

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



Datasets used to adapt the pre-trained ViT:

Training Datasets:

- **FairFace**, ~97k samples
age and gender labelled
- **Lagenda**, ~67k samples
age and gender labelled
- **RAF-DB**, ~17k samples
emotion and gender
labelled
- **CelebaHQ**, ~17k
samples gender labelled



- . Low amount of celebrity data in training set
- . Diverse ethnicities represented
- . High annotator agreements for the labels
- . Cross-dataset generalization with UTKFace and VggFace2

Testing Datasets:

- **FairFace (test-split)**
- **RAF-DB (test-split)**
- **VggFace2**, ~170k samples (synthetic) age and gender labelled
- **UTKFace**, ~24k samples age and gender labelled



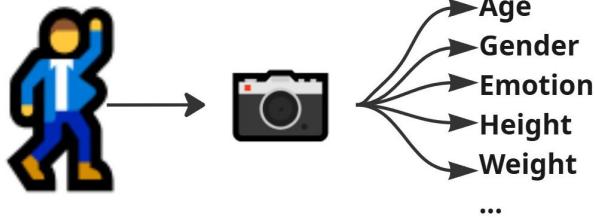
- . **Intra task-class imbalance** for age and emotion
- . **Task imbalance**, emotion heavily under-represented (5% of total)
- . **Missing labels**, no samples are labelled for all tasks

Soft Biometric Recognition, What and Why?

Definition:

Soft biometrics are non-unique human attributes that can be indirectly collected from images.

Unlike traditional biometrics such as fingerprints or iris patterns, soft biometric traits do not uniquely identify an individual but provide rich contextual information



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

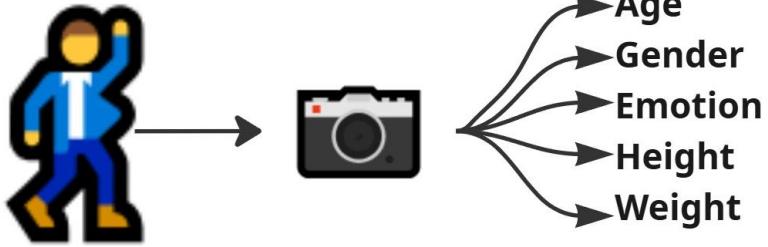


Healthcare & Wellness: A wellness app monitors a user's vocal tone or facial expression through their phone to detect signs of stress or fatigue

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, 10-29, 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:

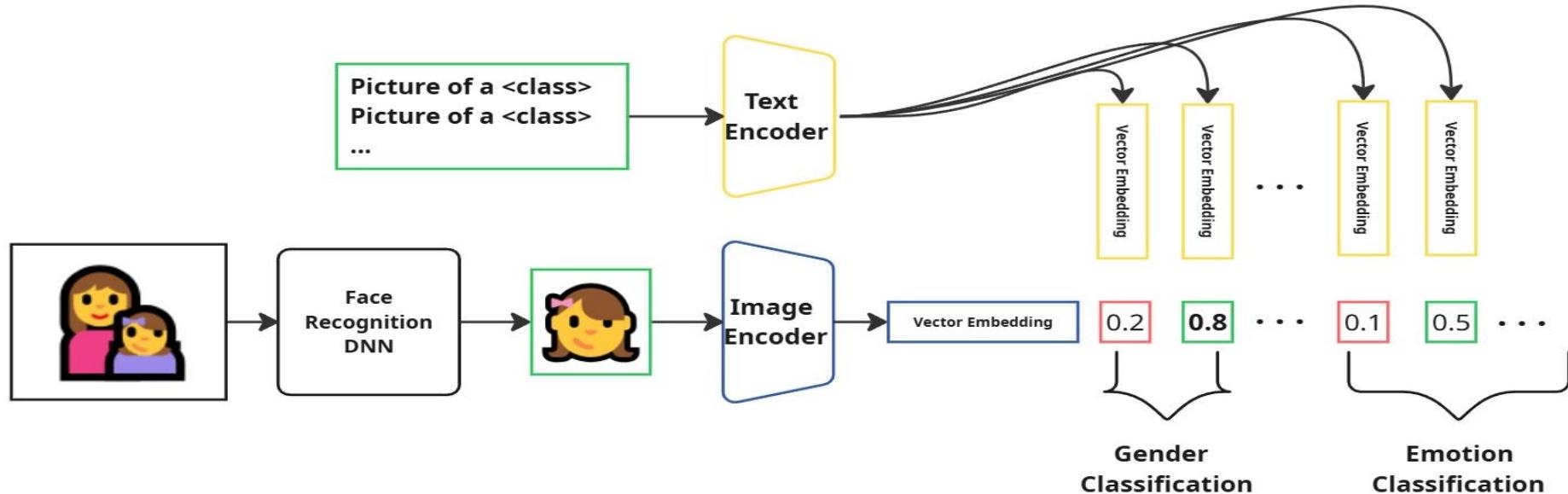


Different age groups
and ethnicities

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 (like the **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
Total Parameters	671,137,793
GFLOPs	699.76

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

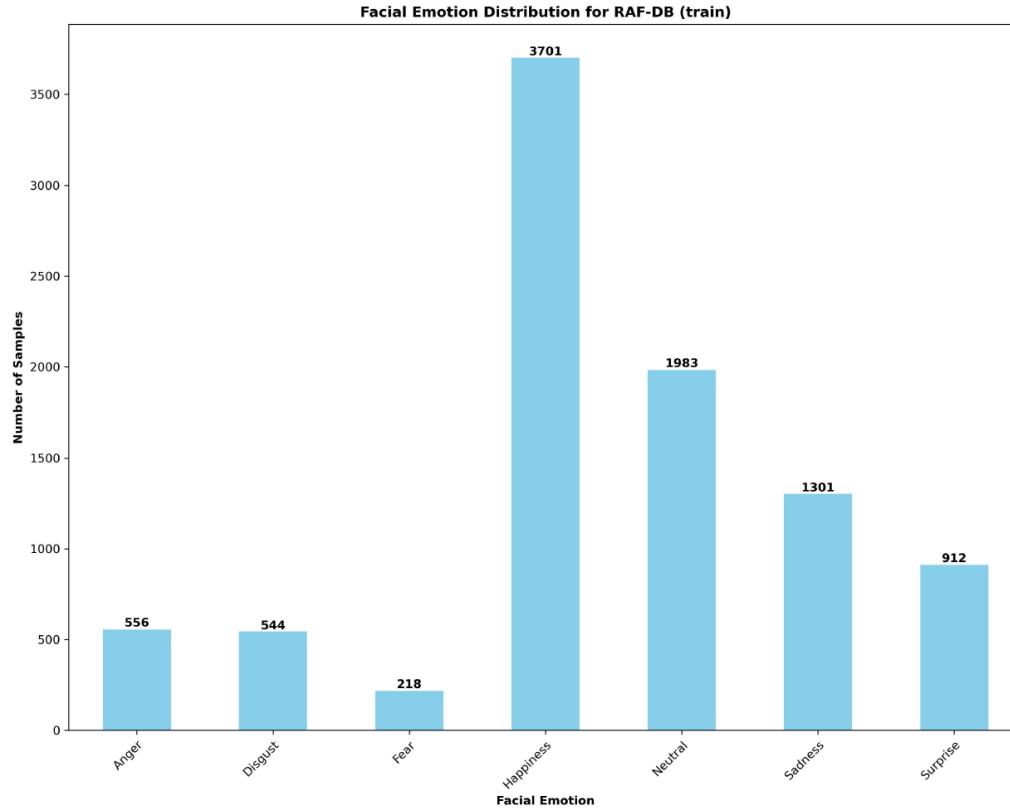
VggFace2	~ 170k	Gender & Age	
UTKFace	~ 24k	Gender & Age	
FairFace & RAF-DB; Test-splits			

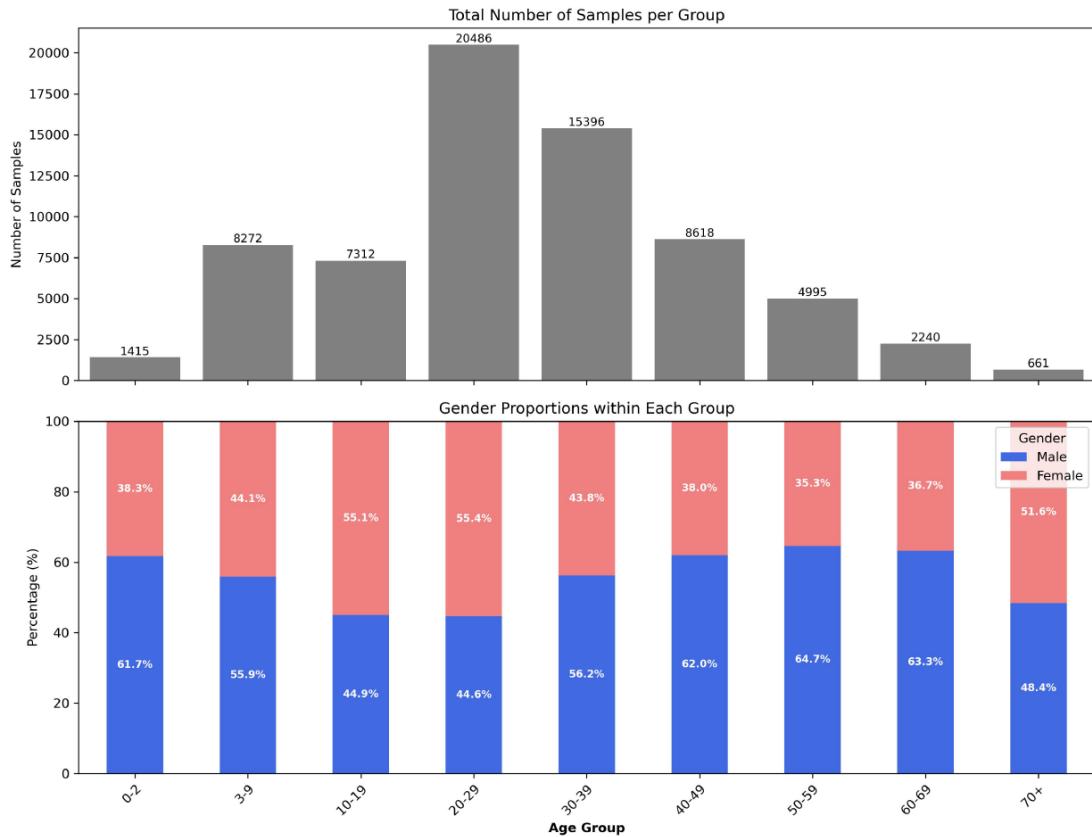


- . Low amount of celebrity data
- . Diverse ethnicities represented
- . High annotator agreements
- . Cross-dataset generalization

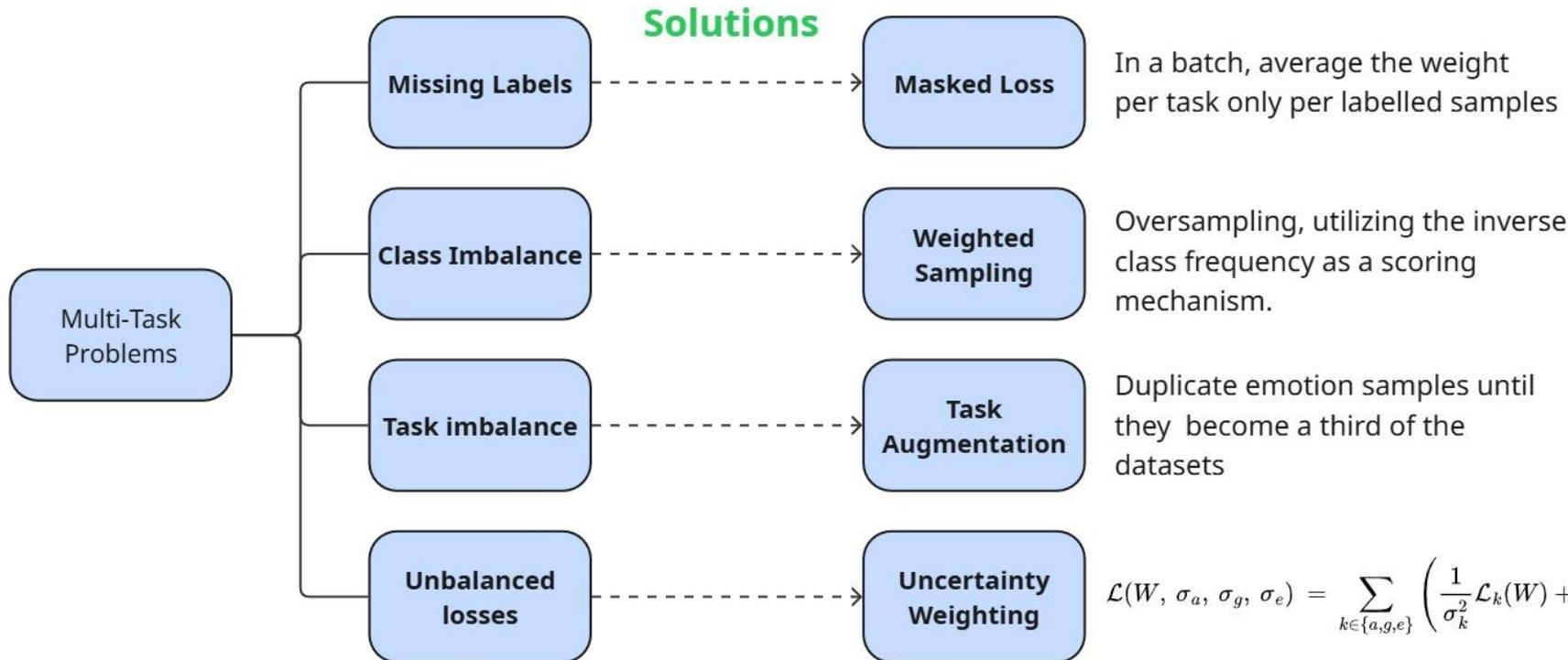


- . Intra task-class imbalance
- . Task imbalance
- . Missing labels



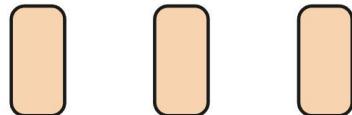


Multi-task learning, problems and solutions



$$\mathcal{L}(W, \sigma_a, \sigma_g, \sigma_e) = \sum_{k \in \{a,g,e\}} \left(\frac{1}{\sigma_k^2} \mathcal{L}_k(W) + \log(\sigma_k^2) \right)$$

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



MLP HEAD MLP HEAD MLP HEAD

ATTENTION POOLING

TRANSFORMERS

TRANSFORMERS

TRANSFORMERS

TRANSFORMERS

 Trained
 Frozen

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? Not really...

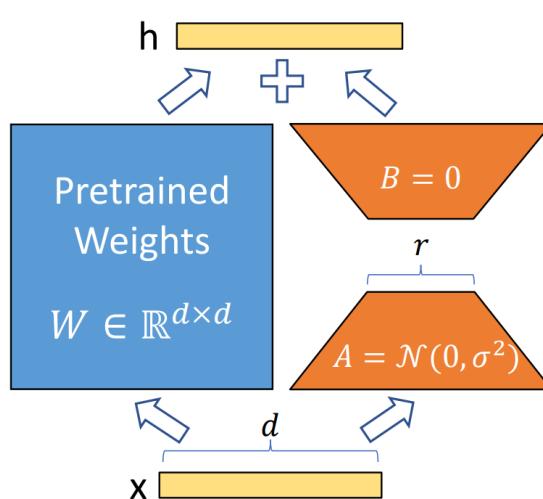
- . **Hardware limitations:** our available hardware does not allow a full fine-tune of the model, even with mixed precision training and gradient checkpointing. It may be possible by a drastic reduction of the batch sizes, but then we incur on the problem of noisy gradients, accentuated by the multi-task setting.

Do we want to go deeper? Yes but...

- . **Increase capacity + relatively small dataset:** when we increase the model capacity by unfreezing more transformers block, if we do not also increase the size of the dataset we may likely encounter **overfitting**.
- . **Risk of catastrophic forgetting:** updating the entire network risks catastrophic forgetting, where **task-specific gradients could destroy the powerful, generalized knowledge** acquired during the initial 5.4 billion pair pre-training.

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.



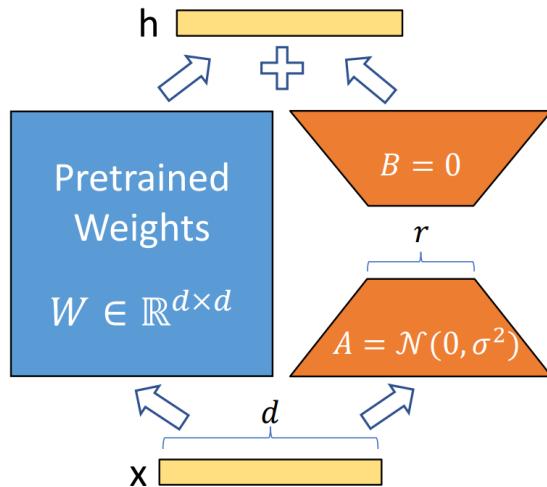
$$h = Wx + BAx$$

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-dimensional space

Enhancements for LoRA: LoRA+ & DoRA

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



Training update for LoRA+:

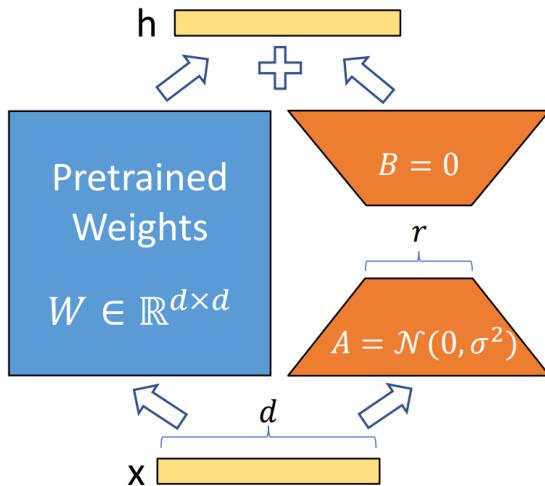
$$B^t = B^{t-1} - \lambda \eta G_B$$
$$A^t = A^{t-1} - \eta G_A$$

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

Enhancements for LoRA: LoRA+ & DoRA

LoRA+, improved performance and faster fine-tuning.

Training update for A and B matrices with LoRA+:

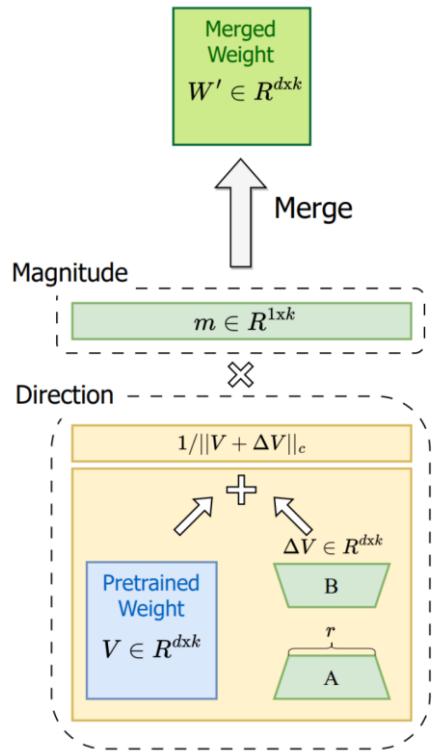


$$B^t = B^{t-1} - \lambda \eta G_B$$

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

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

Enhancements for LoRA: LoRA+ & DoRA

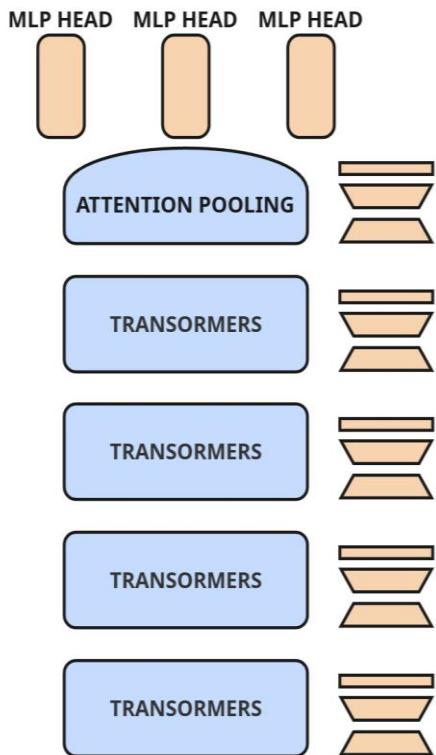


Main Idea: enhance the weight update by separately training the direction and magnitude of the update.

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

Benefits of DoRA:

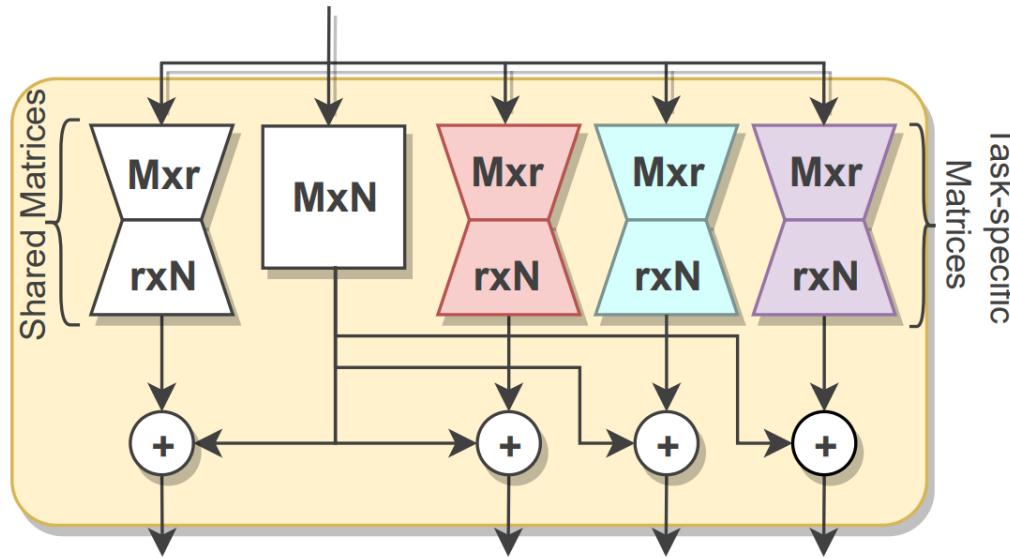
- Same benefits of LoRA
- Increased learning capacity



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



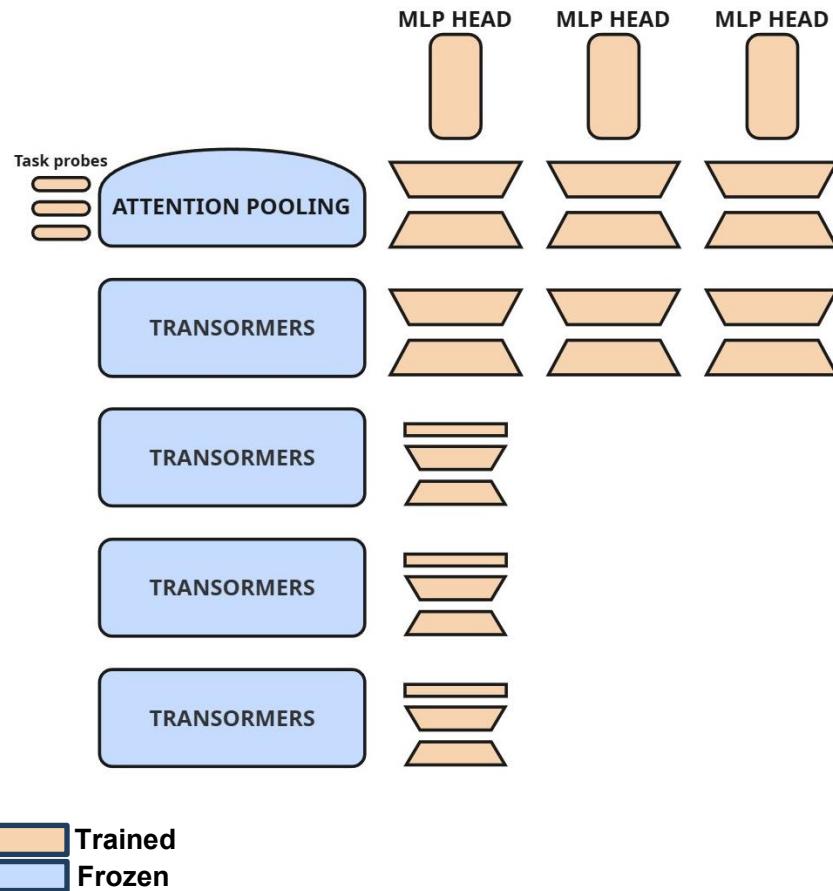
Main Idea: disentangle the parameter space through **Task-Specific LoRA matrices**

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)	61.80%	97.57%	88.78%	82.72%
FT_4 (MTL)	62.56%	97.40%	88.42%	<u>82.79%</u>
LoRA (ST)	62.89%	97.56%	90.83%	83.76%
LoRA (MTL)	62.92%	97.49%	91.21%	<u>83.87%</u>
MTLoRA	64.96%	97.49%	90.06%	<u>83.84%</u>

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

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%	97.00%	57.75%	97.82%
FT_4	63.45%	97.71%	88.42%	62.54%	96.68%	61.68%	97.81%
LoRA	63.73%	97.57%	91.21%	63.34%	96.90%	61.69%	98.00%
MTLoRA	64.11%	97.62%	90.06%	63.96%	96.93%	63.80%	97.93%



Efficiency Comparison

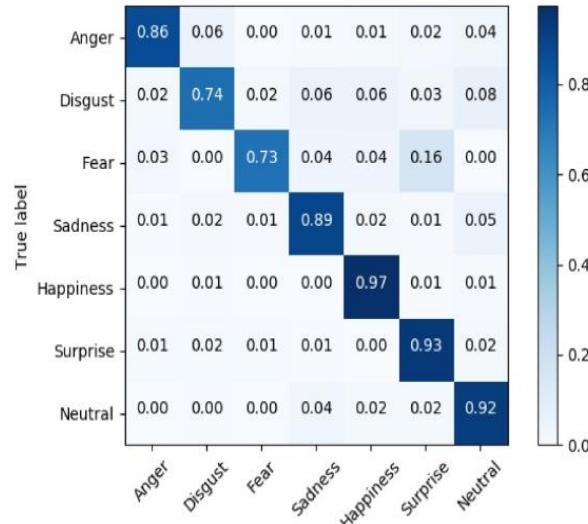
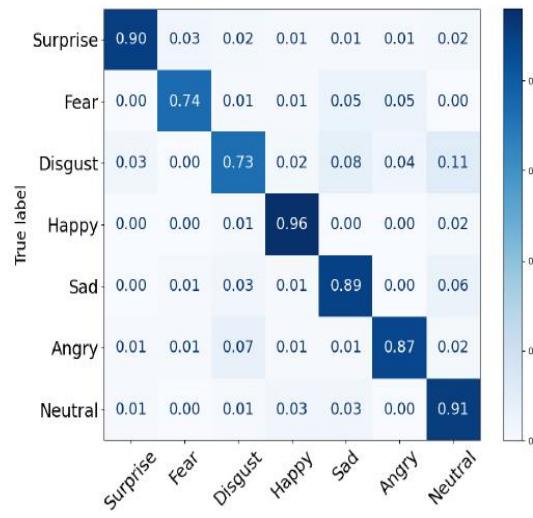
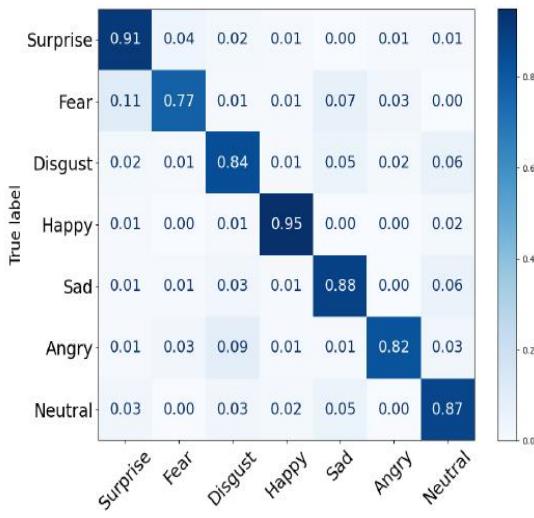
Model	GLOPS	PARAMS	AVG. ACC
Baseline	700	671 M	69.61%
LoRA	352 (-50%)	320 M (-52%)	84.07% (+20%)
MTLoRA	368 (-47%)	329 M (-50%)	83.89% (+20%)



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	-	-	94.76%
ApViT	-	-	92.21%
Baseline	46.11%	97.60%	66.57%
LoRA	63.73%	97.57%	91.21%
MTLoRA	64.11%	97.62%	90.06%



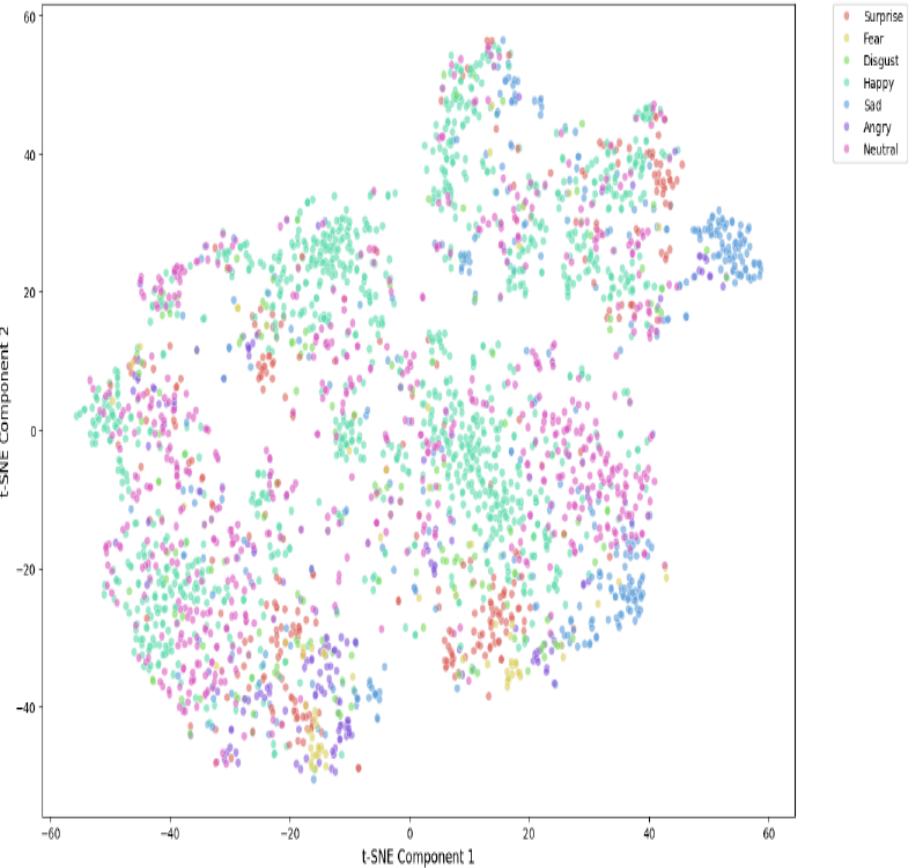


(a) MTLoRA confusion matrix on RAF-DB, balanced accuracy of 86.17 (Acc. 90.06)

85.90 (Acc. 91.21)

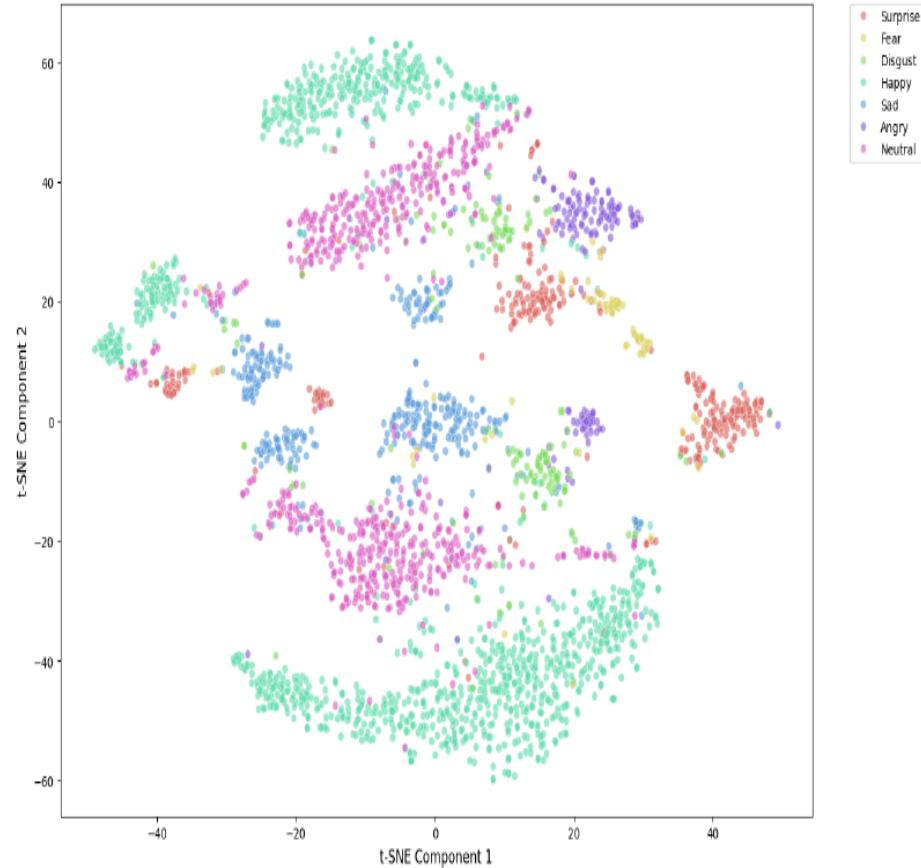
86.36 (Acc. 92.21)

t-SNE of Backbone Features (RAF-DB Test Set)



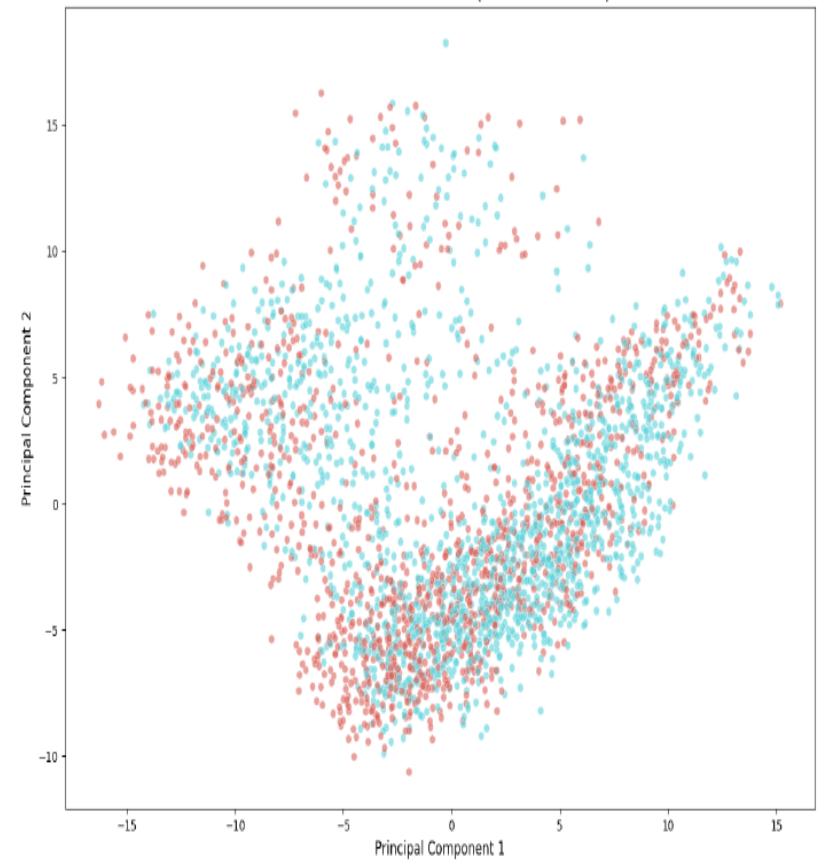
(a) t-SNE Emotion (Untrained)

t-SNE of Backbone Features (RAF-DB Test Set)



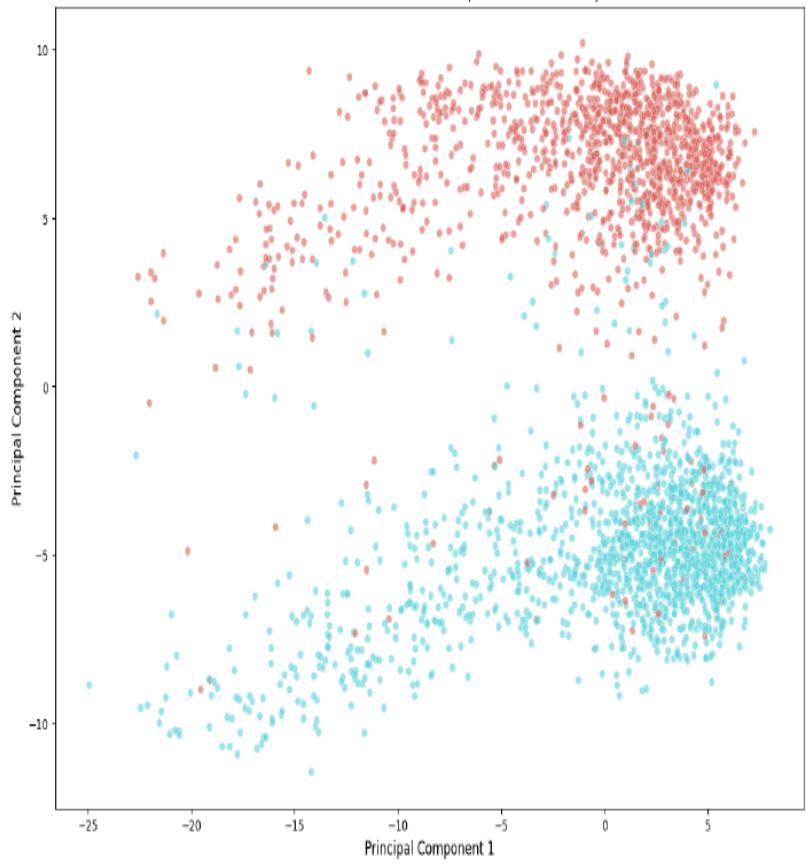
(b) t-SNE Emotion (Trained)

PCA of Backbone Features (RAF-DB Test Set)



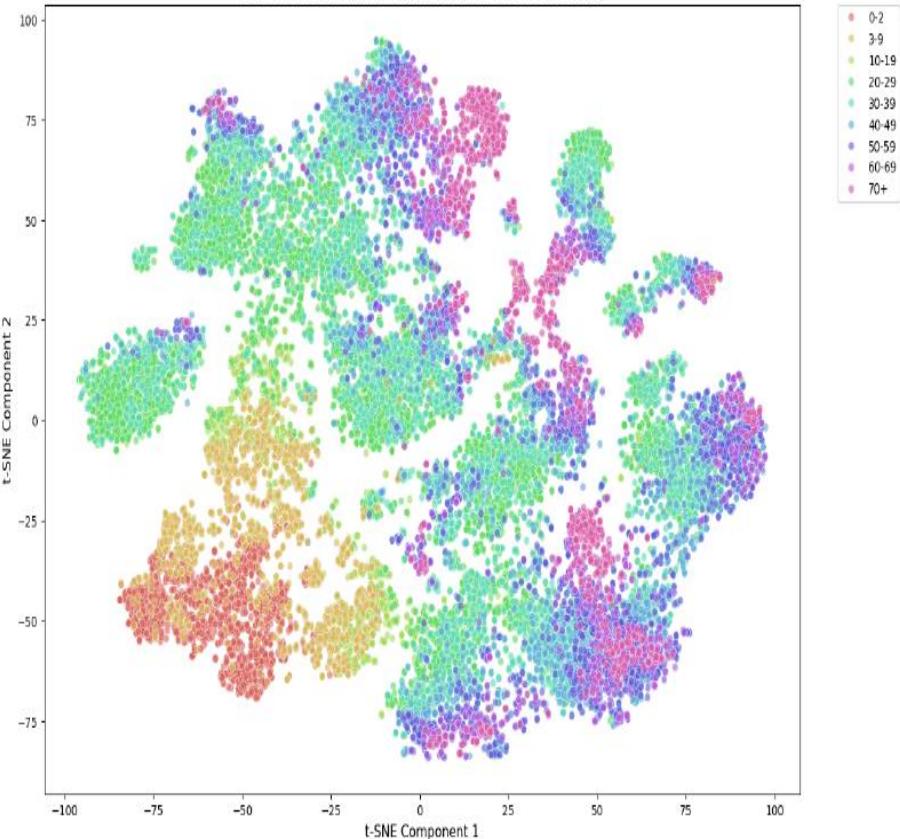
(g) PCA Gender (Untrained)

PCA of Backbone Features (RAF-DB Test Set)



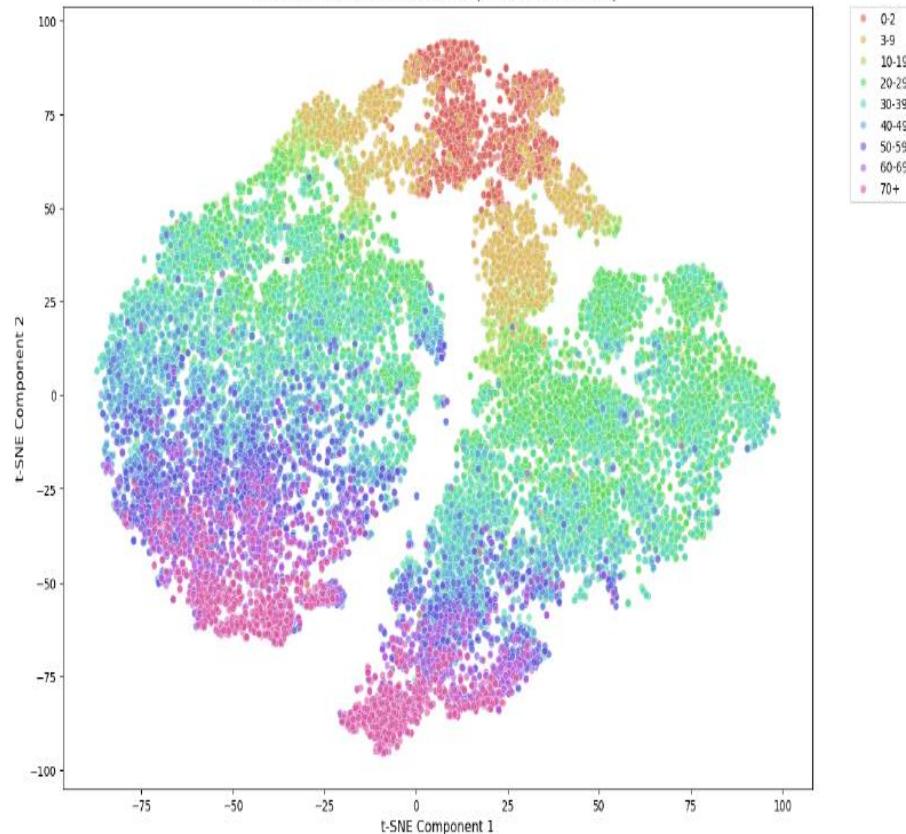
(h) PCA Gender (Trained)

t-SNE of Backbone Features (UTKFace Test Set)



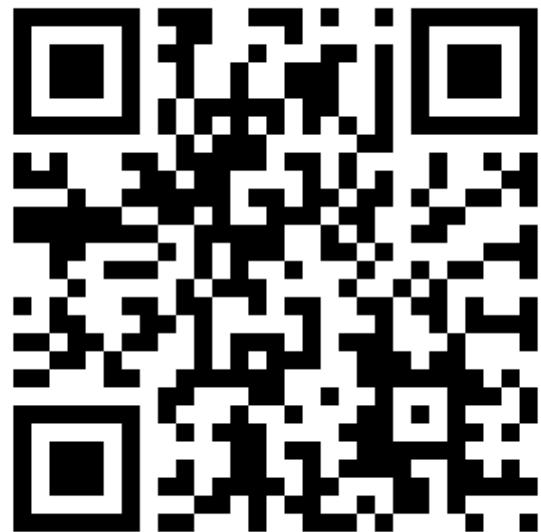
(a) t-SNE Age (Untrained)

t-SNE of Backbone Features (UTKFace Test Set)



(b) t-SNE Age (Trained)

Dimostratore



/login tesi2025



UNIVERSITÀ DEGLI STUDI DI SALERNO

