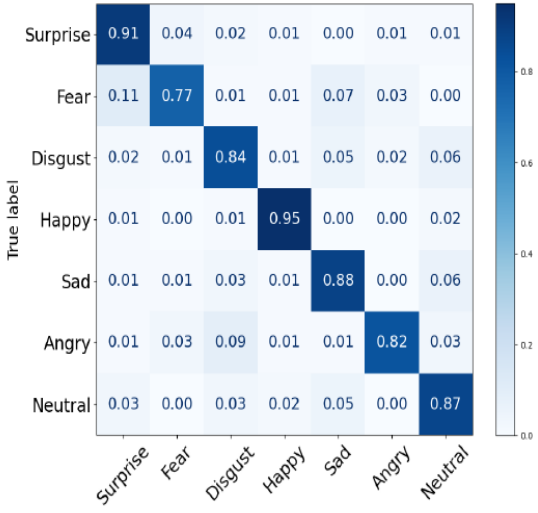
Table 4.11: Summary of Model Average Performance (%)

Model	Emotion Acc	Avg. Age Acc	Avg. Gender Acc	**Overall Avg**
ZS	66.57	47.36	96.67	69.61
LP	84.91	61.28	97.54	78.76
FT_4	88.43 ↓	63.00 ↑	97.41 ↓	82.94
LoRA	91.21 ↑	63.53 ↑	97.49 ↓	**84.07**
MTLoRA	90.06 ↓	64.03 ↑	97.49 ↓	83.88

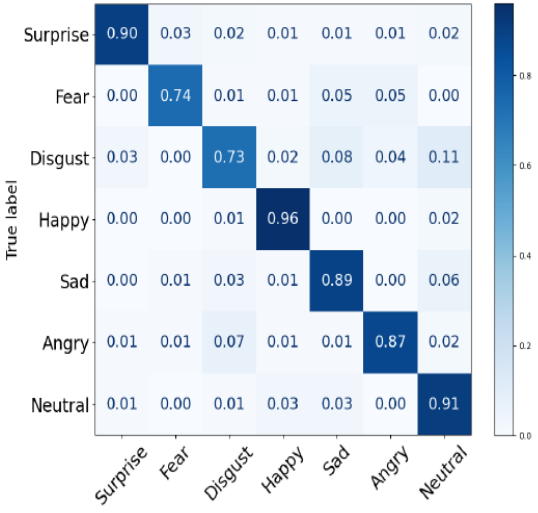
↓ Denotes a decrease compared to single-task equivalent.
↑ Denotes an increase compared to single-task equivalent.

Method	Age (Acc. %)		Gender (Acc. %)		Emotion (Acc. %)
	FairFace	UTKFace	FairFace	UTKFace	RAF-DB
SOTA (Age/Gender Focused)					
$MIVOLO_{224}$ [22]	61.07	3.7 MAE	95.73	**97.69**	-
$MIVOLO_{384}$ [23]	62.28	...	97.5	...	-
CLIP ViT-L/14 336px*[14]	63.45	...	97.1	...	-
SOTA (Emotion Focused)					
ResEmoteNet [18]	-	-	-	-	**94.76**
APViT [19]	-	-	-	-	92.21
POSTER++ [21]	-	-	-	-	91.98
Our Models					
MTLoRA	**64.11**	63.96	**97.62**	96.93	90.06
LoRA	63.73	63.34	97.57	96.90	91.21

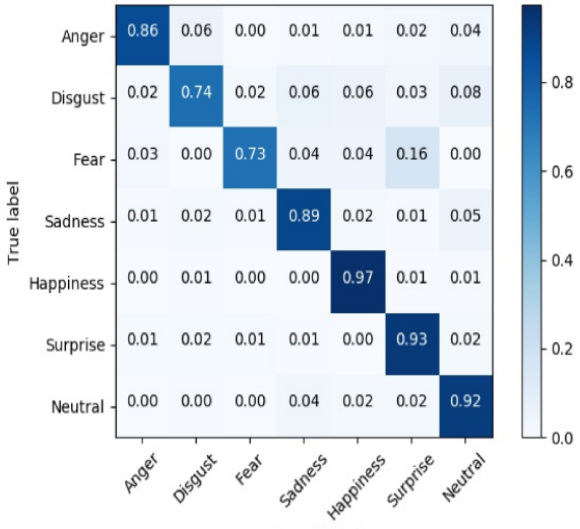
Table 4.12: Performance comparison with state-of-the-art (SOTA) methods for age, gender, and emotion recognition on standard benchmarks.



(a) MTLoRA confusion matrix on RAF-DB, balanced accuracy of 86.17 (Acc. 90.06)



(b) LoRA confusion matrix on RAF-DB, balanced accuracy of 85.90 (Acc. 91.21)



(c) APViT confusion matrix on RAF-DB, balanced accuracy of 86.36 (Acc. 92.21)