

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

Relatori:

*Prof. Mario Vento*

*Prof. Antonio Greco*

Candidato: *Antonio Sessa*

Matricola: *0622702311*



# Index

01      **Introduction**

02      Methodology

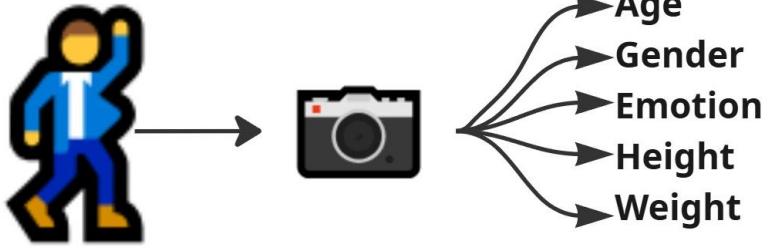
03      Results



# Soft Biometric Recognition, what and why?

## Definition

Soft biometrics are non-unique human attributes that can be indirectly 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

# Our domain, facial attributes

Facial Emotion  
Recognition



**Labels (7 classes):**

Happy, Surprise, Disgust,  
Angry, Fear, Sad, Neutral

**Challanges:**



Class imbalances

Small datasets

Low annotator agreements

Age group  
Classification



**Labels (9 classes):**

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

**Challanges:**



Class imbalances

High class intra-variance

Gender  
Recognition



**Labels (2 classes):**

Male and Female

**Challanges:**

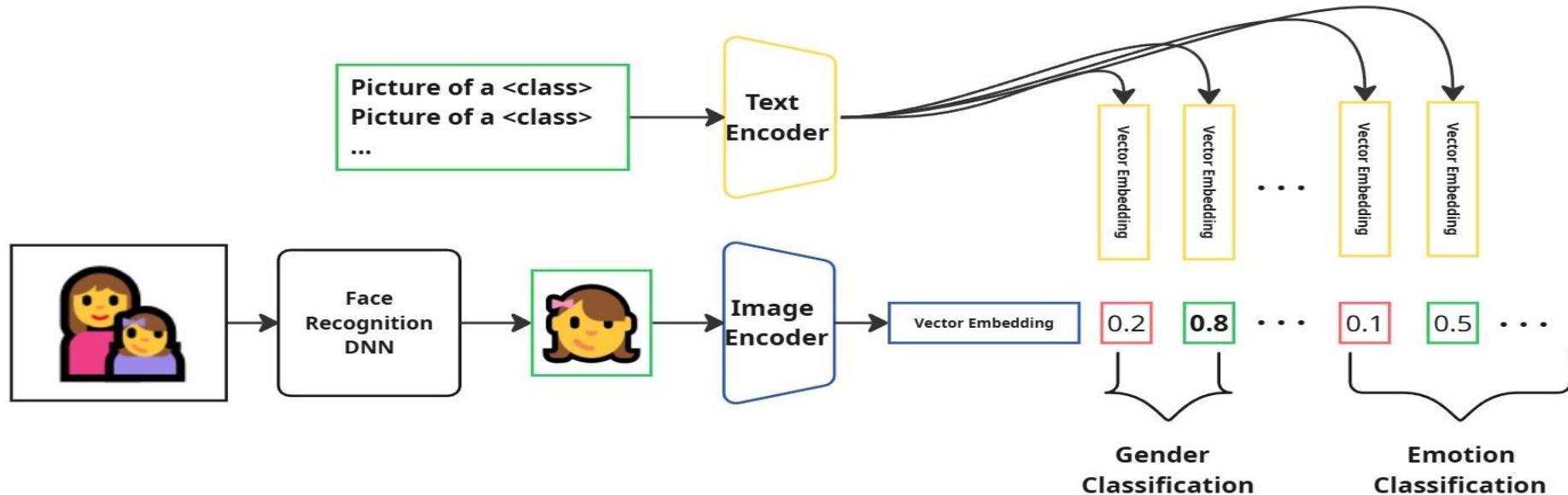


Different age groups  
and ethnicities

# Our Approach, Vision Language Models

Vision language model (VLM) are large neural networks, trained on billion of image-text pairs, that can be used in a **zero-shot manner**.

We can create a soft-biometric recognition system using a **VLM** (we choose **Perception Encoders**).



# Hard-Prompting, not good enough:

## Problems of hard-prompting

👎 Poor accuracy

👎 High memory footprint

👎 High latency

Since **visual understanding stems from the vision encoder**, we can **omit the text encoder** and use the image encoder as a foundation vision model, doing so we **halve the inference time and memory footprint**.

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.

Component	Parameters
Text Encoder	353,986,561
Visual Encoder	318,212,106
<b>Total Parameters</b>	<b>671,137,793</b>
<b>GFLOPs</b>	<b>699.76</b>

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

**01** Introduction

**02** Methodology

**03** Results



# Datasets used to adapt the pre-trained ViT

## Training Set

FairFace	~ 97k	Gender & Age	
Lagenda	~ 67k	Gender & Age	
RAF-DB	~ 17k	Emotion & Gender	
CelebaHQ	~ 17k	Gender	

## Test Set

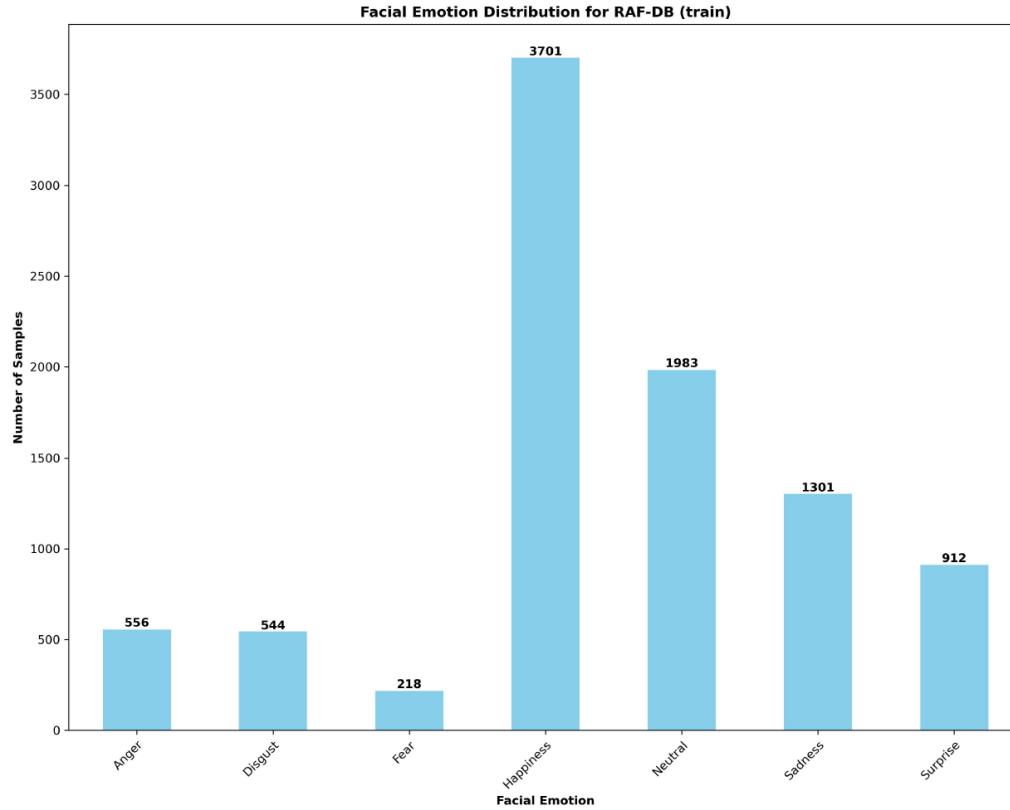
VggFace2	~ 170k	Gender & Age	
UTKFace	~ 24k	Gender & Age	
FairFace & RAF-DB; Test-splits			

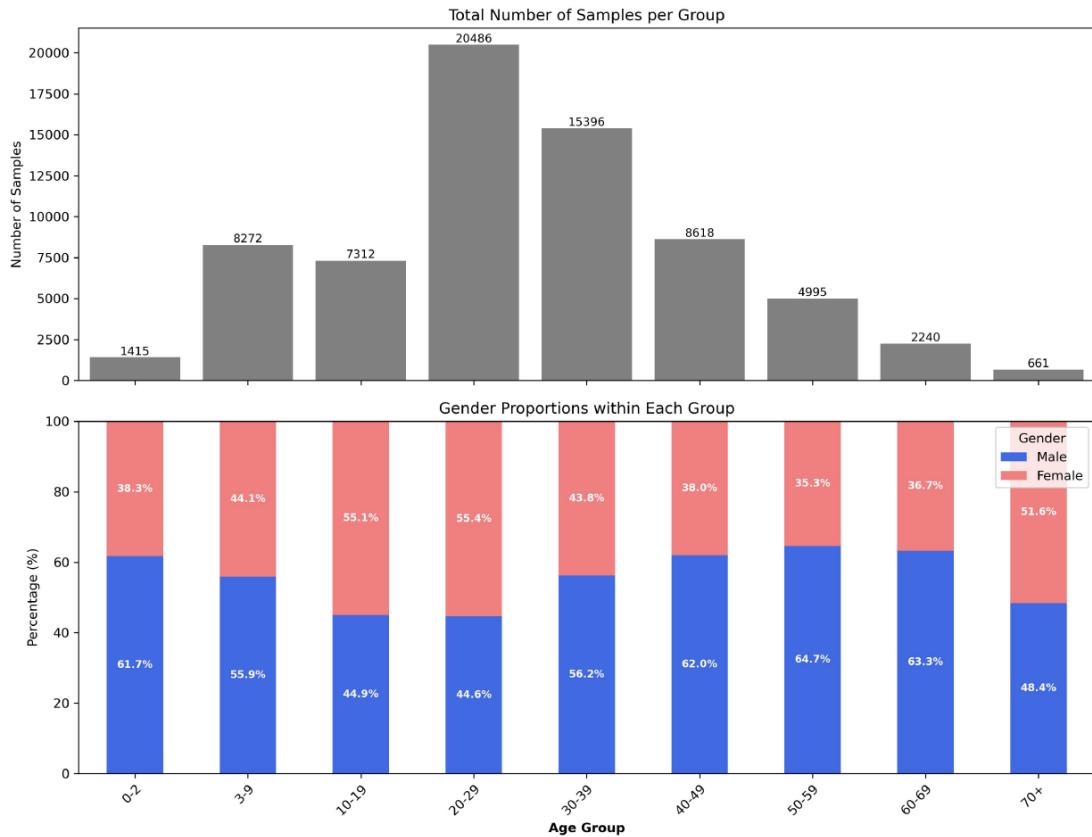


- . Low amount of celebrity data
- . Diverse ethnicities represented
- . High annotator agreements
- . Cross-dataset generalization

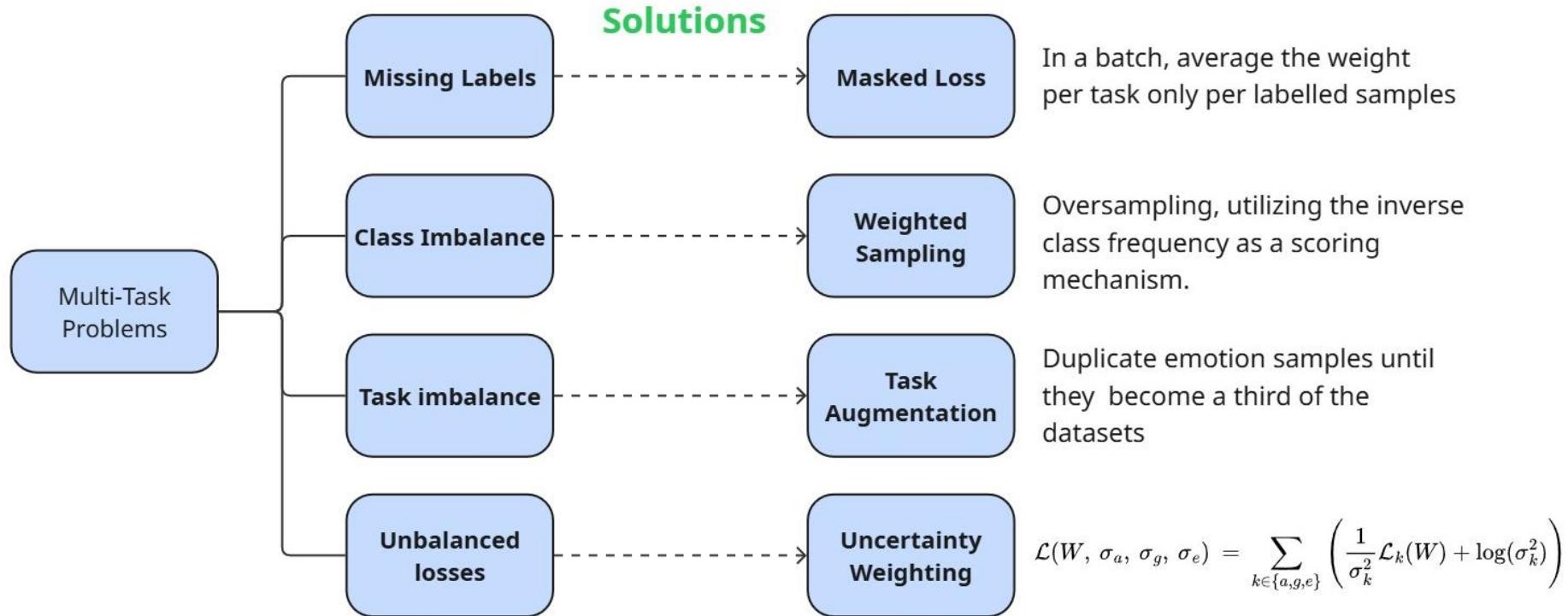


- . Intra task-class imbalance
- . Task imbalance
- . Missing labels

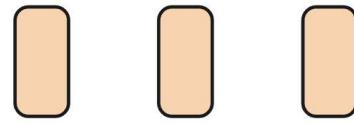




# Multi-task learning, problems and solutions



MLP HEAD    MLP HEAD    MLP HEAD



ATTENTION POOLING

TRANSFORMERS

TRANSFORMERS

TRANSFORMERS

TRANSFORMERS

Trained  
Frozen

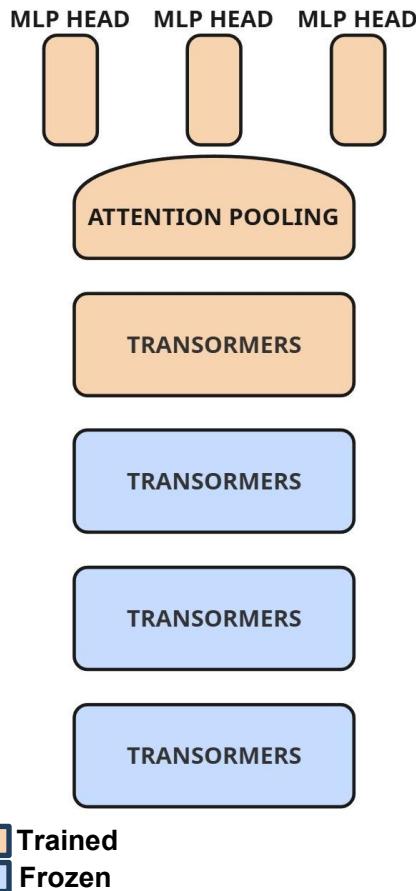


UNIVERSITÀ DEGLI STUDI DI SALERNO

**Linear Probing:** attaching a MLP classifier per task to the output of the model, to harness the out-of-the box features of the vision transformers. As there can be no benefit to train the classification head simultaneously, as there is no parameter sharing, they are **trained sequentially** on fully labelled datasets split.

**Training: 0.33% of parameters**





**Partial fine-tune:** unfreezing the attention pooling layer and last four transformers block of the ViT, using a differential LR strategy (using 1/10 of the LR for transformers blocks)

**Training: 20.13% of parameters**

# Can we go deeper with **full fine-tune**?

## Not really: Not enough VRAM

- ? Use gradient checkpointing and mixed precision? Still not enough.
- ? Use a smaller batch size? Noisy gradients, accentuated by the multi-task setting.

## Do we want to go deeper? **Yes\***

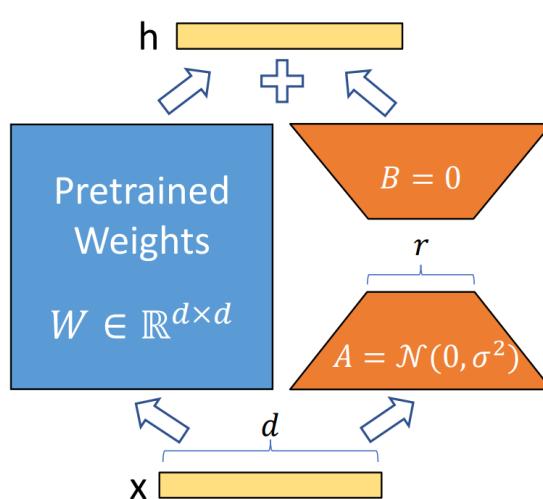
Adjusting the early-layer feature representations within the vision encoder may offer significant benefits.

**Increase capacity** with relatively **small dataset**, possible **overfitting**.

- \* Full fine-tuning could allow task-specific gradients to destroy the network powerful, generalized knowledge, “**Catastrophic forgetting**”

# Parameter Efficient Fine-tune with Low Rank Adaptation

**Main Idea:** the weight updates for large pre-trained model can be effectively represented in a low-dimensional subspace.



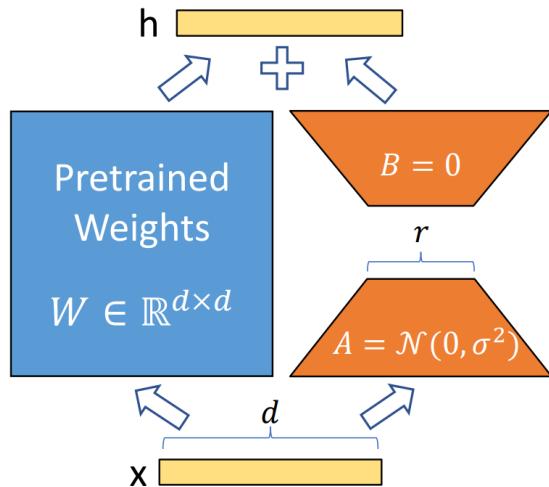
$$h = Wx + BAx$$

## Benefits of LoRA:

- Small checkpoints
- **No added inference latency and memory footprint**
- **Lower VRAM consumption** due to parameter-efficient updates
- **Prevents «catastrophic forgetting»** by limiting weight update in a low-rank space

# Enhancements for LoRA: LoRA+ & DoRA

**Main Idea of LoRA+:** we can achieve a better LoRA fine-tune by employing a differential learning rate strategy for the LoRA matrices.

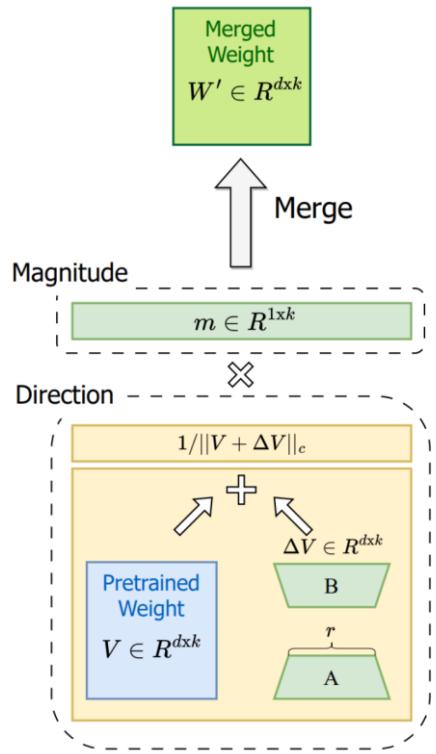


**Training update for LoRA+:**

$$B^t = B^{t-1} - \lambda \eta G_B$$
$$A^t = A^{t-1} - \eta G_A$$

In our implementation, we set  $\lambda = 6$

# Enhancements for LoRA: LoRA+ & DoRA

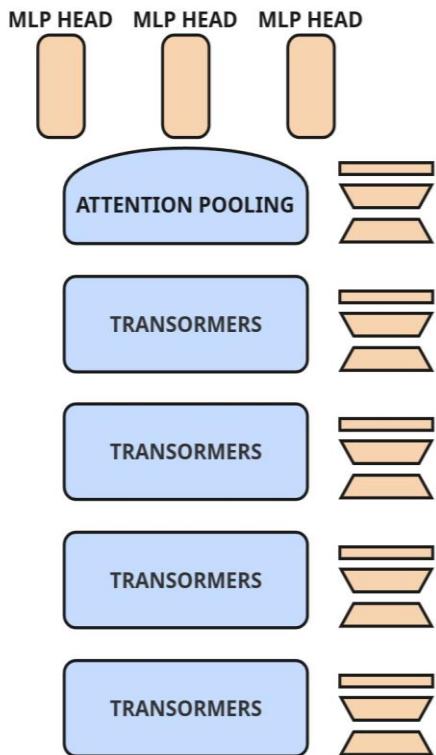


**Main Idea:** enhance the weight update by separately training the direction and magnitude of the update.

$$h = x \left( m \cdot \frac{W + BA}{\|W + BA\|_C} \right)$$

## Benefits of DoRA:

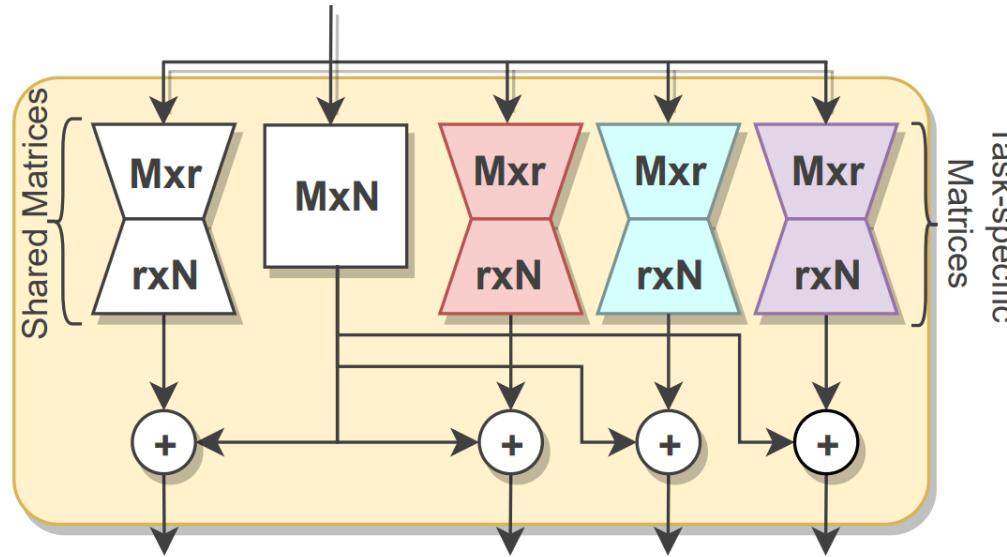
- Same benefits of LoRA
- Increased learning capacity



**DoRA & LoRA+ :** adding DoRA adapters with **rank=64** to each linear layer of the Vision Transformer and using a learning rate x6 to train the B matrices of the adapters.  
**Training: 8.47% of parameters (less than half of FT4)**



# PEFT & Multi-task: MTLoRA



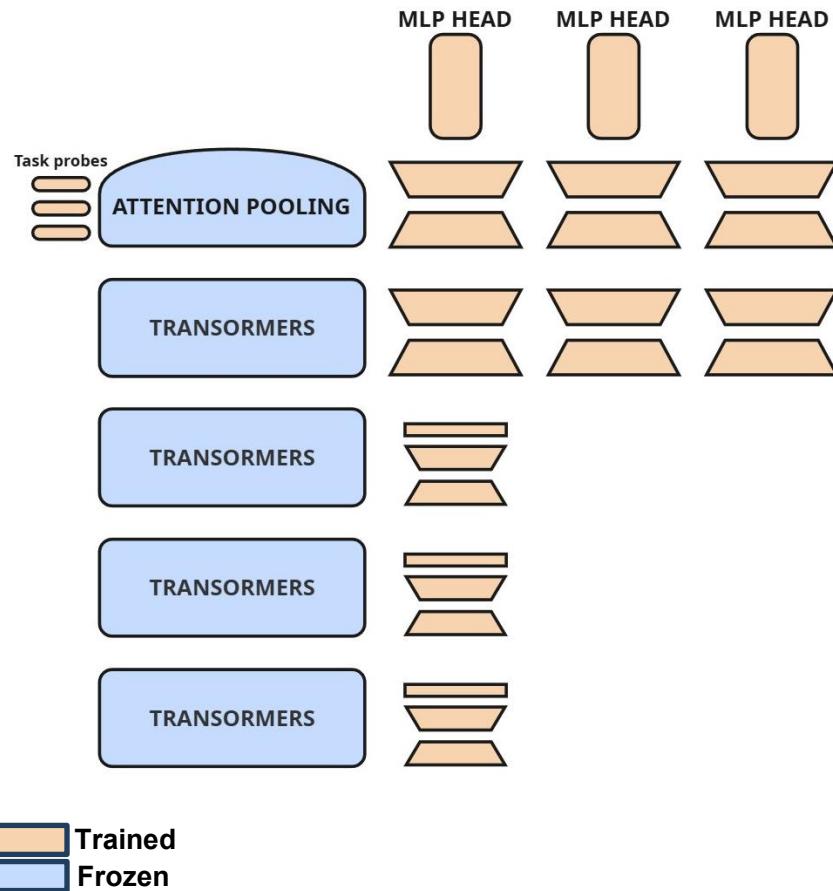
**Main Idea:** disentangle the parameter space through **Task-Specific LoRA matrices**

### Benefit:

- Train specialized parameters for a task, and use them to create **task-specifics feature-maps**

### Deficit:

- The task specific LoRA matrices cannot be merged, **small increase in memory footprint and inference latency**



**MTLoRA:** TS-LoRA ( $r=64$ ) applied to last transformer block and attention pooling layers. TA-DoRA using the same setup described earlier.  
**Training: 10.38% of parameters**



**01** Introduction

**02** Methodology

**03** Results



# Multi-task Vs. Single-task

Model	Avg. Age	Avg. Gender	Emotion	Global Avg	
Baseline	47.36%	96.67%	66.57%	69.59%	
LP	60.10%	97.51%	84.82%	80.81%	
FT_4 (ST)	62.80%	97.57%	88.78%	83.05%	-0.11%
FT_4 (MTL)	63.00%	97.41%	88.43%	82.94%	+0.03%
LoRA (ST)	63.72%	97.56%	90.83%	84.04%	
LoRA (MTL)	63.53%	97.49%	91.21%	<u>84.07%</u>	
MTLoRA	64.03%	97.51%	90.06%	83.86%	-0.18%

**Multi-task approach was successful in preventing negative-transfer**, all of our multi-task model achieve **perfomance parity compared to their single-task equivalent**

# Efficiency Comparison

Model	GLOPS	PARAMS	AVG. ACC
Baseline	700	671 M	69.61%
LoRA	352 (-50%)	320 M (-52%)	84.07% (+20%)
MTLoRA	368 (-47%)	329 M (-50%)	83.89% (+20%)



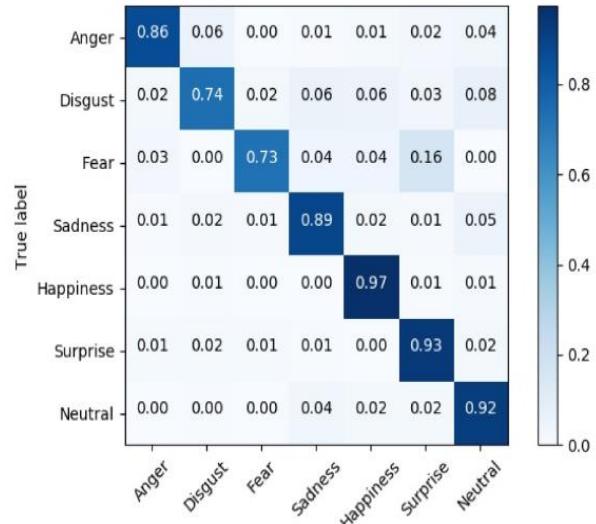
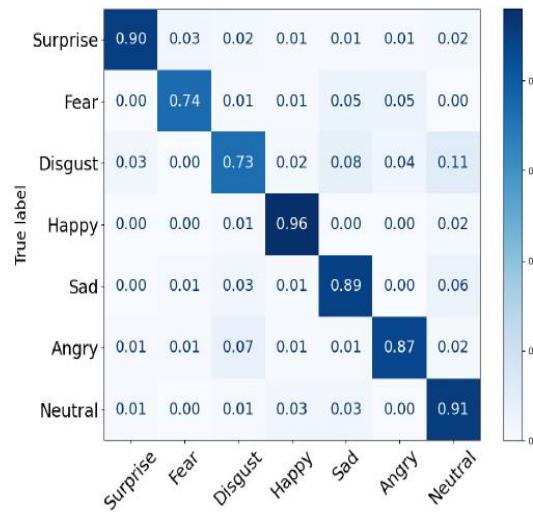
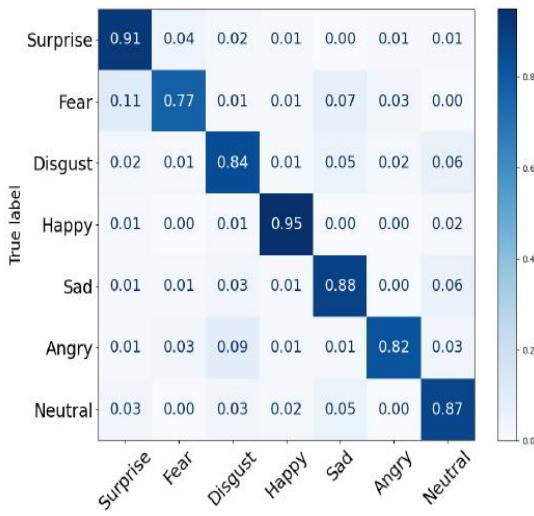
# Multi-task model detailed results

Model	FairFace (Age)	FairFace (Gender)	RAF-DB (Emotion)	UTKFace (Age)	UTKFace (Gender)	VggFace2 (Age)	VggFace2 (Gender)
Baseline	46.11%	97.60%	66.57%	48.43%	96.63%	42.01%	95.78%
LP	61.00%	97.70%	84.82%	61.56%	<b>97.00%</b>	57.75%	97.82%
FT_4 (MTL)	63.45%	<b>97.71%</b>	88.42%	62.54%	96.68%	61.68%	97.81%
LoRA (MTL)	63.73%	97.57%	<b>91.21%</b>	63.34%	96.90%	61.69%	<b>98.00%</b>
MTLoRA	<b>64.11%</b>	97.62%	90.06%	<b>63.96%</b>	96.93%	<b>62.74%</b>	<b>98.00%</b>

# Comparison with the state of the art

Model	FairFace (Age)	FairFace (Gender)	RAF-DB (Emotion)
MIVOLO	62.28%	97.50%	-
CLIP-ViT-L/14	63.45%	97.10%	-
ResEmoteNet	-	-	<b>94.76%</b>
ApViT	-	-	92.21%
Baseline	46.11%	97.60%	66.57%
LoRA	63.73%	97.57%	91.21%
MTLoRA	<b>64.11%</b>	<b>97.62%</b>	90.06%





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

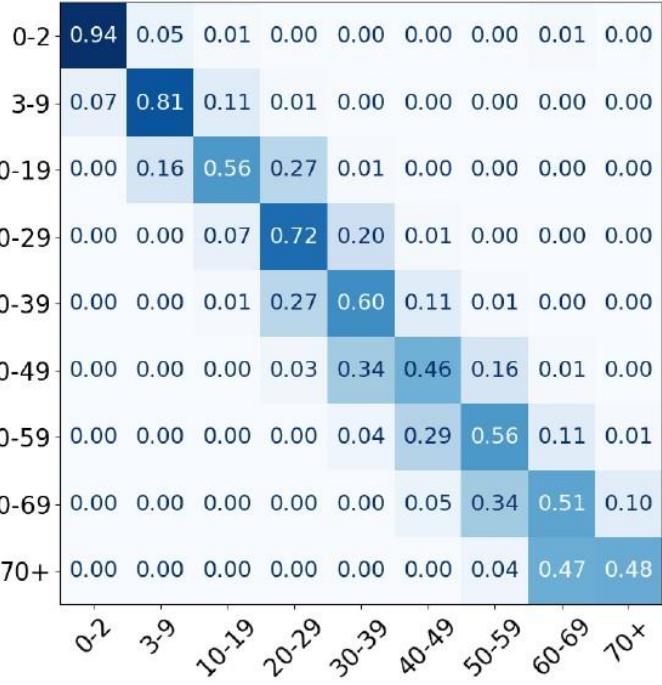
**-0.19%**

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

**-0.46%**

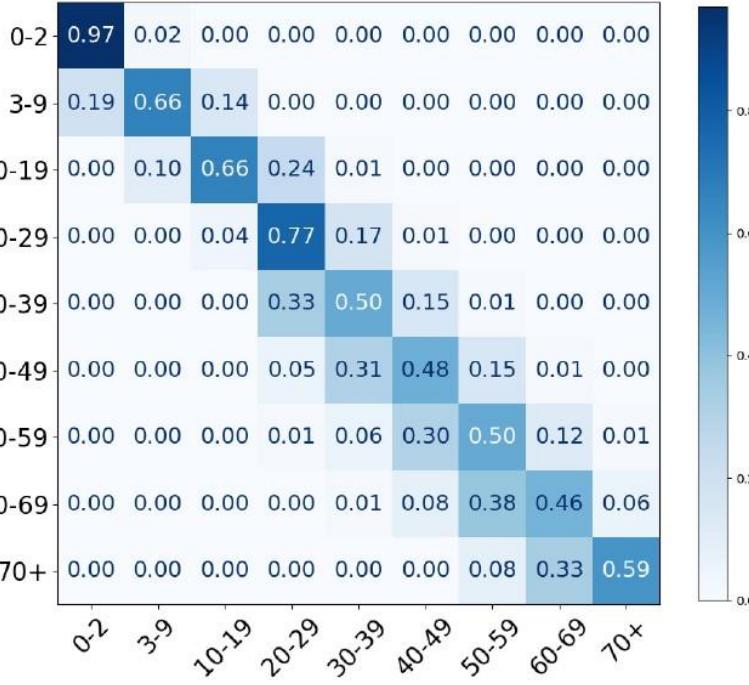
(c) APViT confusion matrix on RAF-DB, balanced accuracy of 86.36 (Acc. 92.21)

True label

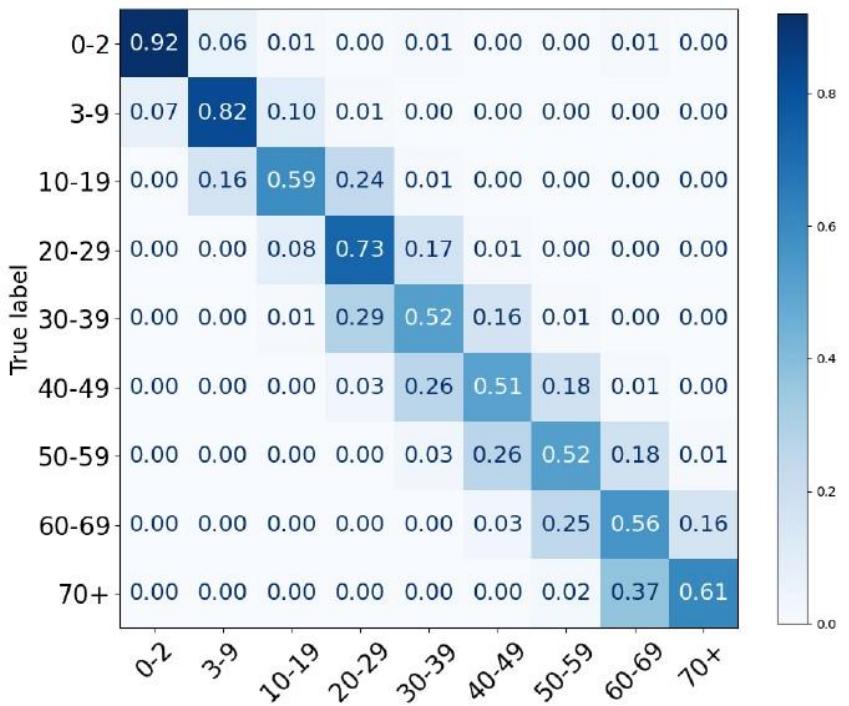


(a) MTLoRA confusion matrix on FairFace, balanced accuracy of **62.78** (Acc. 64.11)

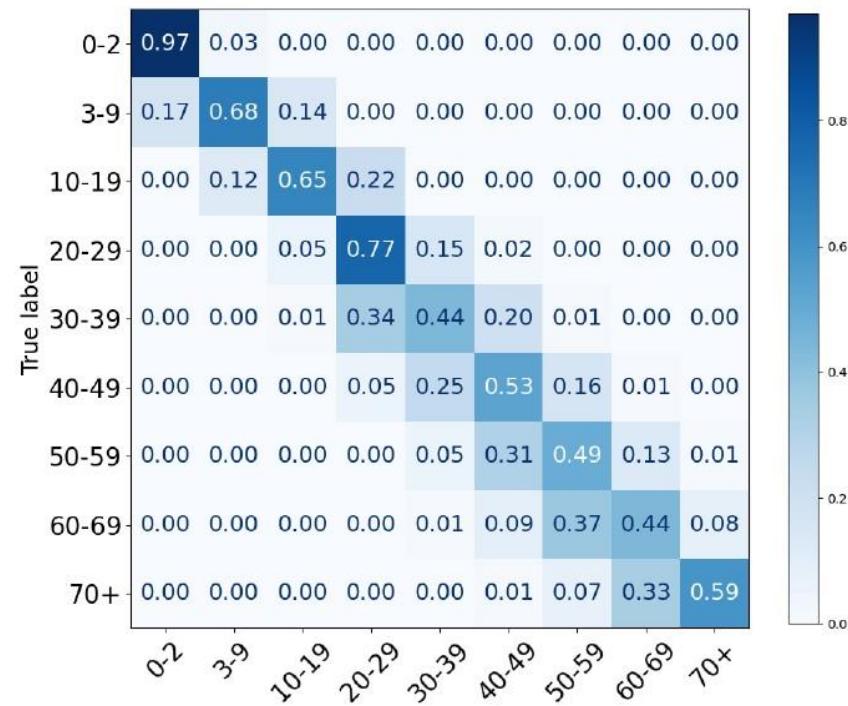
True label



(b) MTLoRA confusion matrix on UTKFace, balanced accuracy of **61.98** (Acc. 63.96)

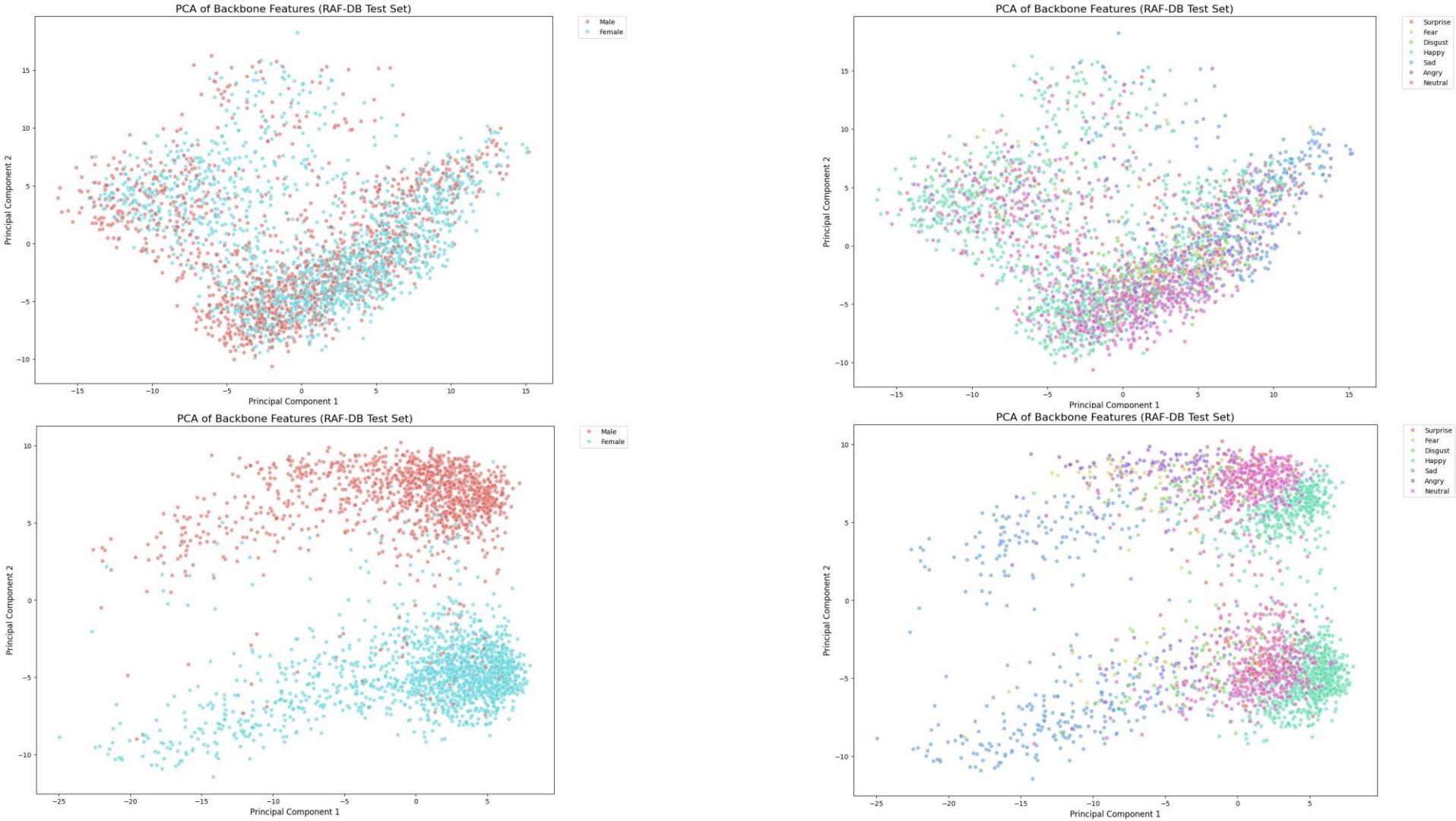


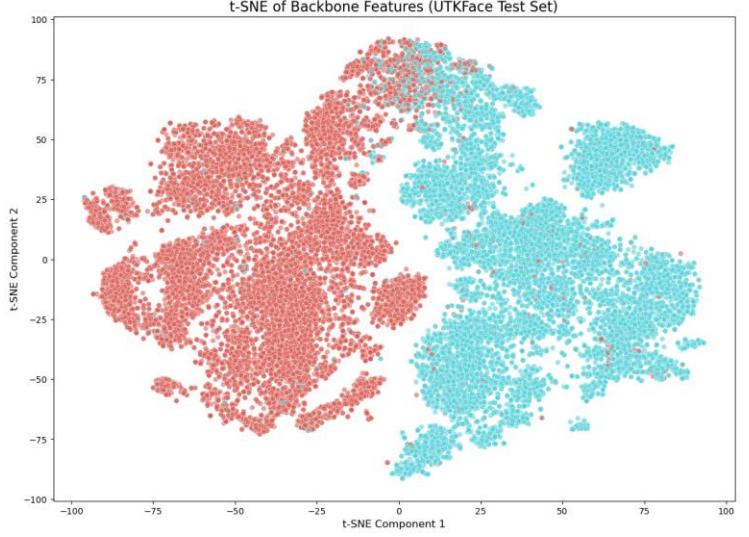
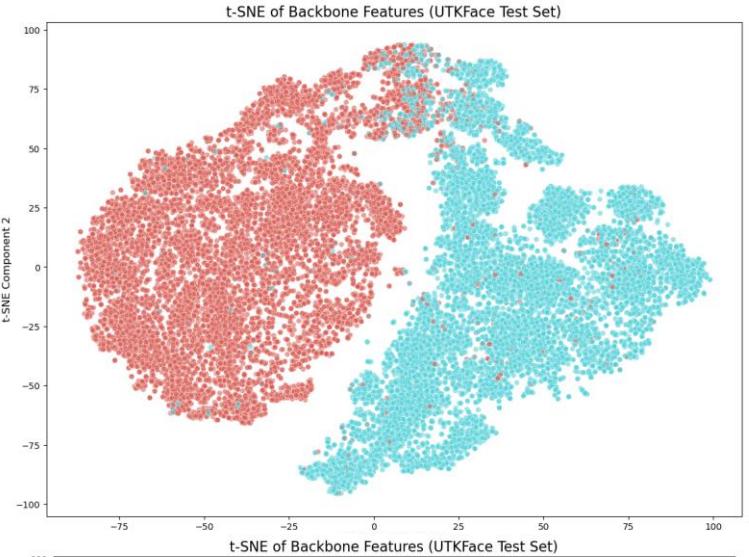
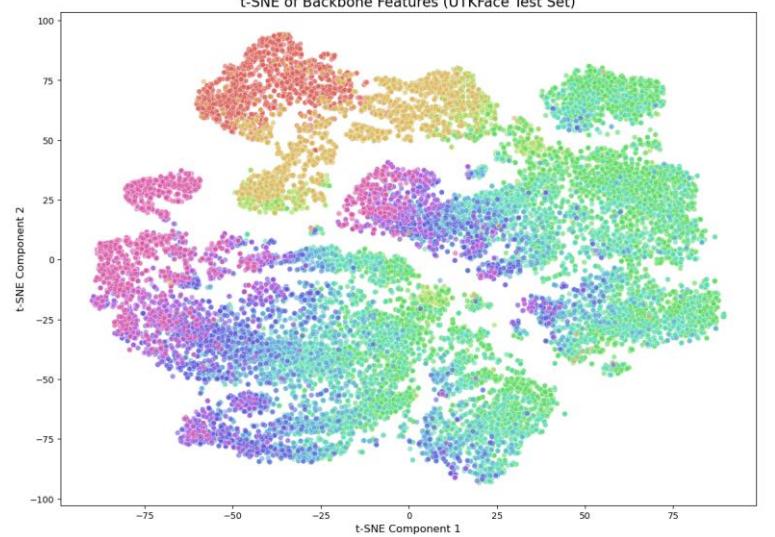
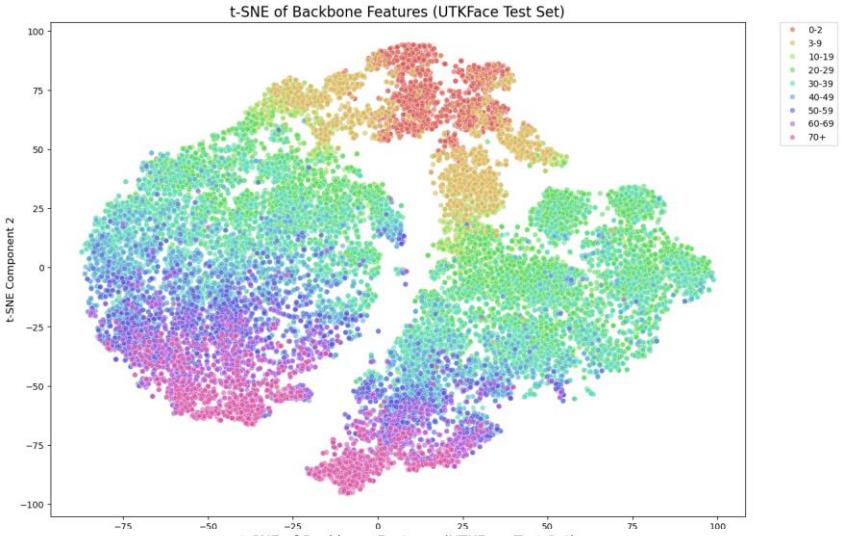
(c) LoRA confusion matrix on FairFace, balanced accuracy of **64.26** (Acc. 63.73)



(d) LoRA confusion matrix on UTKFace, balanced accuracy of **61.91** (Acc. 63.34)







# Dimostratore



<https://huggingface.co/spaces/Antuke/FaR-FT-PE>



UNIVERSITÀ DEGLI STUDI DI SALERNO

