科技社团党建

党建强会行动中ing...

CSIG · 云讲堂

# 大模型时代的行人属性识别

## *Pedestrian Attribute Recognition in the Big Model Era*

王逍　2024.08.27 19:00

*xiaowang@ahu.edu.cn*

安徽大学　计算机科学与技术学院

# *Overview*



- **Background of Pedestrian Attribute Recognition (PAR)**
  - Task definition, Review of PAR (Major Challenges, Datasets, Mainstream Algorithms, … )
- **When Big Models Meet PAR**
  - CLIP, MAE, LLM, …
  - VTB, PromptPAR, SequencePAR, LLM-PAR, …
  - Applications on Other Tasks
- **Conclusion & Discussion**

# *Overview*



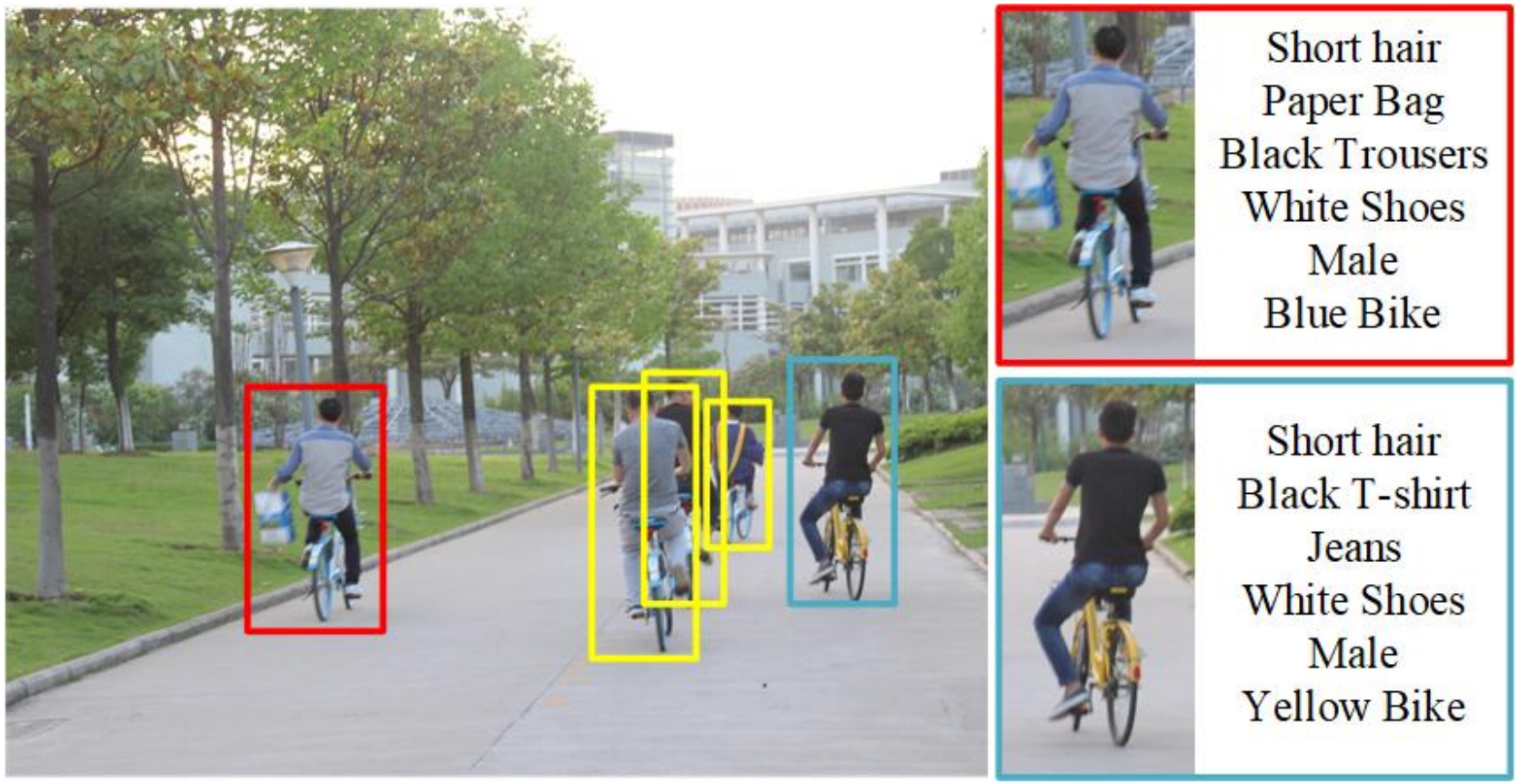
- **Background of Pedestrian Attribute Recognition (PAR)**
  - Task definition, Review of PAR (Major Challenges, Datasets, Mainstream Algorithms, … )
- **When Big Models Meet PAR**
  - CLIP, MAE, LLM, …
  - VTB, PromptPAR, SequencePAR, LLM-PAR, …
  - Applications on Other Tasks
- **Conclusion & Discussion**

# Background of PAR



➢ *Task Definition*

***Pedestrian Attribute Recognition: A Survey,*** *Pattern Recognition, 2021*
*Xiao Wang, Shaofei Zheng, Rui Yang, Aihua Zheng, Zhe Chen, Jin Tang, and Bin Luo*




https://arxiv.org/pdf/1901.07474

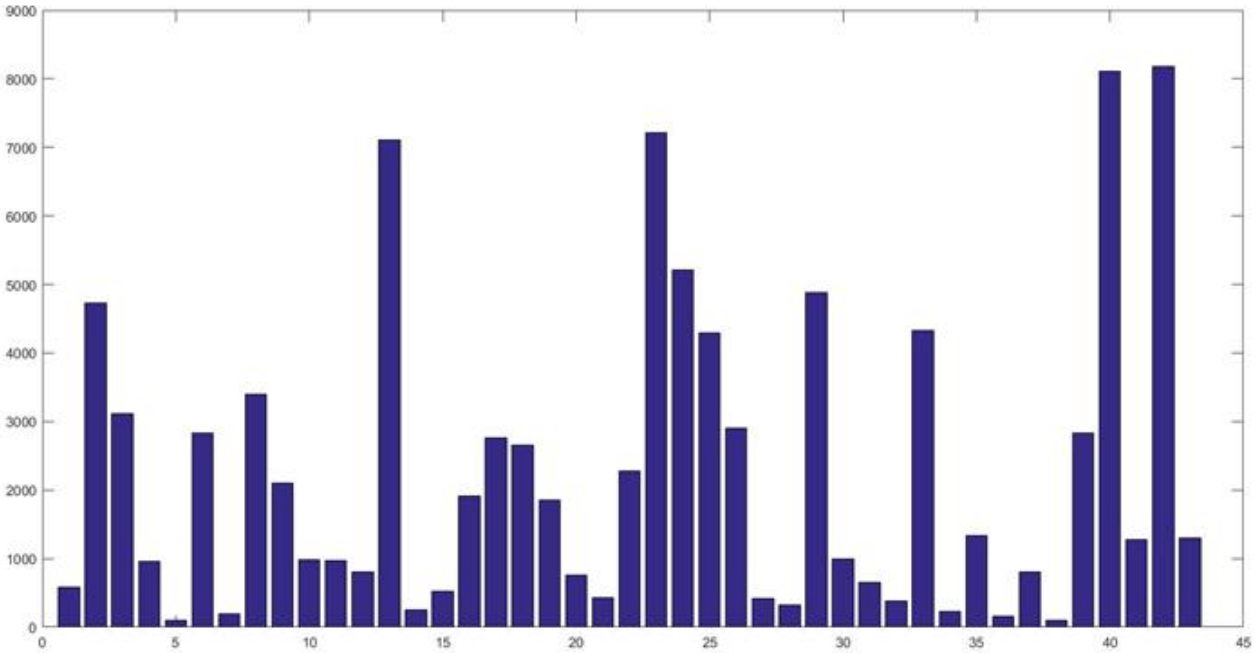
# Background of PAR



## ➢ *Major Challenges*

https://arxiv.org/pdf/1901.07474

(a) Shadow (b) Blur (c) Occlusion (d) Low Resolution (e) Multi-views (f) Night time




(g) Unbalanced Data Distribution (from PETA data, 43 attributes, 9500 training images)

Short hair
Black T-shirt
Jeans
White Shoes
Male
Yellow Bike

☐ **Semantic gaps** between *vision features* and *attribute labels*

# Background of PAR



## ➢ Benchmark Datasets

https://arxiv.org/pdf/1901.07474

| Dataset | #Pedestrians | #Attributes | Source | Project Page |
|---|---|---|---|---|
| PETA [34] | 19000 | 61 binary and 4 multi-class attributes | outdoor & indoor | URL |
| RAP [35] | 41585 | 69 binary and 3 multi-class attributes | indoor | URL |
| RAP-2.0 [36] | 84928 | 69 binary and 3 multi-class attributes | indoor | URL |
| PA-100K [18] | 100000 | 26 binary attributes | outdoor | URL |
| WIDER [16] | 13789 | 14 binary attributes | WIDER images [43] | URL |
| Market-1501 [37] | 32668 | 26 binary and 1 multi-class attributes | outdoor | URL |
| DukeMTMC [37] | 34183 | 23 binary attributes | outdoor | URL |
| PARSE-27K [5], [39] | 27000 | 8 binary and 2 multi-class orientation attributes | outdoor | URL |
| APiS [40] | 3661 | 11 binary and 2 multi-class attributes | KITTI [44] , CBCL Street Scenes [45], INRIA [1] and SVS | URL |
| HAT [41] | 9344 | 27 binary attributes | image site Flickr | URL |
| CRP [42] | | | outdoor | URL |
| CAD [38] | 1856 | 23 binary attributes and 3 multi-class attributes | image site Sartorialist* and Flickr | URL |
| BAP [8] | 8035 | 9 binary attributes | H3D [46] dataset PASCAL VOC 2010 [47] | URL |
| MARS-Attributes [48] | 20,478 tracklets (1,261 people) | 20 attributes | MARS | URL |
| DukeMTMC-VID-Attributes [48] | 4,832 tracklets (1,402 people) | 18 attributes | DukeMTMC-VID | URL |
| UAV-Human [49] | 22,263 | 7 attributes | outdoor | URL |
| UPAR [50] | - | 40 attributes | PA100K, PETA, RAPv2, and Market1501 | URL |

**Video-based PAR datasets**

- **MSP60K Dataset**: https://github.com/Event-AHU/OpenPAR

# Background of PAR

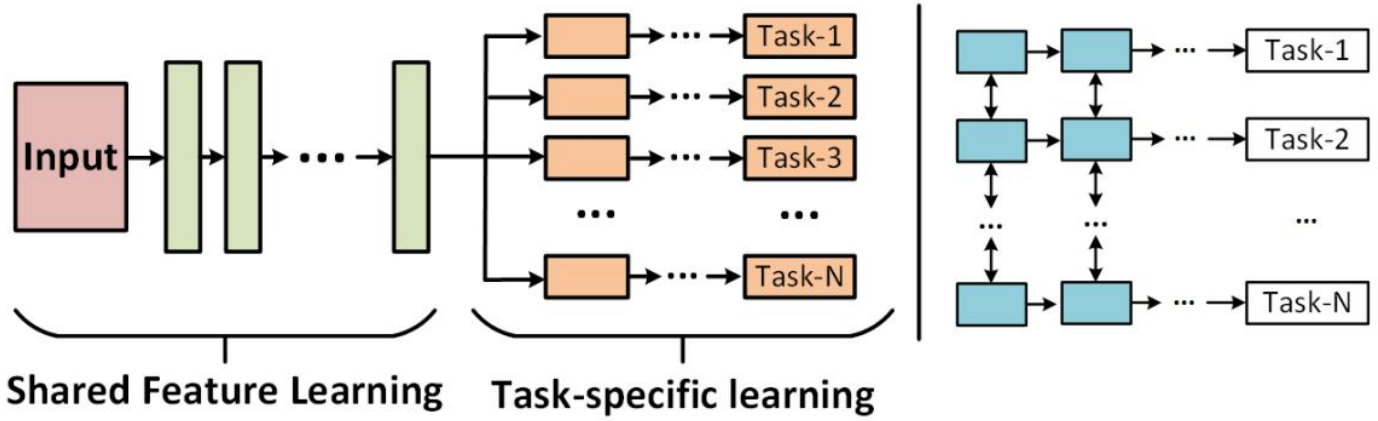


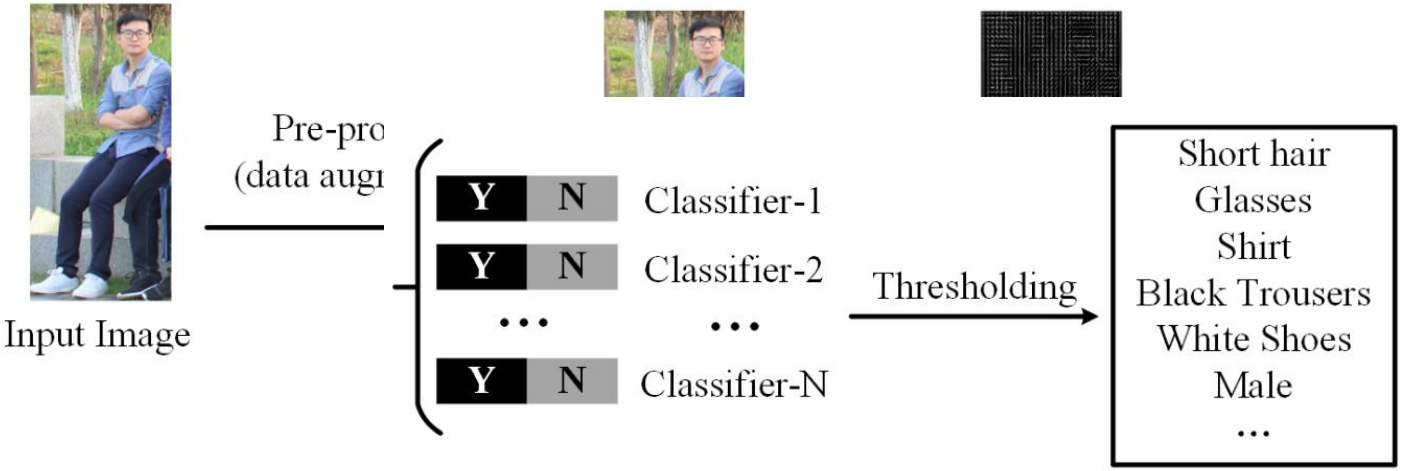
## ➢ Mainstream Algorithms

| | |
|---|---|
| Global-based | ACN, DeepSAR, DeepMAR, MTCNN |
| Part-based | Poselets, RAD, PANDA, MLCNN, AAWP, ARAP, DeepCAMP, PGDM, DHC, LGNet |
| Attention-based | HPN, VeSPA, DIAA, CAM |
| New Loss Function-based | WPAL, AWMT, |
| Sequential-based | CNN-RNN, JRL, GRL, JCM, RCRA |
| Curriculum Learning-based | MTCT, CILICIA |
| Graphic Model-based | DSCA, A-AOG, VSGA |
| Others | PatchIt, FaFS, GAM |

*Vision-language fusion, LLM, Pre-trained Big Model*

https://arxiv.org/pdf/1901.07474



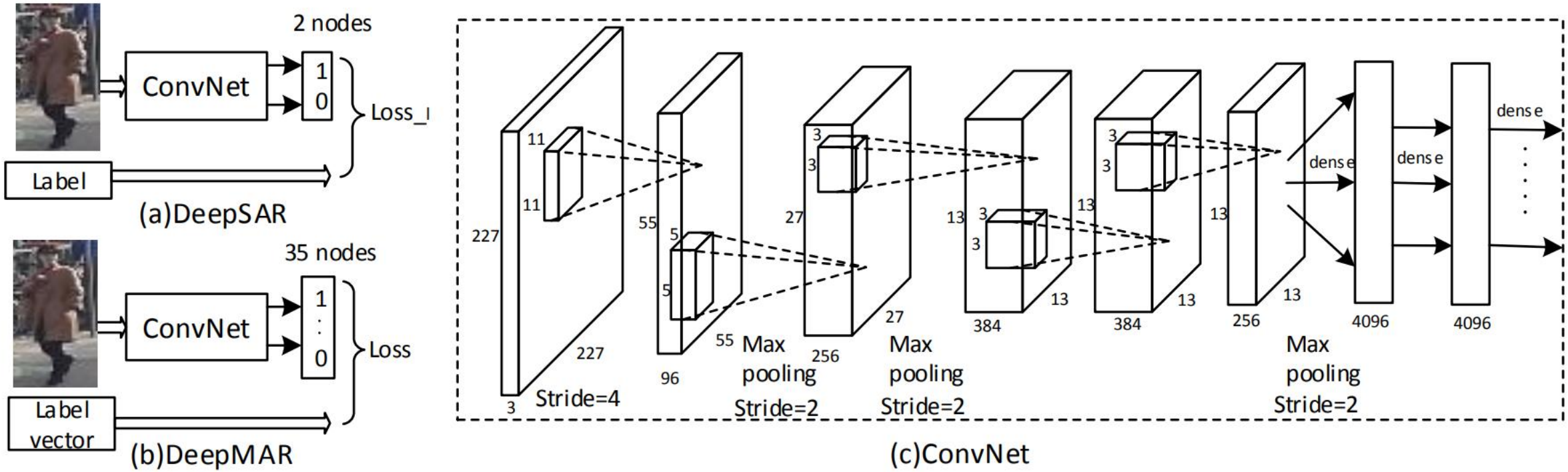
# Background of PAR



## ➢ *Mainstream Algorithms*

*Global-based PAR*

- Li, Dangwei, et al. "**Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios**." 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR). IEEE, 2015.



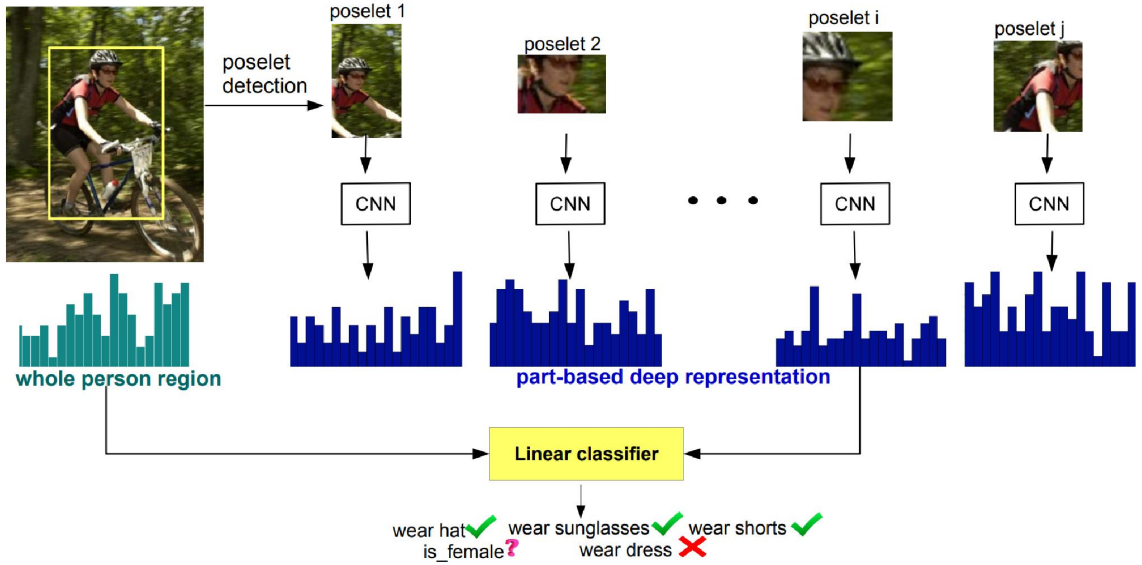
# Background of PAR



## ➢ *Mainstream Algorithms*

Part-based PAR

■ Zhang, Ning, et al. "***Panda: Pose aligned networks for deep attribute modeling***." *CVPR*-2014.

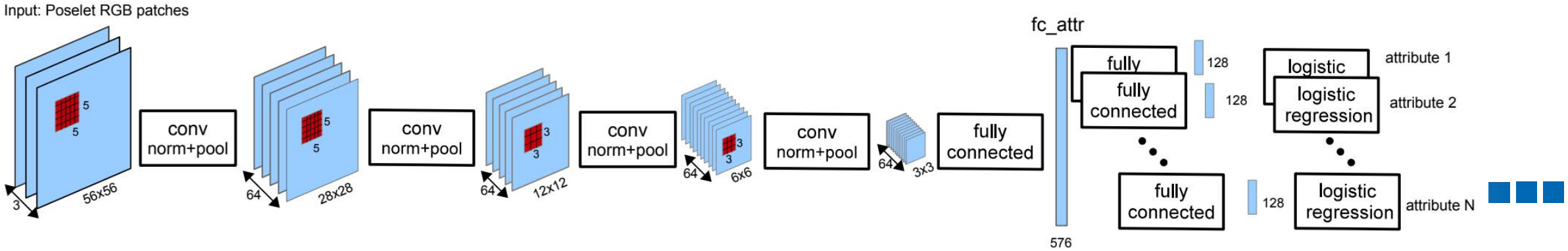



| Attribute | male | long hair | glasses | hat | tshirt | longsleeves | shorts | jeans | long pants | Mean AP |
|---|---|---|---|---|---|---|---|---|---|---|
| Poselets[4] | 82.4 | 72.5 | 55.6 | 60.1 | 51.2 | 74.2 | 45.5 | 54.7 | 90.3 | 65.18 |
| DPD[27] | 83.7 | 70.0 | 38.1 | 73.4 | 49.8 | 78.1 | 64.1 | 78.1 | 93.5 | 69.88 |
| Joo *et al.* [14] | 88.0 | 80.1 | 56.0 | **75.4** | **53.5** | 75.2 | 47.6 | 69.3 | 91.1 | 70.7 |
| PANDA | **91.7** | **82.7** | **70.0** | 74.2 | 49.8 | **86.0** | **79.1** | **81.0** | **96.4** | **78.98** |

Table 1: Attribute classification results on the Berkeley Attributes of People Dataset as compared to the methods of Bourdev *et al.* [4] and Zhang *et al.* [27] .

| Attribute | male | long hair | hat | glasses | dress | sunglasses | short sleeves | baby | mean AP |
|---|---|---|---|---|---|---|---|---|---|
| Poselets150[4] | 86.00 | 75.31 | 29.03 | 36.72 | 34.73 | 50.16 | 55.25 | 41.26 | 51.06 |
| DPD[27] | 85.84 | 72.40 | 27.55 | 23.94 | 48.55 | 34.36 | 54.75 | 41.38 | 48.60 |
| DeCAF [8] | 82.47 | 65.03 | 19.15 | 14.91 | 44.68 | 26.91 | 56.40 | 50.19 | 44.97 |
| DL-DPM | 88.27 | 77.64 | **43.44** | 36.70 | 55.72 | 55.03 | 67.95 | 64.89 | 61.20 |
| PANDA | **94.10** | **83.17** | 39.52 | **72.25** | **59.41** | **66.62** | **72.09** | **78.76** | **70.74** |

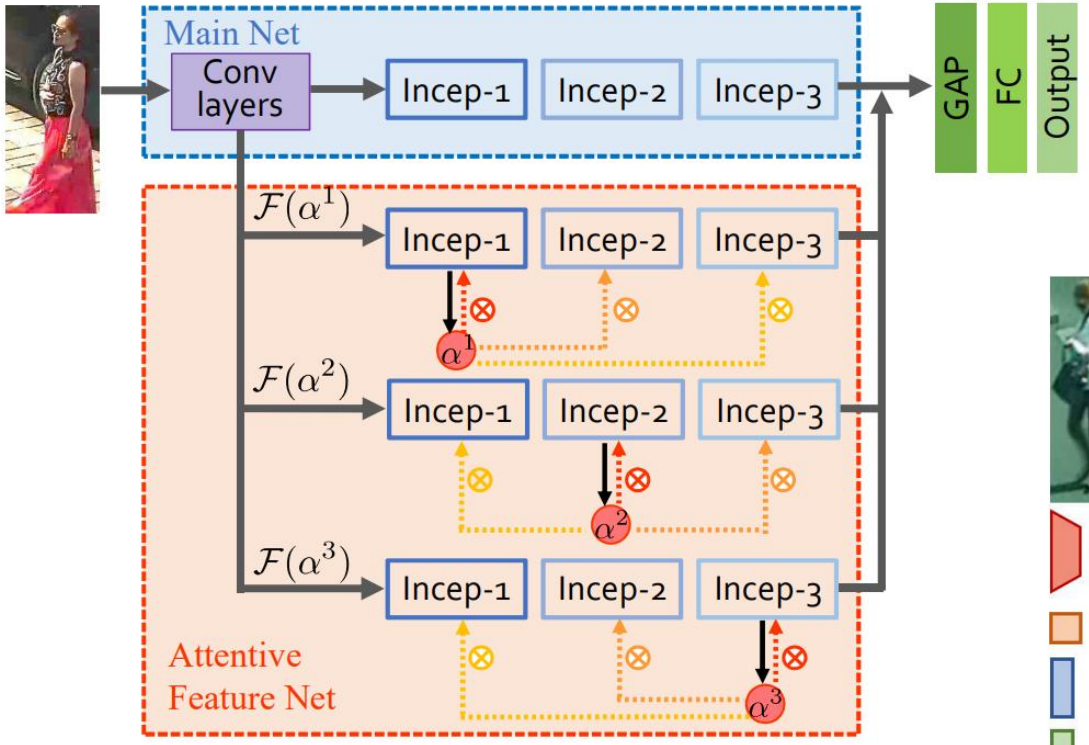
Table 2: Average Precision on the Attributes25K-test dataset.



# Background of PAR

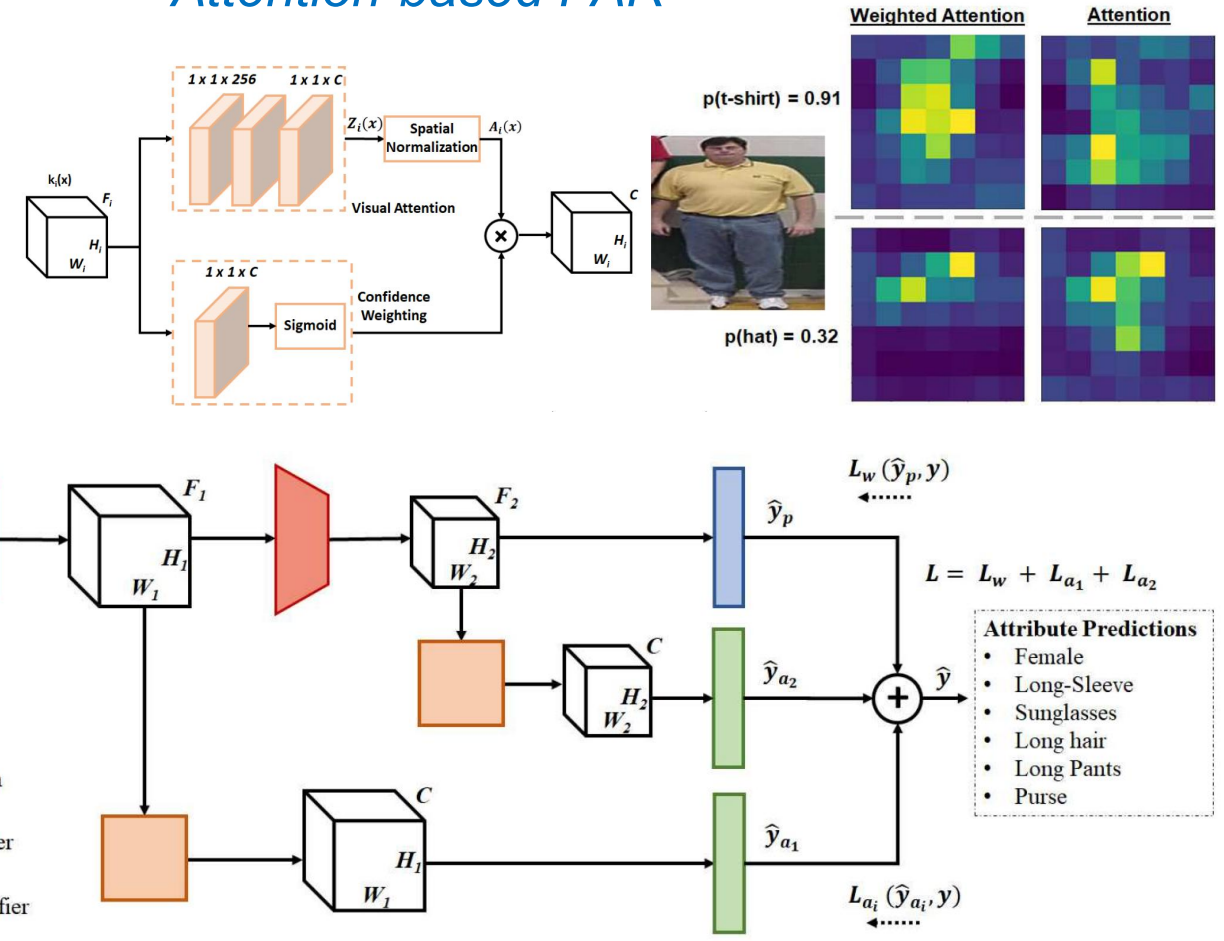


## ➢ *Mainstream Algorithms*




Liu, Xihui, et al. "Hydraplus-net: Attentive deep features for pedestrian analysis." *ICCV*-2017.

## *Attention-based PAR*



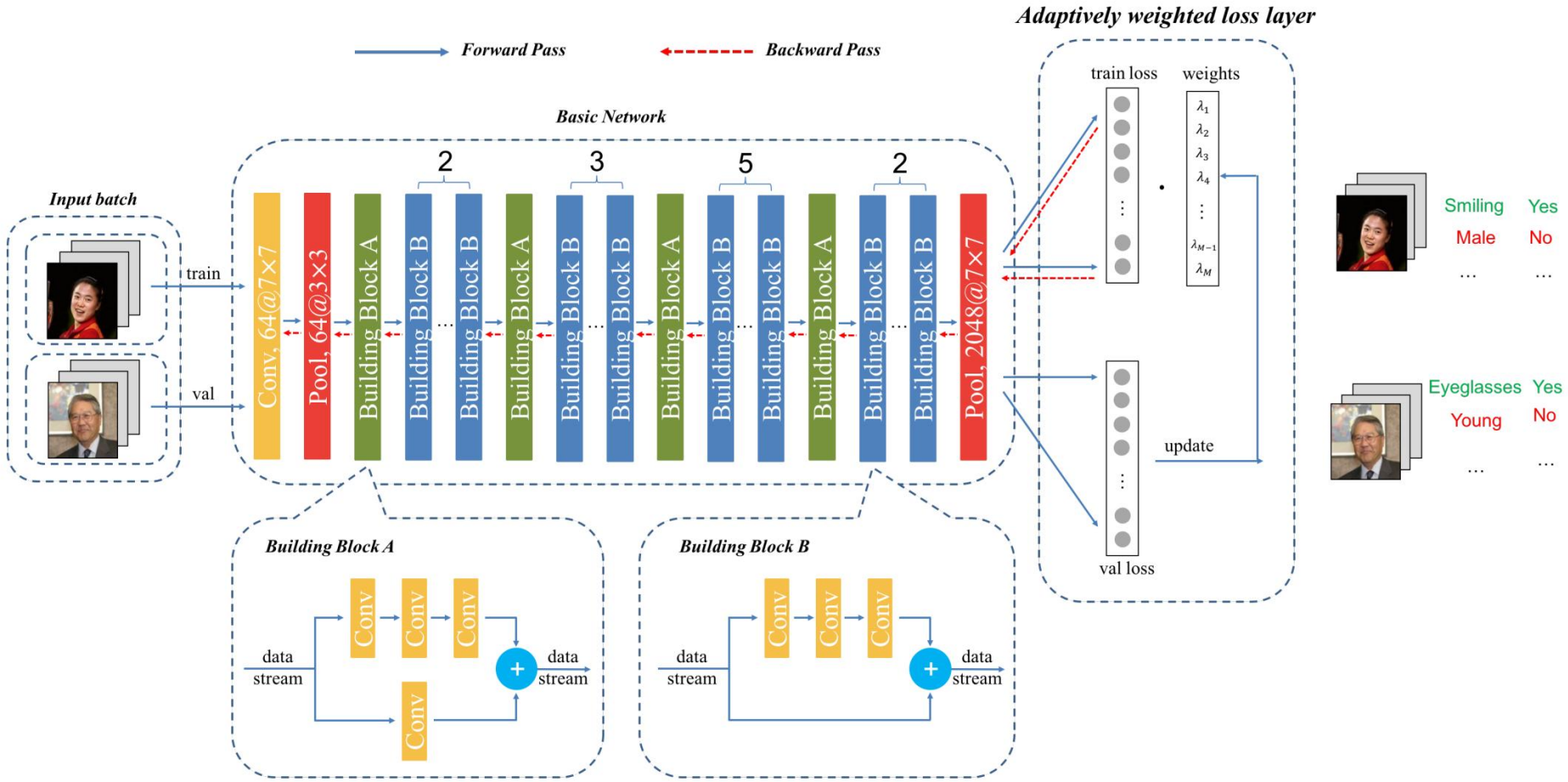

Sarafianos, et al. "Deep imbalanced attribute classification using visual attention aggregation." ECCV-2018.

# Background of PAR



## ➢ Mainstream Algorithms

New Loss Function-based PAR



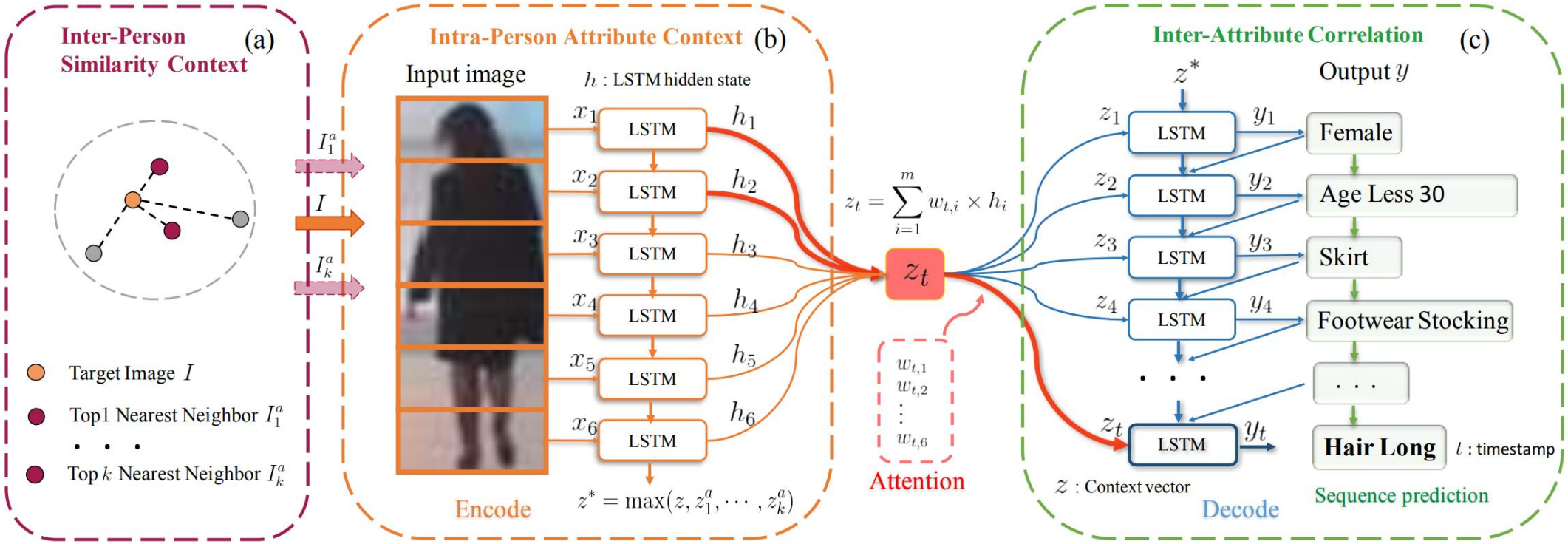

He, Keke, et al. "Adaptively weighted multi-task deep network for person attribute classification." *ACM Multimedia*. 2017.

# Background of PAR



➢ ***Mainstream Algorithms***

*Sequential-based PAR*

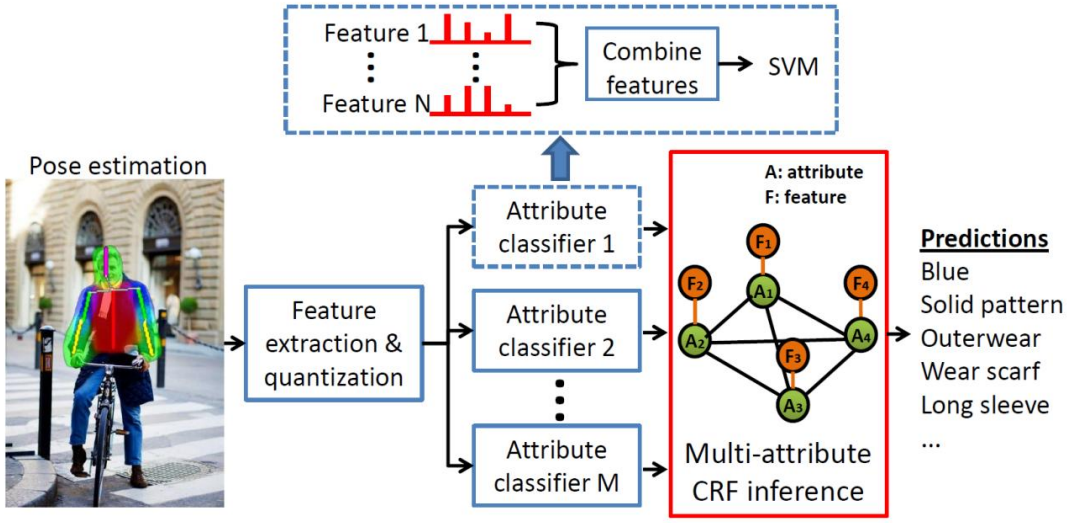



Wang, Jingya, et al. "Attribute recognition by joint recurrent learning of context and correlation." *ICCV*-2017.

# Background of PAR

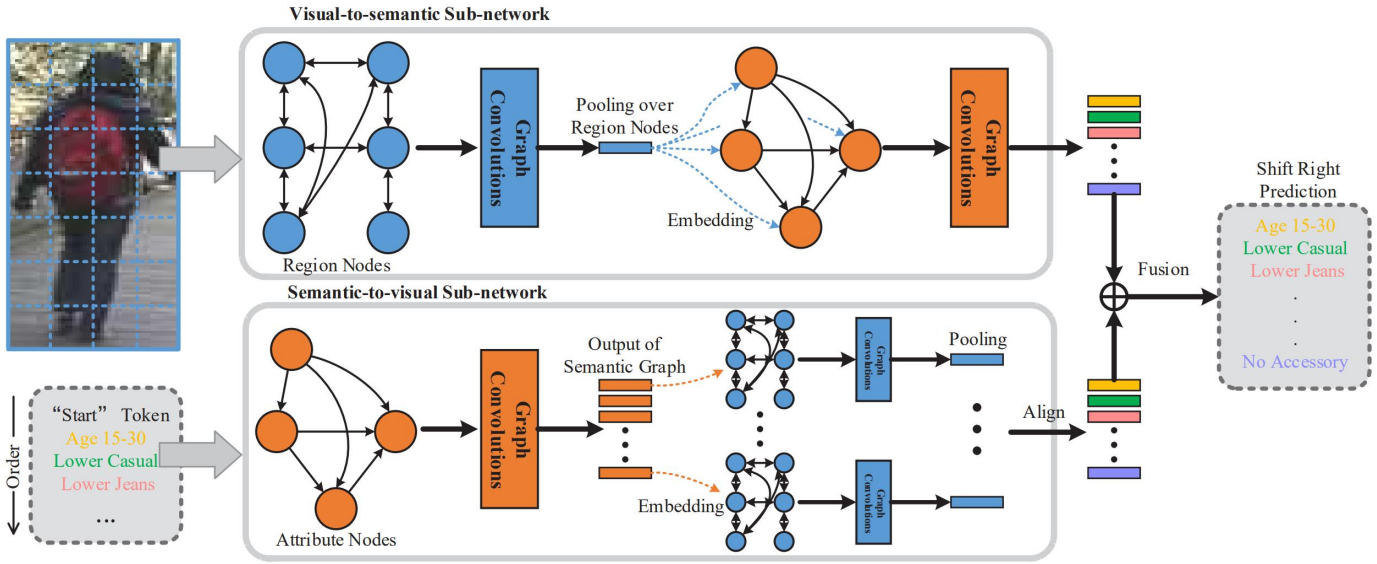


## ➢ *Mainstream Algorithms*




Chen, Huizhong, "Describing clothing by semantic attributes." ECCV-2012.

## Graphic Model-based PAR

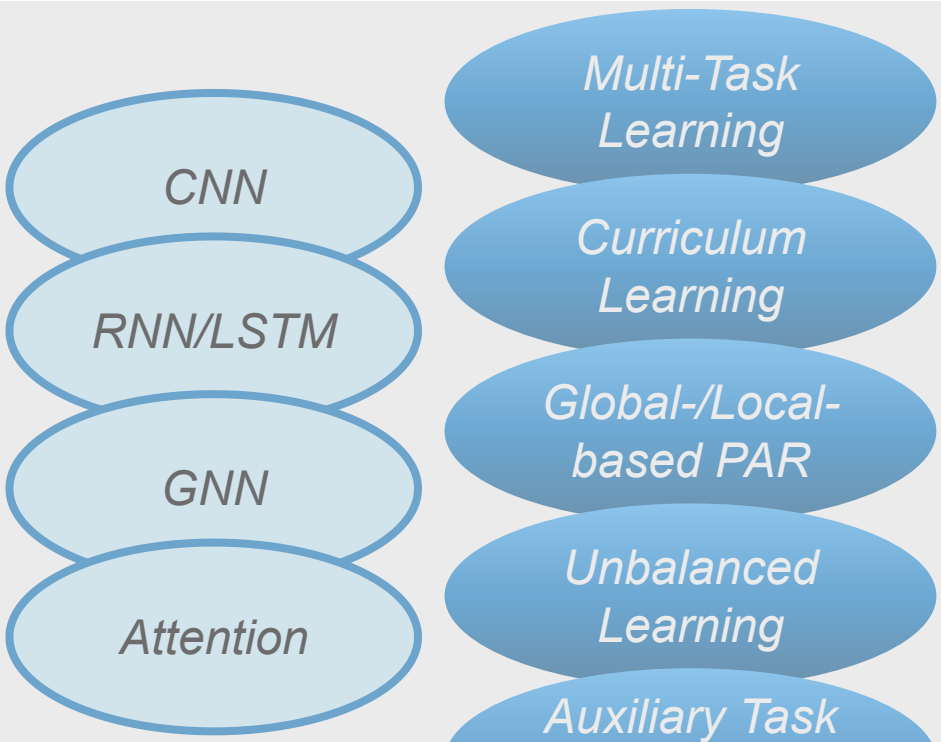


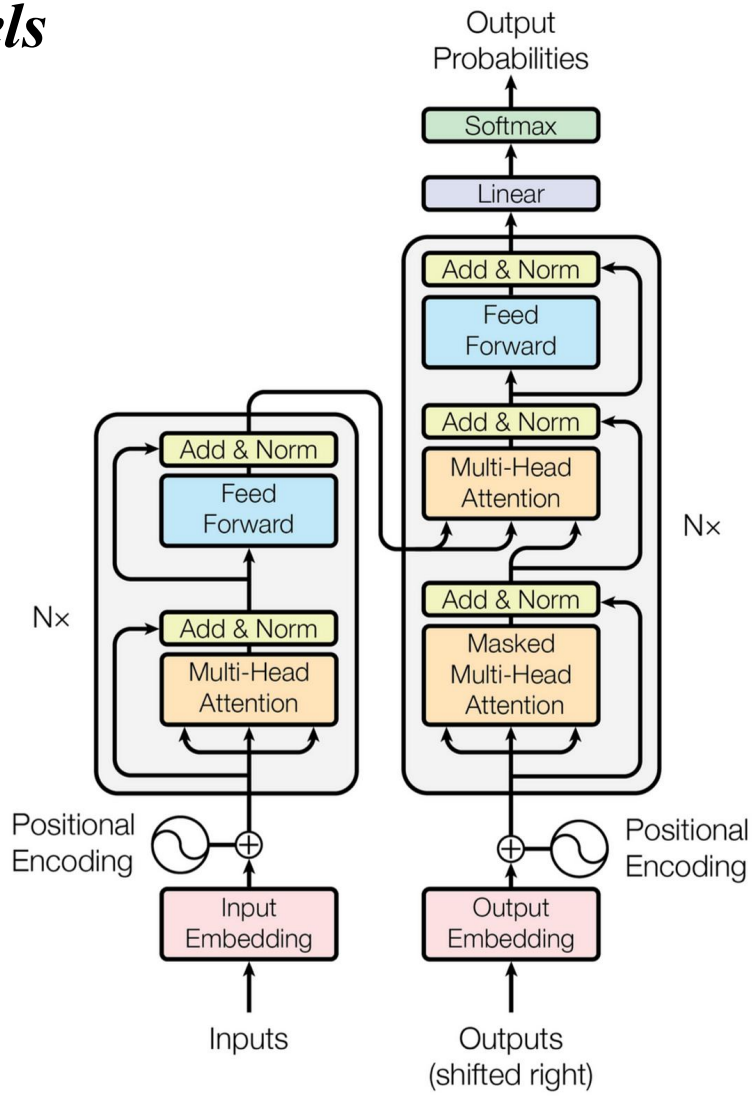

Li, Qiaozhe, et al. "Visual-semantic graph reasoning for pedestrian attribute recognition." *AAAI*-2019.

# Background of PAR



## ➢ Summary of Existing PAR Models




**Pre-trained LLM**

**Pre-trained Vision-Language Models**

# *Overview*



- Background of Pedestrian Attribute Recognition (PAR)
  - Task definition, Review of PAR (Major Challenges, Datasets, Mainstream Algorithms, … )

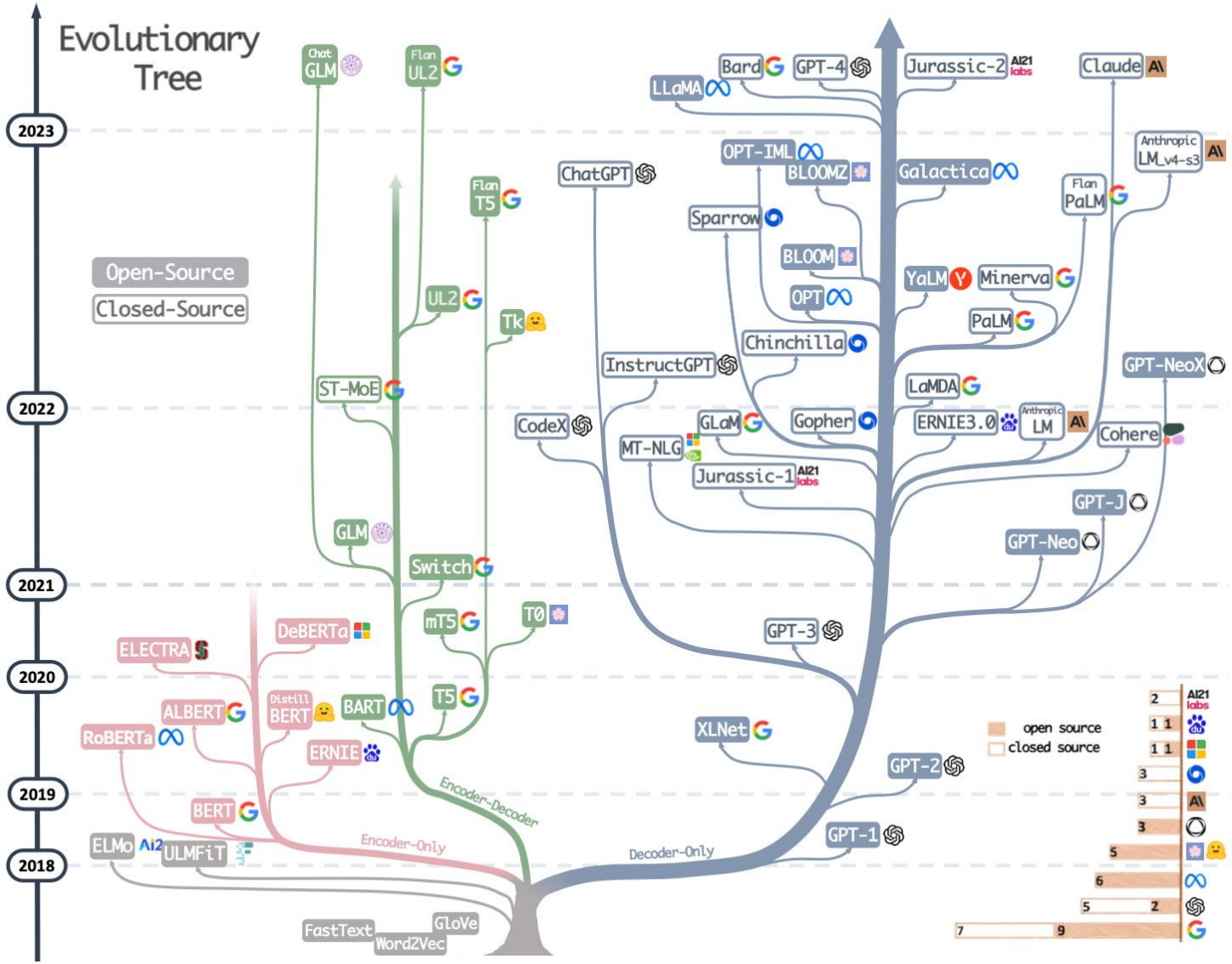
- **When Big Models Meet PAR**
  - CLIP, MAE, LLM, …
  - VTB, PromptPAR, SequencePAR, LLM-PAR, …
  - Applications on Other Tasks

- Conclusion & Discussion

CSIG · 云讲堂

# When Big Models Meet PAR

EVENT-AHU RESEARCH




**GPT-1,**
**GPT-2,**
**GPT-3,**
**GPT-3.5 (ChatGPT),**
**GPT-4,**
*******

**GPT series**

**LLaMA,**
**LLaMA-2,**
**LLaMA-3,**
**LLaMA-3.1**

**LLaMA**

**Baichuan-1,**
**Baichuan-2,**
**Baichuan-3**

**百川系列**

**CLIP**
**EVA**
**Flamingo**
**LLaVA**
**…**

**多模态大模型**

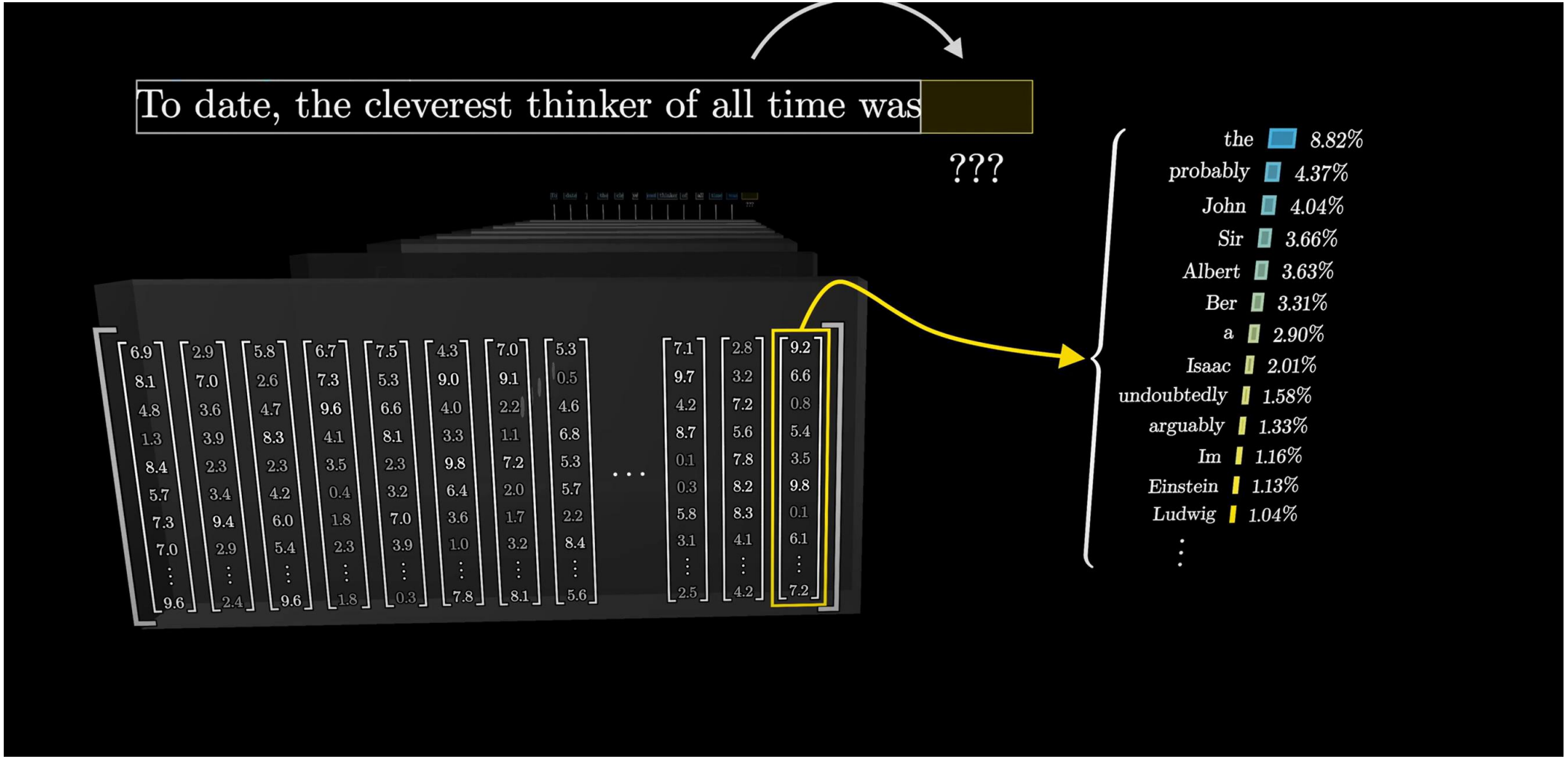
**讯飞 – 星火大模型**
**百度 – 文心**
**阿里 – Qwen**
**PCL – “大圣”**
**华为 – 盘古**
**清华 – 智谱**

**国产大模型**

# When Big Models Meet PAR



## ➢ Next Token Prediction --- Transformer based LLM

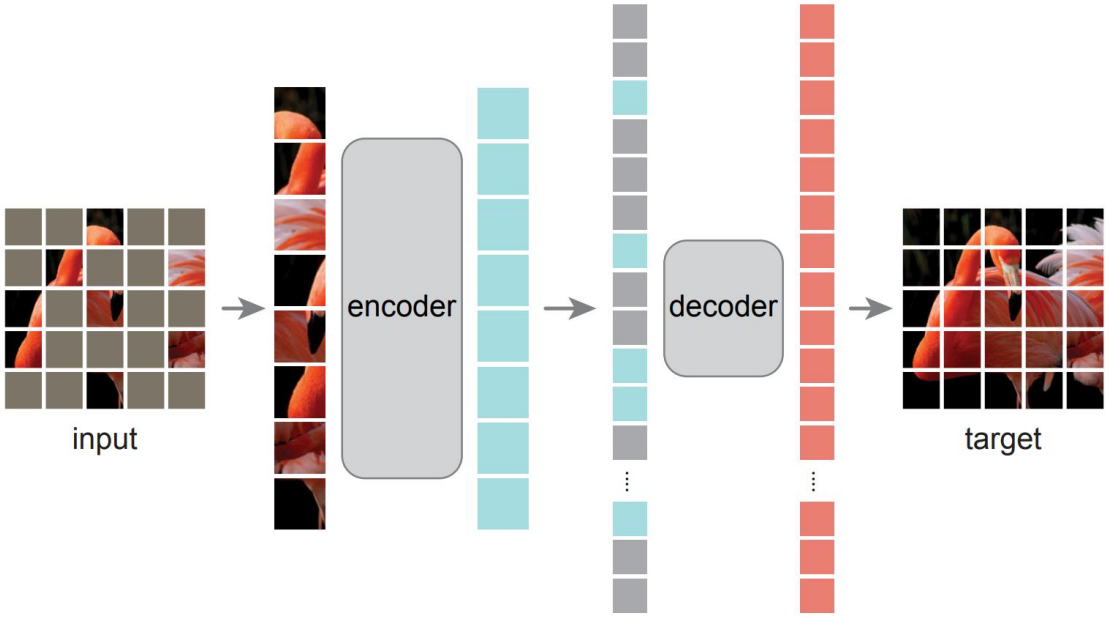


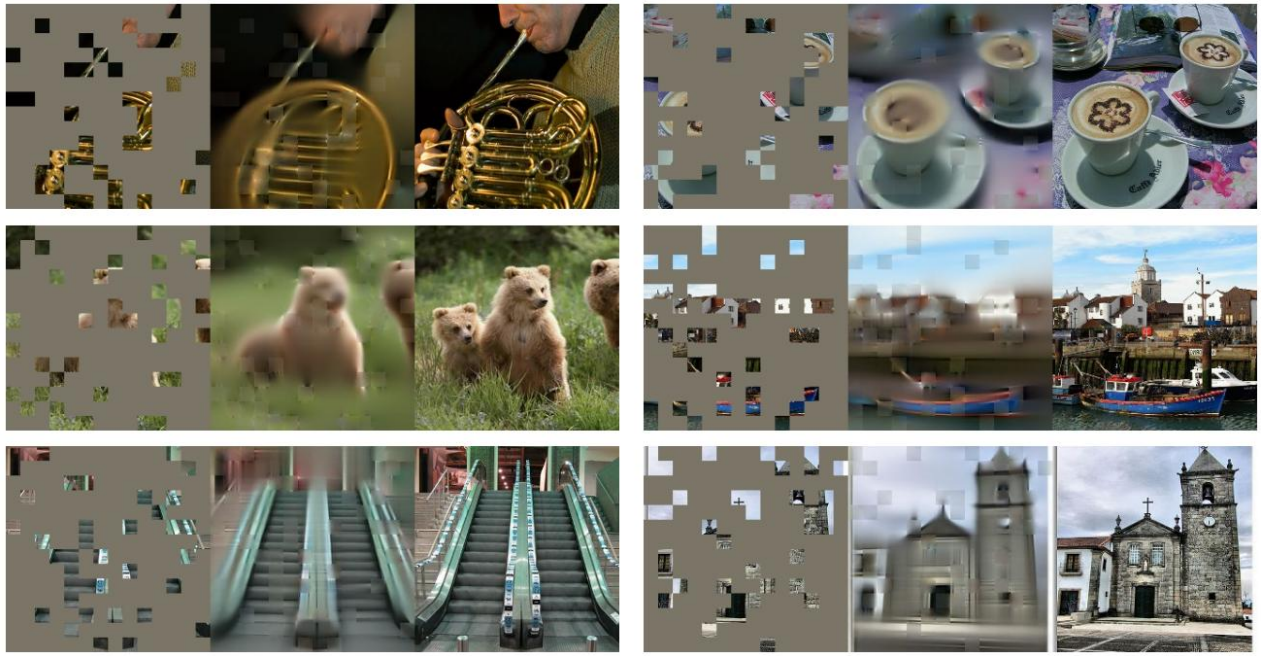
# When Big Models Meet PAR



## ➢ Masked Auto-Encoder --- MAE

He, Kaiming, et al. "Masked autoencoders are scalable vision learners." *CVPR*-2022.

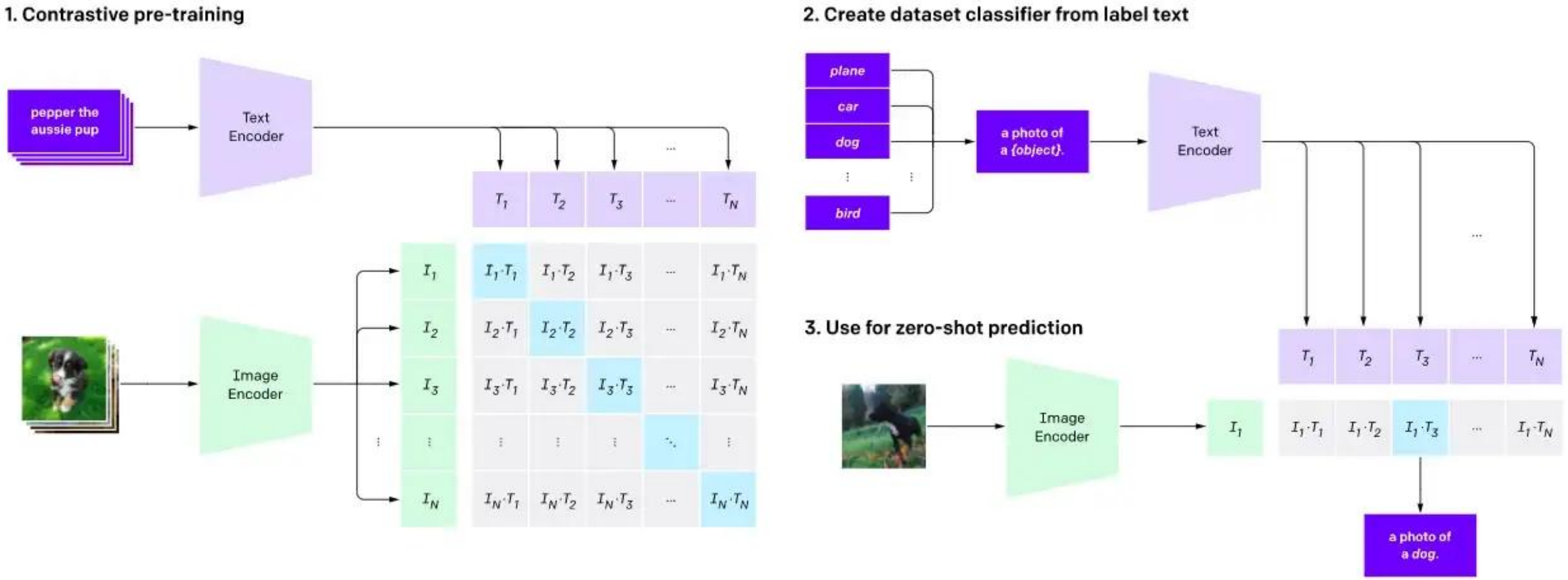


# When Big Models Meet PAR



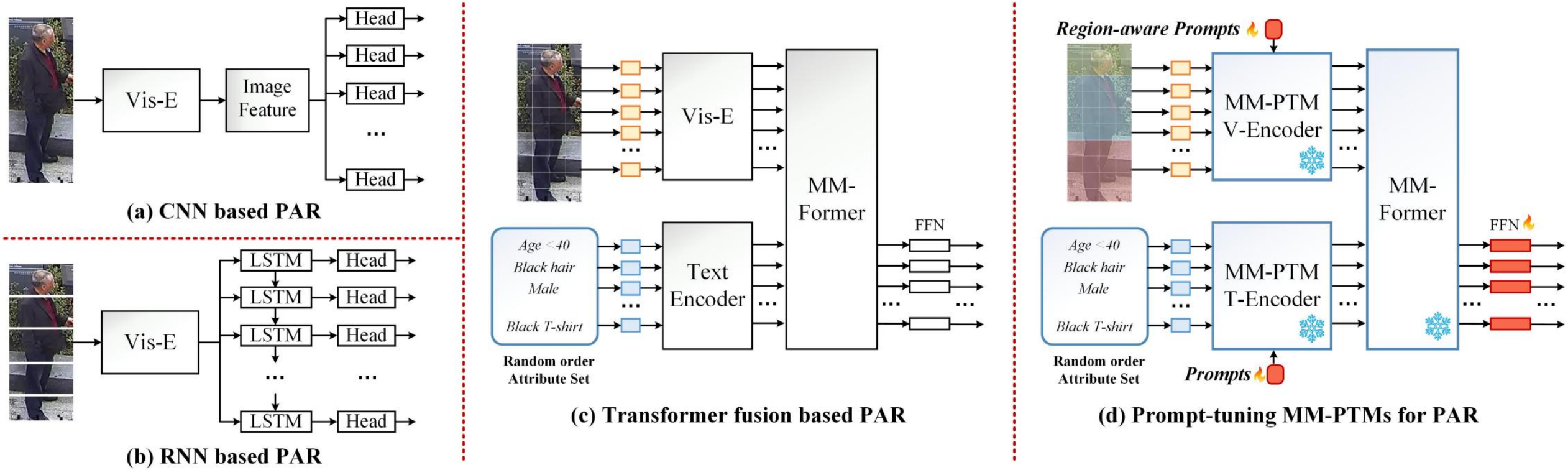
## ➢ Vision-Text Contrastive Learning --- CLIP



# When Big Models Meet PAR



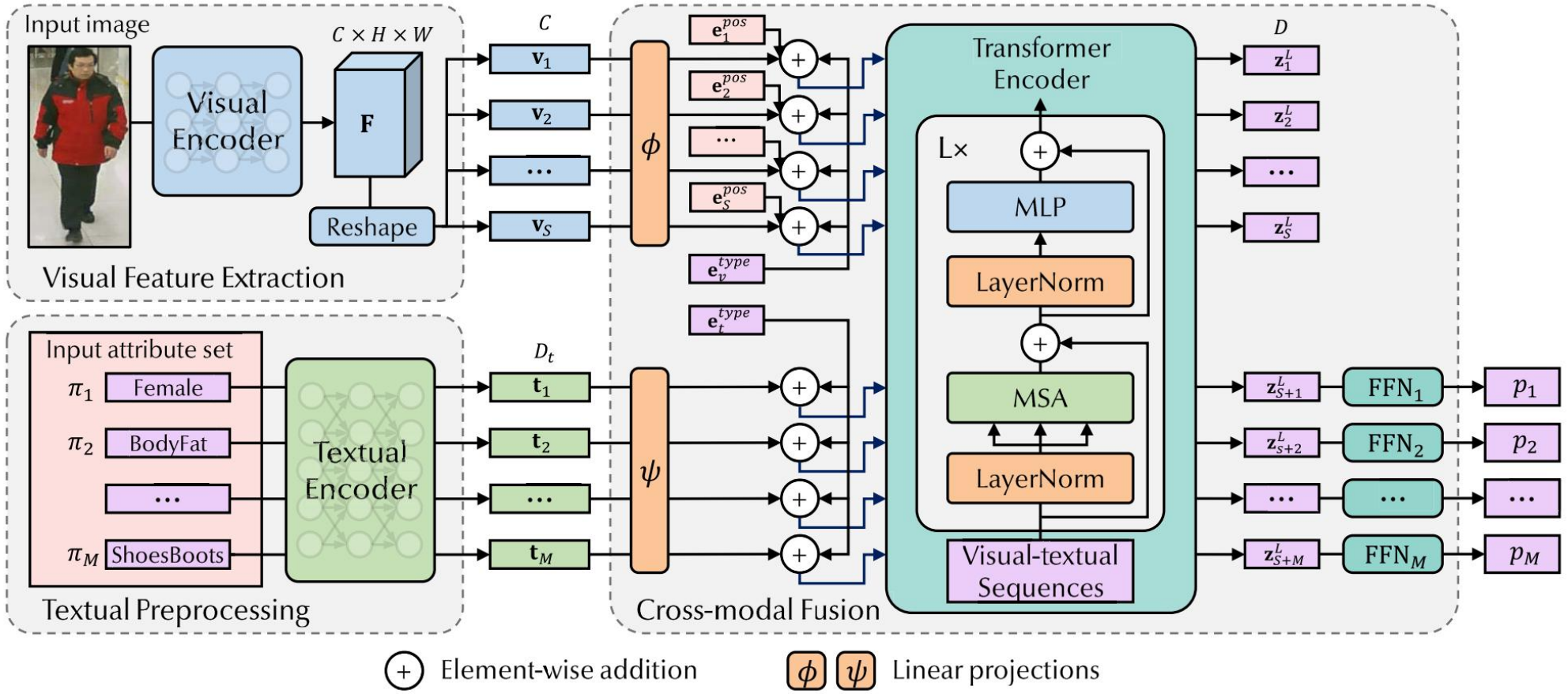
## ➢ Comparison between Different PAR Frameworks



# When Big Models Meet PAR



## ➢ Vision-Text Fusion for PAR




Cheng, Xinhua, et al. "A simple visual-textual baseline for pedestrian attribute recognition." *IEEE TCSVT* - 2022

# When Big Models Meet PAR



## ➢ Vision-Text Fusion for PAR

| Image | Label | Split | Expand |
|---|---|---|---|
|  | Age31-45 | Age 31 to 45 | A pedestrian whose age is between 31 and 45 |
| | BodyThin | Body thin | A pedestrian whose body is thin |
| | Customer | Customer | A pedestrian whose identity is customer |
| | Hs-BlackHair | Head black hair | A pedestrian whose head is black hair |
| | Ub-Jacket | Upper body jacket | A pedestrian whose upper body is wearing jacket |
| | Lb-LongTrousers | Lower body long trousers | A pedestrian whose lower body is wearing long trousers |

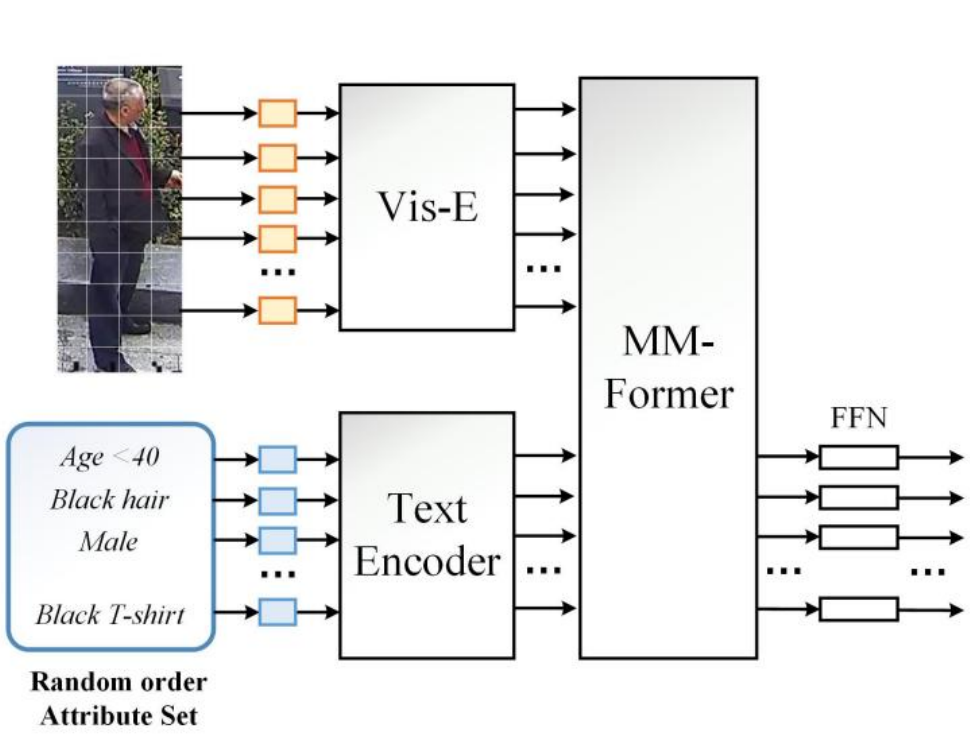
| | | RAP | | | | | PA100K | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Rethinking [46] | ResNet50 | 78.48 | 67.17 | 82.84 | 76.28 | 78.94 | 79.38 | 78.56 | **89.41** | 84.78 | 86.55 |
| MT-CAS [50] | ResNet34 | - | - | - | - | - | 77.20 | 78.09 | 88.46 | 84.86 | 86.62 |
| PD-Net [51] | Inception-V3 | - | - | - | - | - | 80.40 | 78.80 | 87.50 | 86.91 | 87.20 |
| JLAC [52] | ResNet50 | 83.69 | 69.15 | 79.31 | 82.40 | 80.82 | 82.31 | 79.47 | 87.45 | 87.77 | 87.61 |
| $SSC_{soft}$ [53] | ResNet50 | 82.83 | 68.16 | 74.74 | **87.54** | 80.27 | 81.70 | 78.85 | 85.80 | 88.92 | 86.89 |
| DAFL [54] | ResNet50 | **83.72** | 68.18 | 77.41 | 83.39 | 80.29 | 83.54 | 80.13 | 87.01 | 89.19 | 88.09 |
| JRL [9] | AlexNet | 74.74 | - | 75.08 | 74.96 | 74.62 | - | - | - | - | - |
| GRL [29] | Inception-V3 | 81.20 | - | 77.70 | 80.90 | 79.29 | - | - | - | - | - |
| RC [23] | Inception-V3 | 78.47 | - | 82.67 | 76.65 | 79.54 | - | - | - | - | - |
| RA [23] | Inception-V3 | 81.16 | - | 79.45 | 79.23 | 79.34 | - | - | - | - | - |
| VTB | ResNet50 | 81.43 | 69.21 | 78.22 | 83.99 | 80.63 | 81.02 | 79.52 | 86.89 | 88.53 | 87.31 |
| VTB | ViT-B/16 | 82.67 | **69.44** | 78.28 | 84.39 | **80.84** | **83.72** | **80.89** | 87.88 | **89.30** | **88.21** |

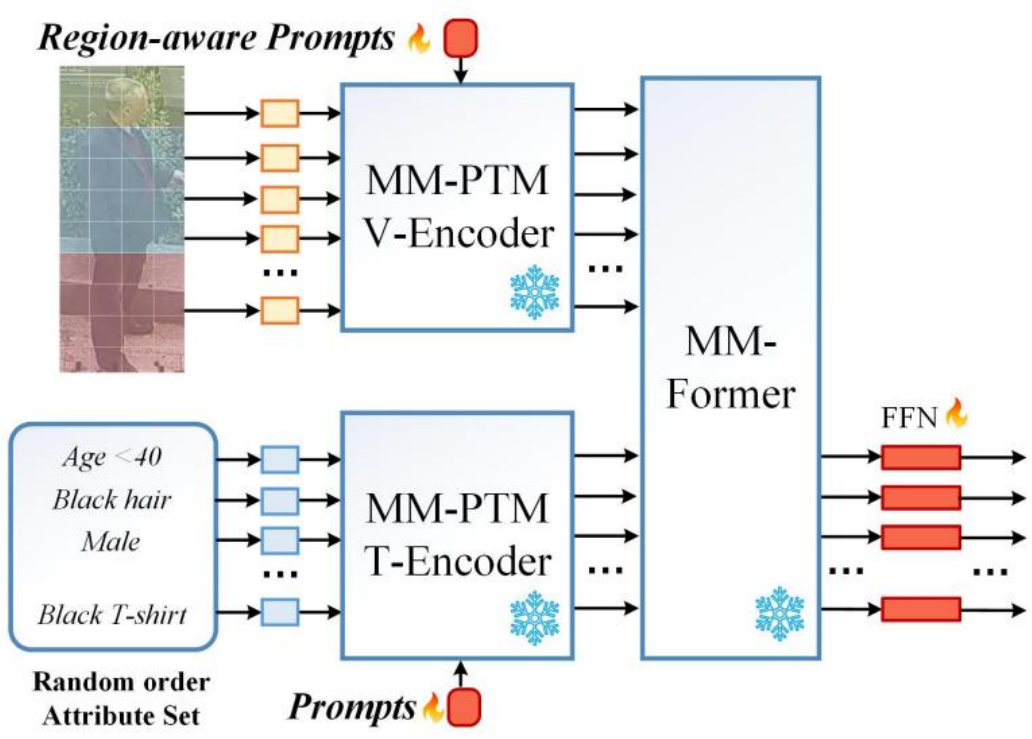
# When Big Models Meet PAR



## ➢ Motivation of Our Proposed PromptPAR

- *VTB adopts ViT and BERT for visual and text feature extraction independently, which weaking the relations between the dual modalities.*
- *It adopts full-parameter fine-tuning which may computational expensive.*

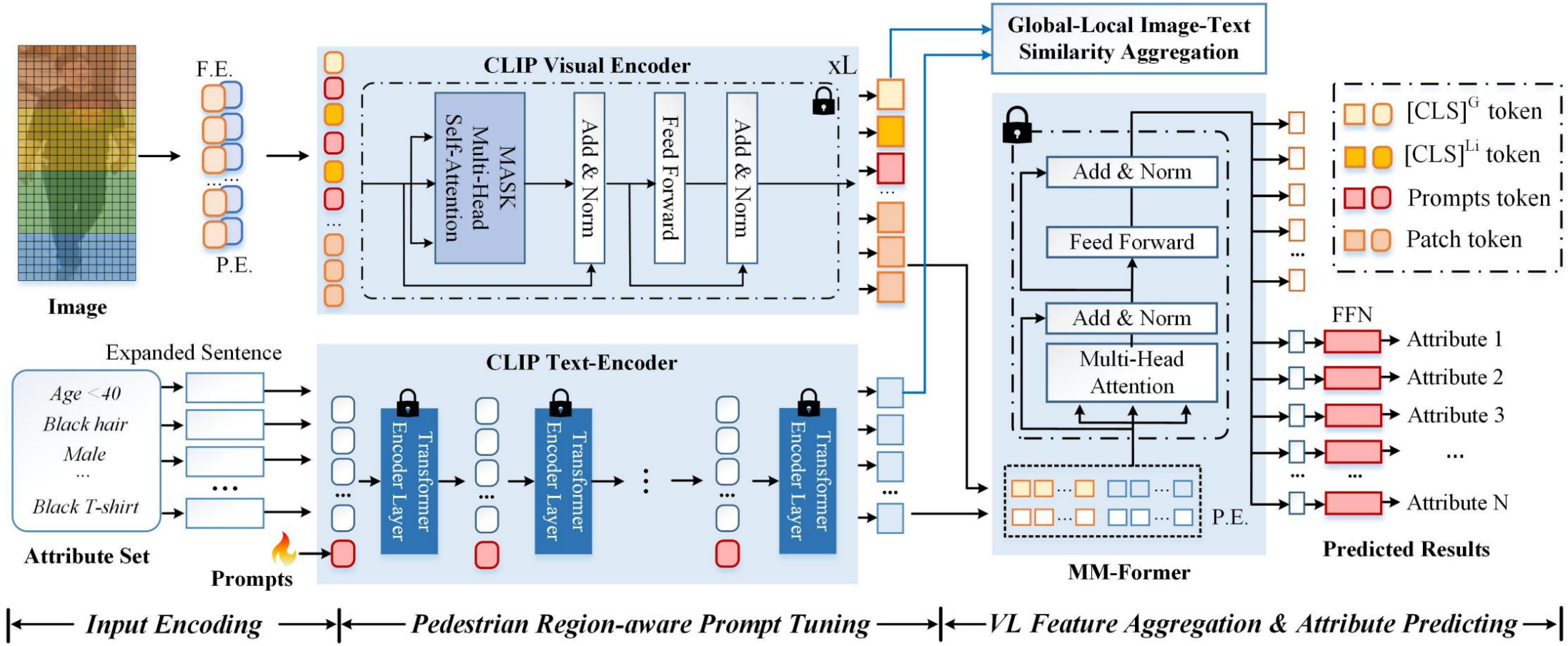


# When Big Models Meet PAR



## ➢ PromptPAR Framework



# When Big Models Meet PAR



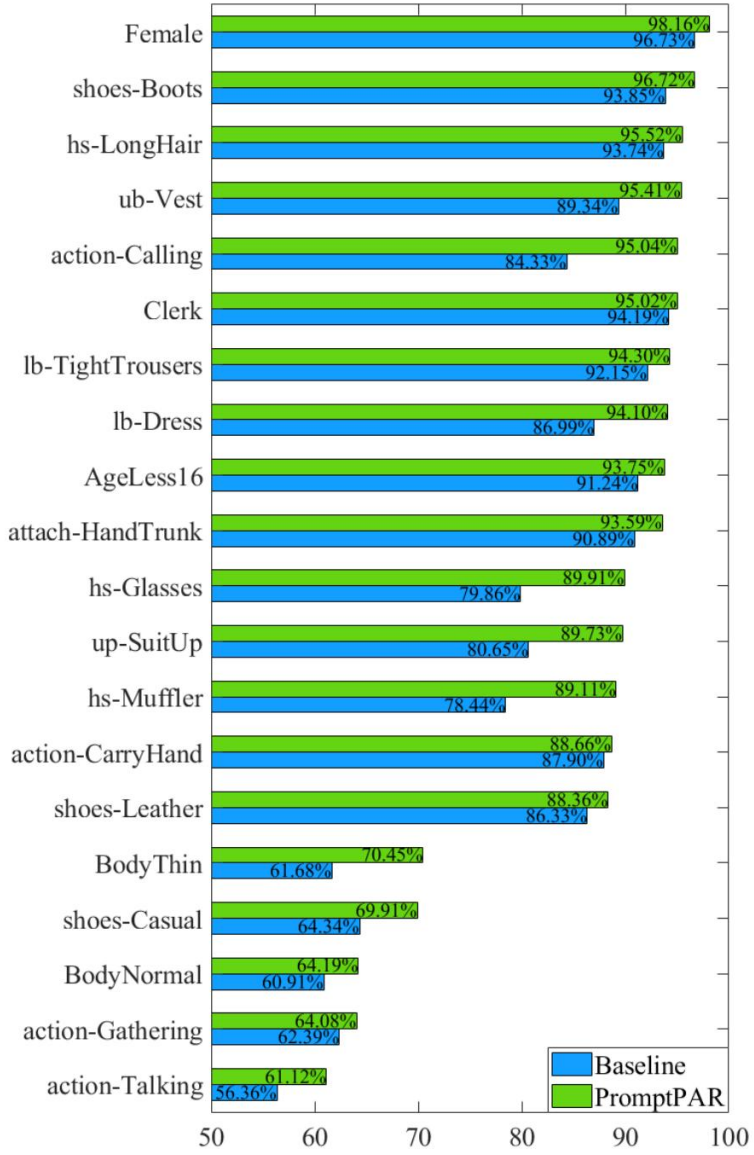
## ➢ Experimental Results

*No shared personal identity between the training and inference data.*

- PETA, PA100K, RAPv1, RAPv2, WIDER; PETA-ZS, RAP-ZS

Results on PETA and PA100K datasets

| Methods | Backbone | PETA | | | | | PA100K | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | mA | Acc | Prec | Recall | F1 | mA | Acc | Prec | Recall | F1 |
| DeepMAR (ACPR 2015) [60] | CaffeNet | 82.89 | 75.07 | 83.68 | 83.14 | 83.41 | 72.70 | 70.39 | 82.24 | 80.42 | 81.32 |
| HPNet (ICCV 2017) [28] | Inception | 81.77 | 76.13 | 84.92 | 83.24 | 84.07 | 74.21 | 72.19 | 82.97 | 82.09 | 82.53 |
| JRL (ICCV 2017) [10] | AlexNet | 82.13 | - | 82.55 | 82.12 | 82.02 | - | - | - | - | - |
| GRL (IJCAI 2018) [11] | Inception-V3 | 86.70 | - | 84.34 | 88.82 | 86.51 | - | - | - | - | - |
| MsVAA (ECCV 2018) [62] | ResNet101 | 84.59 | 78.56 | 86.79 | 86.12 | 86.46 | - | - | - | - | - |
| RA (AAAI 2019) [12] | Inception-V3 | 86.11 | - | 84.69 | 88.51 | 86.56 | - | - | - | - | - |
| VRKD (IJCAI 2019) [63] | ResNet50 | 84.90 | 80.95 | 88.37 | 87.47 | 87.91 | 77.87 | 78.49 | 88.42 | 86.08 | 87.24 |
| AAP (IJCAI 2019) [64] | ResNet50 | 86.97 | 79.95 | 87.58 | 87.73 | 87.65 | 80.56 | 78.30 | **89.49** | 84.36 | 86.85 |
| VAC (CVPR 2019) [65] | ResNet50 | - | - | - | - | - | 79.16 | 79.44 | 88.97 | 86.26 | 87.59 |
| ALM (ICCV 2019) [66] | BN-Inception | 86.30 | 79.52 | 85.65 | 88.09 | 86.85 | 80.68 | 77.08 | 84.24 | 88.84 | 86.46 |
| JLAC (AAAI 2020) [67] | ResNet50 | 86.96 | 80.38 | 87.81 | 87.09 | 87.50 | 82.31 | 79.47 | 87.45 | 87.77 | 87.61 |
| SCRL (TCSVT 2020) [68] | ResNet50 | 87.2 | - | **89.20** | 87.5 | 88.3 | 80.6 | - | 88.7 | 84.9 | 82.1 |
| SSCsoft (ICCV 2021) [69] | ResNet50 | 86.52 | 78.95 | 86.02 | 87.12 | 86.99 | 81.87 | 78.89 | 85.98 | 89.10 | 86.87 |
| IAA-Caps (PR 2022) [70] | OSNet | 85.27 | 78.04 | 86.08 | 85.80 | 85.64 | 81.94 | 80.31 | 88.36 | 88.01 | 87.80 |
| MCFL (NCA 2022) [71] | ResNet-50 | 86.83 | 78.89 | 84.57 | 88.84 | 86.65 | 81.53 | 77.80 | 85.11 | 88.20 | 86.62 |
| DRFormer (NC 2022) [72] | ViT-B/16 | **89.96** | **81.30** | 85.68 | **91.08** | **88.30** | 82.47 | 80.27 | 87.60 | 88.49 | 88.04 |
| VAC-Combine (IJCV 2022) [36] | ResNet50 | - | - | - | - | - | 82.19 | 80.66 | 88.72 | 88.10 | 88.41 |
| DAFL (AAAI 2022) [37] | ResNet50 | 87.07 | 78.88 | 85.78 | 87.03 | 86.40 | 83.54 | 80.13 | 87.01 | 89.19 | 88.09 |
| CGCN (TMM 2022) [73] | ResNet | 87.08 | 79.30 | 83.97 | 89.38 | 86.59 | - | - | - | - | - |
| CAS-SAL-FR (IJCV 2022) [74] | ResNet50 | 86.40 | 79.93 | 87.03 | 87.33 | 87.18 | 82.86 | 79.64 | 86.81 | 87.79 | 85.18 |
| VTB (TCSVT 2022) [19] | ViT-B/16 | 85.31 | 79.60 | 86.76 | 87.17 | 86.71 | 83.72 | 80.89 | 87.88 | 89.30 | 88.21 |
| VTB* (TCSVT 2022) [19] | ViT-L/14 | 86.34 | 79.59 | 86.66 | 87.82 | 86.97 | **85.30** | **81.76** | 87.87 | **90.67** | **88.86** |
| PromptPAR (Ours) | ViT-L/14 | **88.76** | **82.84** | **89.04** | **89.74** | **89.18** | **87.47** | **83.78** | **89.27** | **91.70** | **90.15** |



# When Big Models Meet PAR



## ➢ Experimental Results

Results on RAPv1 and RAPv2

| Methods | Backbone | RAPv1 | | | | | RAPv2 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | mA | Acc | Prec | Recall | F1 | mA | Acc | Prec | Recall | F1 |
| DeepMAR (ACPR 2015) [60] | CaffeNet | 73.79 | 62.02 | 74.92 | 76.21 | 75.56 | - | - | - | - | - |
| HPNet (ICCV 2017) [28] | Inception | 76.12 | 65.39 | 77.33 | 78.79 | 78.05 | - | - | - | - | - |
| JRL (ICCV 2017) [10] | AlexNet | 74.74 | - | 75.08 | 74.96 | 74.62 | - | - | - | - | - |
| GRL (IJCAI 2018) [11] | Inception-V3 | 81.20 | - | 77.70 | 80.90 | 79.29 | - | - | - | - | - |
| MsVAA (ECCV 2018) [62] | ResNet101 | - | - | - | - | - | 78.34 | 65.57 | **77.37** | 79.17 | 78.26 |
| RA (AAAI 2019) [12] | Inception-V3 | 81.16 | - | 79.45 | 79.23 | 79.34 | - | - | - | - | - |
| VRKD (IJCAI 2019) [63] | ResNet50 | 78.30 | 69.79 | **82.13** | 80.35 | 81.23 | - | - | - | - | - |
| AAP (IJCAI 2019) [64] | ResNet50 | 81.42 | 68.37 | 81.04 | 80.27 | 80.65 | - | - | - | - | - |
| VAC (CVPR 2019) [65] | ResNet50 | - | - | - | - | - | 79.23 | 64.51 | 75.77 | 79.43 | 77.10 |
| ALM (ICCV 2019) [66] | BN-Inception | 81.87 | 68.17 | 74.71 | 86.48 | 80.16 | 79.79 | 64.79 | 73.93 | 82.03 | 77.77 |
| JLAC (AAAI 2020) [67] | ResNet50 | 83.69 | 69.15 | 79.31 | 82.40 | 80.82 | 79.23 | 64.42 | 75.69 | 79.18 | 77.40 |
| SSCsoft (ICCV 2021) [69] | ResNet50 | 82.77 | 68.37 | 75.05 | 87.49 | 80.43 | - | - | - | - | - |
| IAA-Caps (PR 2022) [70] | OSNet | 81.72 | 68.47 | 79.56 | 82.06 | 80.37 | - | - | - | - | - |
| MCFL (NCA 2022) [71] | ResNet50 | 84.04 | 67.28 | 73.44 | **87.75** | 79.96 | - | - | - | - | - |
| DRFormer (NC 2022) [72] | ViT-B/16 | 81.81 | **70.60** | 80.12 | 82.77 | 81.42 | - | - | - | - | - |
| VAC-Combine (IJCV 2022) [36] | ResNet50 | 81.30 | 70.12 | **81.56** | 81.51 | **81.54** | - | - | - | - | - |
| DAFL (AAAI 2022) [37] | ResNet50 | 83.72 | 68.18 | 77.41 | 83.39 | 80.29 | 81.04 | 66.70 | 76.39 | 82.07 | 79.13 |
| CGCN (TMM 2022) [73] | ResNet50 | **84.70** | 54.40 | 60.03 | 83.68 | 70.49 | - | - | - | - | - |
| CAS-SAL-FR (IJCV 2022) [74] | ResNet50 | 84.18 | 68.59 | 77.56 | 83.81 | 80.56 | - | - | - | - | - |
| VTB (TCSVT 2022) [19] | ViT-B/16 | 82.67 | 69.44 | 78.28 | 84.39 | 80.84 | 81.34 | 67.48 | 76.41 | 83.32 | 79.35 |
| PARformer (TCSVT 2023) [76] | Swin-L | 84.13 | 69.94 | 79.63 | **88.19** | 81.35 | - | - | - | - | - |
| VTB* (TCSVT 2022) [19] | ViT-L/14 | 83.69 | 69.78 | 78.09 | 85.21 | 81.10 | **81.36** | **67.58** | 76.19 | **84.00** | **79.52** |
| PromptPAR (Ours) | ViT-L/14 | **85.45** | **71.61** | 79.64 | 86.05 | **82.38** | **83.14** | **69.62** | **77.42** | **85.73** | **81.00** |

# When Big Models Meet PAR



## ➢ Experimental Results

COMPONENT ANALYSIS ON THE RAP-V1 DATASET. MA, ACC, AND F1 RESULTS ARE REPORTED.

| No. | ViT | PTM | FTune | PTune(V) | PTune(T) | RegionPTune | RAPV1 | | | PETA | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | mA | Acc | F1 | mA | Acc | F1 |
| 1 | ✓ | | ✓ | | | | 82.79 | 68.95 | 80.43 | 85.31 | 79.60 | 86.71 |
| 2 | | ✓ | | ✓ | | | 85.22 | 71.12 | 82.12 | 88.54 | 82.25 | 88.77 |
| 3 | | ✓ | | ✓ | ✓ | | 85.37 | 71.27 | 82.20 | 88.54 | 82.46 | 88.94 |
| 4 | | ✓ | | ✓ | ✓ | ✓ | 85.45 | 71.61 | 82.38 | 88.76 | 82.82 | 89.18 |

COMPARISON WITH STATE-OF-THE-ART METHODS ON PETA-ZS AND RAP-ZS DATASETS.

| Methods | Backbone | PETA-ZS | | | | | RAP-ZS | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | mA | Acc | Prec | Recall | F1 | mA | Acc | Prec | Recall | F1 |
| MsVAA (ECCV 2018) [62] | ResNet101 | 71.53 | 58.67 | 74.65 | 69.42 | 71.94 | 72.04 | 62.13 | 75.67 | 75.81 | 75.74 |
| VAC (CVPR 2019) [65] | ResNet50 | 71.91 | 57.72 | 72.05 | 70.64 | 70.90 | 73.70 | 63.25 | 76.23 | 76.97 | 76.12 |
| ALM (ICCV 2019) [66] | BN-Inception | 73.01 | 57.78 | 69.50 | 73.69 | 71.53 | 74.28 | 63.22 | 72.96 | 80.73 | 76.65 |
| JLAC (AAAI 2020) [67] | ResNet50 | 73.60 | 58.66 | 71.70 | 72.41 | 72.05 | 76.38 | 62.58 | 73.14 | 79.20 | 76.05 |
| Jia et al. (Arxiv 2021) [61] | ResNet50 | 71.62 | 58.19 | 73.09 | 70.33 | 71.68 | 72.32 | 63.61 | **76.88** | 76.62 | 76.75 |
| MCFL (NCA 2022) [71] | ResNet50 | 72.91 | 57.04 | 68.47 | 74.35 | 71.29 | 74.37 | 63.37 | 71.21 | 83.86 | 77.02 |
| VTB (TCSVT 2022) [19] | ViT-B/16 | 75.13 | 60.50 | 73.29 | 74.40 | 73.38 | 75.76 | 64.73 | 74.93 | 80.85 | 77.35 |
| VTB* (TCSVT 2022) [19] | ViT-L/14 | **77.18** | **63.12** | **74.77** | **77.24** | **75.50** | **79.17** | **68.34** | 76.81 | **84.51** | **80.07** |
| PromptPAR (Ours) | ViT-L/14 | **80.08** | **66.02** | **76.53** | **80.49** | **77.77** | **80.43** | **70.39** | **78.48** | **85.57** | **81.52** |

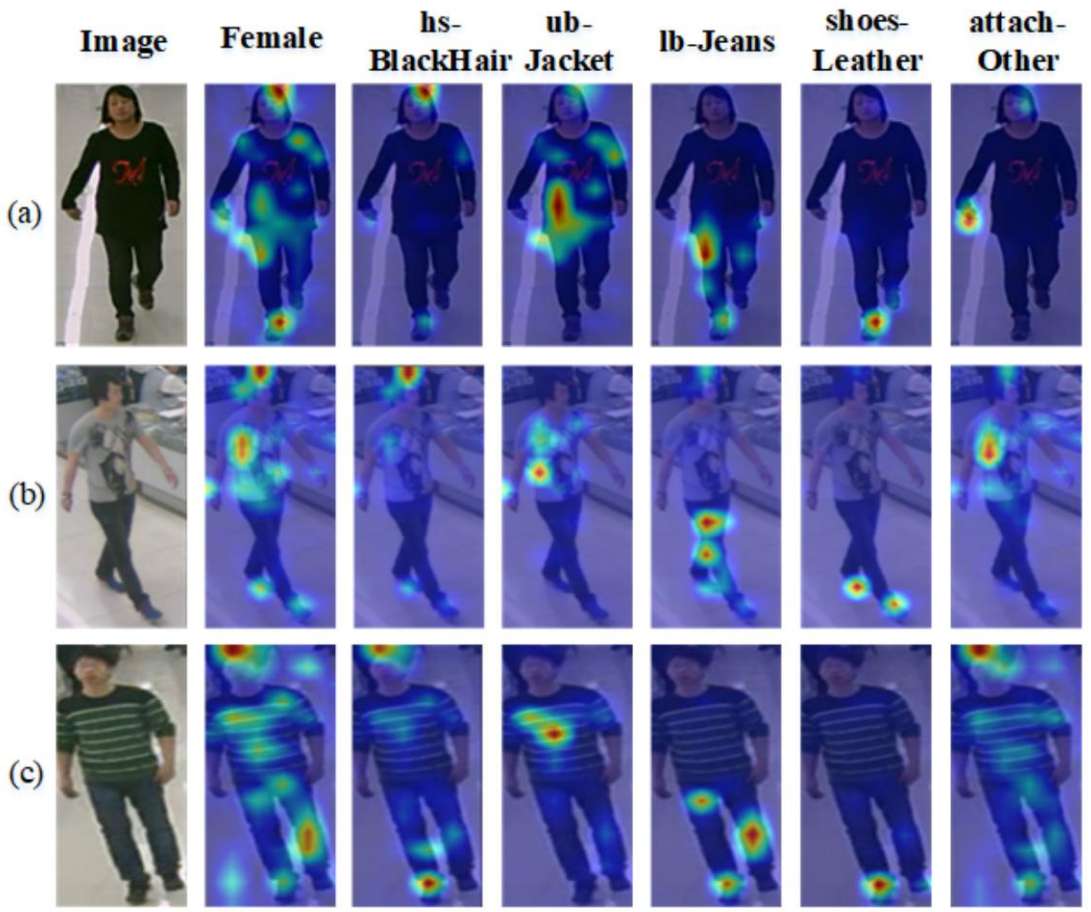
# When Big Models Meet PAR

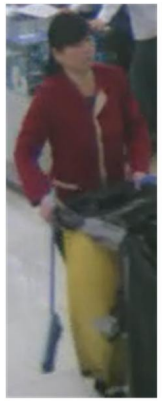


## ➢ Experimental Results

**Accepted by IEEE TCSVT 2024**

**Pedestrian Attribute Recognition via CLIP based Prompt Vision-Language Fusion**
Xiao Wang, Jiandong Jin, Chenglong Li, Jin Tang, Cheng Zhang, Wei Wang



Female
Age31-45
BodyNormal
Clerkhs-BaldHead
hs-Hat
ub-Sweater
attach-PaperBag
attach-Other
action-Calling
action-Pulling

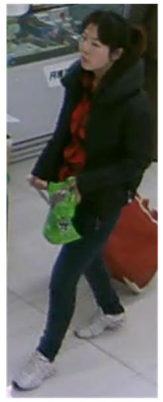
Age31-45
BodyNormal
Customer
hs-Hat
ub-SuitUp
lb-Skirt
shoes-Boots
attach-Backpack

Age31-45
BodyFat
Customer
hs-Hat
ub-Jacket
lb-Skirt
shoes-Sport
attach-Box
attach-HandTrunk
action-Calling

Age17-30
Age31-45
BodyNormal
Customer
hs-Hat
ub-Sweater
ub-SuitUp
lb-TightTrousers
shoes-Sport

Age31-45
BodyFat
Customer
hs-Hat
ub-Jacket
lb-Skirt
shoes-Boots
attach-HandBag

Female
Age31-45
BodyNormal
Customer
hs-Hat
ub-Jacket
lb-ShortSkirt
lb-Jeans
lb-TightTrousers
shoes-Leather
attach-Backpack



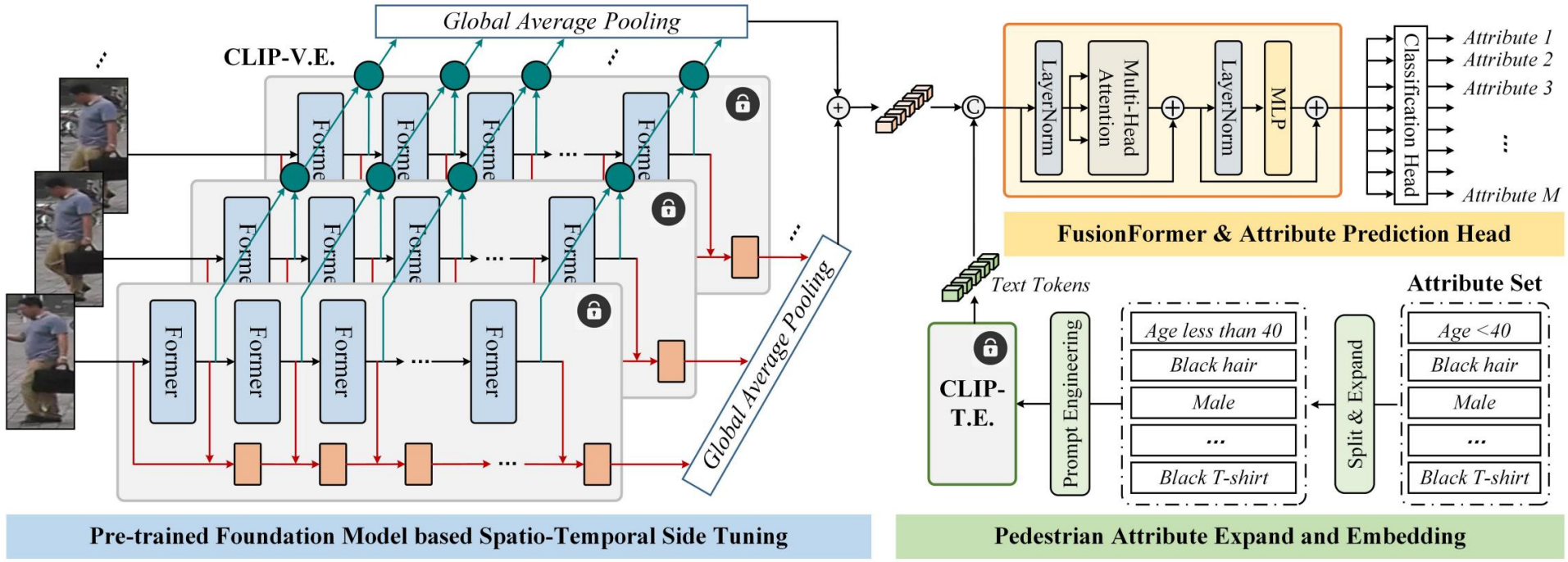
Female
Age17-30
Age31-45
BodyNormal
Customer
hs-BlackHair
hs-Hat
ub-Shirt
ub-Vest
ub-Jacket
ub-SuitUp
lb-TightTrousers
shoes-Boots
action-Calling

Age31-45
BodyNormal
Customer
hs-Hat
ub-Sweater
lb-Skirt
shoes-Sport

# When Big Models Meet PAR



## ➢ Video-based PAR

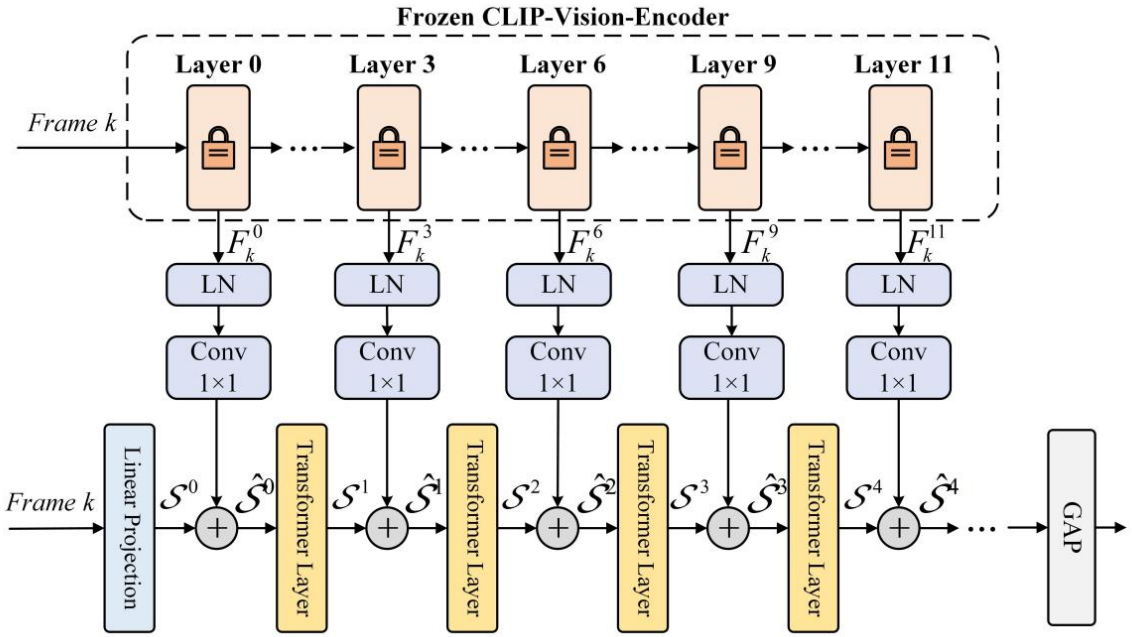

*Spatio-Temporal Side Tuning*


- Jun Zhu, et al. "Learning clip guided visual-text fusion transformer for video-based pedestrian attribute recognition." *CVPR*-2023 Workshop,
- Xiao Wang, et al. "Spatio-Temporal Side Tuning Pre-trained Foundation Models for Video-based Pedestrian Attribute Recognition." *arXiv preprint arXiv:2404.17929* (2024).

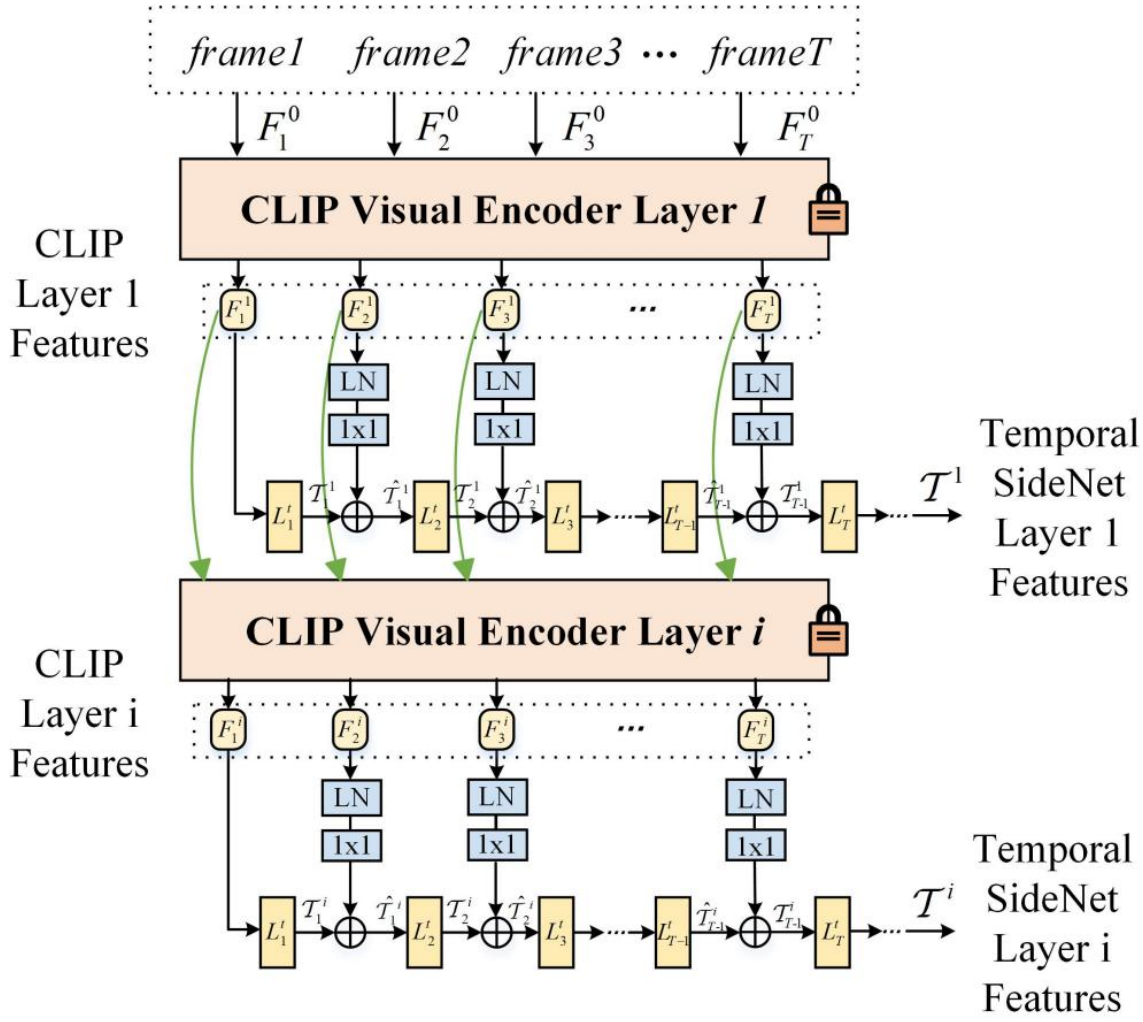
# When Big Models Meet PAR



## ➢ Video-based PAR




Spatial Side Network (SSN)




Temporal Side Network (TSN)

# When Big Models Meet PAR



## ➢ Experimental Results

- MARS-Attribute dataset, DukeMTMC-VID-Attribute dataset

RESULTS ON MARS-ATTRIBUTE AND DUKEMTMC-VID-ATTRIBUTE VIDEO-BASED PAR DATASET.

| Methods | Backbone | MARS-Attribute Dataset | | | | DukeMTMC-VID-Attribute Dataset | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Accuracy | Precision | Recall | F1 score | Accuracy | Precision | Recall | F1 score |
| 3DCNN [61] | - | 81.95 | - | - | 61.87 | 84.24 | - | - | 62.93 |
| CNN-RNN [62] | - | 86.35 | - | - | 70.42 | 88.84 | - | - | 71.63 |
| ALM [63] | BN-Inception | 86.56 | - | - | 68.89 | 88.13 | - | - | 69.66 |
| $SSC_{soft}$ [64] | ResNet50 | 86.00 | - | - | 68.15 | 87.52 | - | - | 68.71 |
| TA(Image) [16] | ResNet50 | 85.85 | - | - | 67.28 | 87.77 | - | - | 68.70 |
| TA(Video) [16] | ResNet50 | 87.01 | - | - | 72.04 | 89.31 | - | - | 73.24 |
| Lee et al. [27] | ResNet50 | 86.75 | - | - | 70.42 | 88.98 | - | - | 72.30 |
| TRA [65] | ResNet50 | 87.05 | - | - | 71.92 | 89.32 | - | - | 75.01 |
| VTB [2] | ViT-B/16 | 90.37 | 78.96 | 78.42 | 78.32 | 90.29 | 73.38 | 76.99 | 74.81 |
| VTF [46] | ViT-B/16 | **92.47** | **81.76** | **82.95** | **81.94** | **92.45** | **77.23** | **81.44** | **78.83** |
| Ours | ViT-B/16 | **93.19** | **82.27** | **84.87** | **83.22** | **93.31** | **78.19** | **83.18** | **80.45** |
| Improvements | - | +0.72 | +0.51 | +1.92 | +1.28 | +0.86 | +0.60 | +0.13 | +0.53 |

# When Big Models Meet PAR



## ➢ Experimental Results

RESULTS ON MARS VIDEO-BASED PAR DATASET. ACCURACY AND F1-SCORE ARE REPORTED FOR ALL THE ASSESSED ATTRIBUTES.

| Attribute | TA(Image) | | 3DCNN | | CNN-RNN | | TA(Video) | | TRA | | VTF | | VTFPAR++ | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Acc | F1 | Acc | F1 | Acc | F1 | Acc | F1 | Acc | F1 | Acc | F1 | Acc | F1 |
| top length | 94.21 | 58.72 | 93.63 | 56.37 | 94.60 | 65.18 | 94.47 | 71.61 | **95.12** | 69.09 | **94.62** | **97.26** | 94.57 | **96.76** |
| bottom type | 93.12 | 81.69 | 89.19 | 72.86 | 93.67 | 84.16 | **94.60** | 86.62 | 93.73 | 85.06 | 92.97 | **97.21** | **94.29** | **97.22** |
| shoulder bag | 80.39 | 72.57 | 71.82 | 61.30 | 82.70 | 75.89 | **83.48** | **76.08** | 82.90 | **76.83** | 80.26 | 65.61 | **84.12** | 70.97 |
| backpack | 89.37 | 85.95 | 82.60 | 76.58 | 90.18 | **87.17** | 90.59 | **87.62** | **90.63** | 86.90 | **91.32** | 82.08 | 90.56 | 83.13 |
| hat | 96.91 | 57.57 | 96.53 | 57.69 | 97.90 | **77.74** | 97.51 | **77.84** | 97.84 | 77.70 | **98.32** | 72.76 | **98.42** | 65.66 |
| handbag | 85.71 | 62.82 | 83.88 | 59.90 | 88.07 | **71.68** | 87.61 | **73.55** | 87.03 | 70.98 | **87.63** | 59.08 | **89.64** | 66.61 |
| hair | 88.61 | 86.91 | 85.12 | 82.77 | 88.78 | **87.11** | 89.54 | **88.17** | 89.54 | 88.23 | **90.51** | 86.37 | **90.56** | 86.49 |
| gender | 91.32 | 90.89 | 86.49 | 85.75 | 92.77 | 92.44 | 92.83 | 92.50 | 92.89 | 92.53 | **94.56** | **92.88** | **94.87** | **93.68** |
| bottom length | 92.69 | 92.29 | 89.96 | 89.35 | 93.70 | 93.33 | 94.22 | **93.90** | 93.41 | 93.06 | **95.68** | 93.69 | **95.21** | **94.99** |
| pose | 72.03 | 56.91 | 62.51 | 47.69 | 72.40 | 58.36 | 73.65 | 61.36 | 72.50 | 56.86 | **90.32** | **74.84** | **92.10** | **76.27** |
| motion | 91.08 | 39.39 | 90.34 | 33.64 | 92.12 | 43.92 | 92.12 | 43.69 | 92.62 | 45.27 | **97.33** | **93.50** | **98.05** | **95.03** |
| top color | 74.73 | 72.72 | 68.04 | 65.63 | 71.90 | 69.28 | 73.43 | 71.44 | 74.41 | 72.34 | **93.85** | **74.97** | **94.67** | **78.08** |
| bottom color | 68.27 | 44.63 | 65.44 | 40.39 | 65.77 | 39.68 | 69.45 | 43.98 | 71.10 | 48.46 | **93.66** | **69.76** | **94.38** | **74.01** |
| age | 83.44 | 38.87 | 81.70 | 36.22 | 84.28 | 39.93 | 84.71 | 40.21 | 84.92 | 43.53 | **93.50** | **87.07** | **93.17** | **86.09** |
| **Average** | 85.85 | 67.28 | 81.95 | 61.87 | 86.35 | 70.42 | 87.01 | 72.04 | 87.05 | 71.92 | **92.47** | **81.94** | **93.19** | **83.22** |

# When Big Models Meet PAR

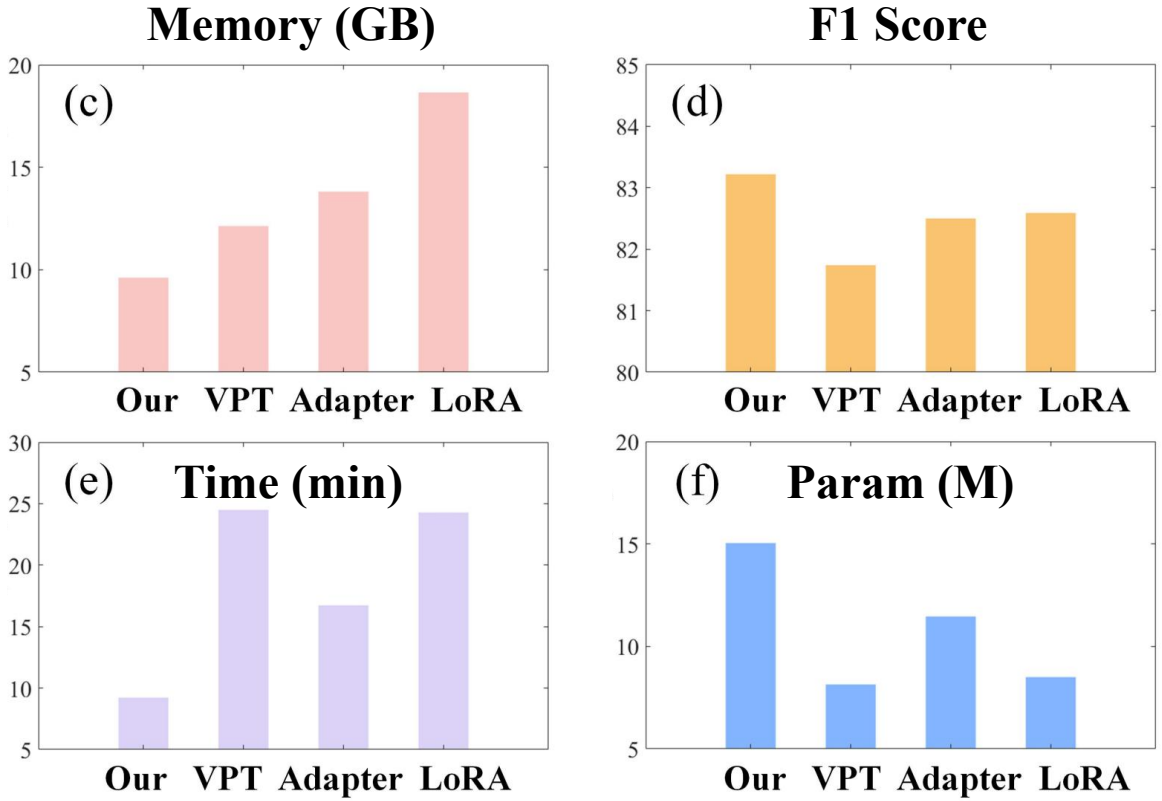


## ➢ Experimental Results

| NO. | FFN | VTFormer | SSN | TSN | Acc | F1 | Params(M) |
|---|---|---|---|---|---|---|---|
| 1 | ✓ | | | | 89.74 | 78.69 | 150.45 |
| 2 | | ✓ | | | 92.47 | 81.94 | 157.53 |
| 3 | | ✓ | ✓ | | 92.90 | 82.84 | 7.85 |
| 4 | | ✓ | | ✓ | 92.99 | 82.90 | 8.32 |
| 5 | | ✓ | ✓ | ✓ | 93.09 | 83.22 | 15.04 |

COMPARISON WITH OTHER PEFT STRATEGIES ON MARS DATASET.

| Method | Precision | Recall | F1 score |
|---|---|---|---|
| LoRA | 81.61 | 84.30 | 82.59 |
| Prompt-Tuning | 82.03 | 82.66 | 81.74 |
| Adapter-Tuning | 82.46 | 83.65 | 82.50 |
| Ours | 82.27 | 84.87 | 83.22 |



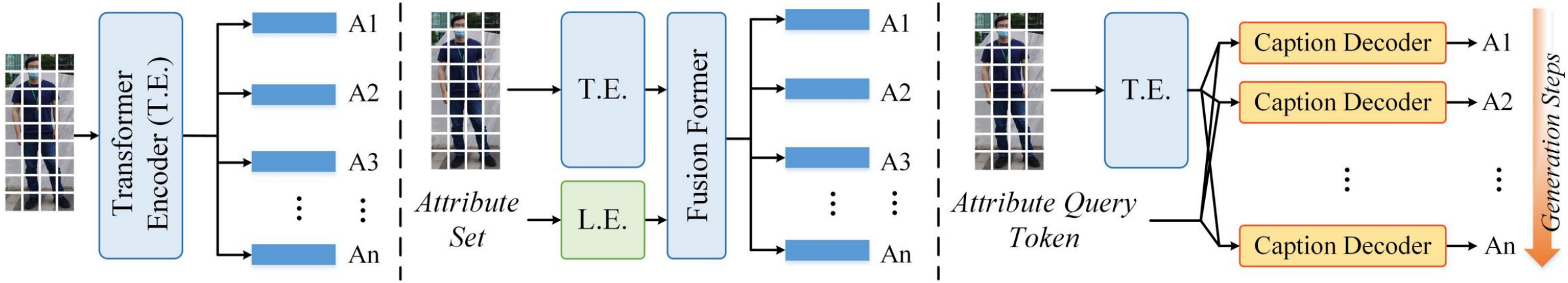

(c-f). Comparison between existing PEFT methods and ours

# When Big Models Meet PAR



## ➢ SequencePAR

Jin, Jiandong, et al. "**Sequencepar: Understanding pedestrian attributes via a sequence generation paradigm.**" *arXiv preprint arXiv:2312.01640* (2023).




(a). Backbone feature extraction using Former for PAR

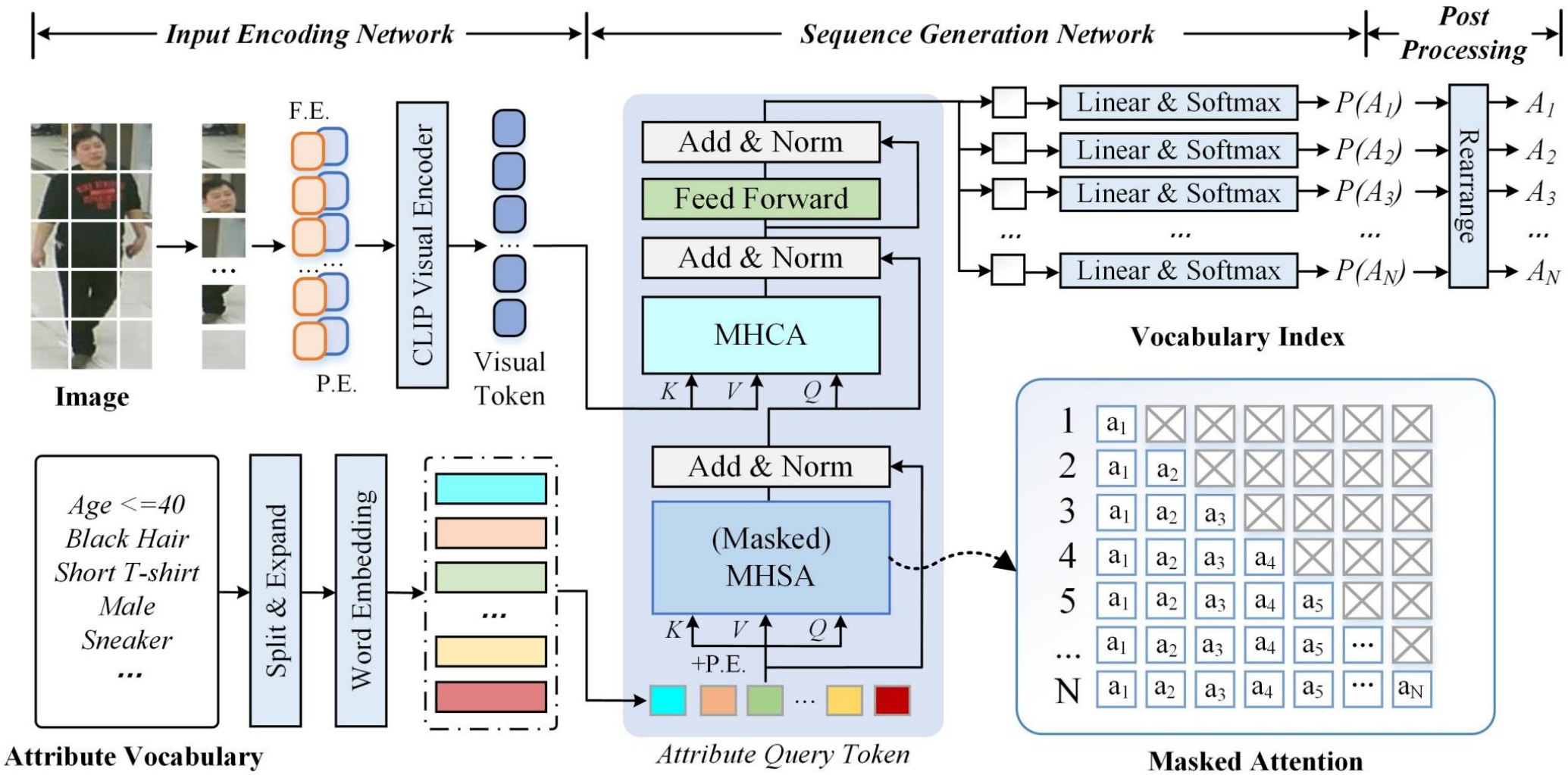
(b). Image-text fusion using Former for PAR

(c). Transformer encoder-decoder paradigm for attribute generation (Ours)

# When Big Models Meet PAR



## ➢ SequencePAR



# When Big Models Meet PAR



## ➢ SequencePAR

COMPARISON OF DIFFERENT DECODING LAYERS FOR PAR ON PETA DATASET. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**.

| Decoder Layers | PETA | | | |
|---|---|---|---|---|
| | Accuracy | Precision | Recall | F1 |
| 1 | 83.96 | 89.90 | 89.95 | 89.80 |
| 3 | 83.84 | 89.82 | 89.78 | 89.68 |
| **6** | **84.92** | **90.44** | **90.73** | **90.46** |
| 9 | 83.72 | 89.67 | 89.78 | 89.60 |
| 12 | 83.90 | 89.93 | 89.84 | 89.76 |

COMPARE THE GREEDY SEARCH AND BEAM SEARCH WITH DIFFERENT BEAM WIDTHS. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**.

| Beam Width | PETA | | | |
|---|---|---|---|---|
| | Accuracy | Precision | Recall | F1 |
| **1** | **84.92** | **90.44** | 90.73 | **90.46** |
| 3 | 84.90 | 90.33 | 90.79 | 90.45 |
| 5 | 84.89 | 90.36 | 90.76 | 90.44 |
| 10 | 84.84 | 90.27 | **90.80** | 90.41 |

COMPARISON WITH STATE-OF-THE-ART METHODS ON PETA-ZS AND RAP-ZS DATASETS.

| Methods | Ref | Backbone | PETA-ZS | | | | RAP-ZS | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Accuracy | Precision | Recall | F1 | Accuracy | Precision | Recall | F1 |
| MsVAA [56] | ECCV 2018 | ResNet101 | 58.67 | 74.65 | 69.42 | 71.94 | 62.13 | 75.67 | 75.81 | 75.74 |
| VAC [60] | CVPR 2019 | ResNet50 | 57.72 | 72.05 | 70.64 | 70.90 | 63.25 | 76.23 | 76.97 | 76.12 |
| ALM [61] | ICCV 2019 | BN-Inception | 57.78 | 69.50 | 73.69 | 71.53 | 63.22 | 72.96 | 80.73 | 76.65 |
| JLAC [14] | AAAI 2020 | ResNet50 | 58.66 | 71.70 | 72.41 | 72.05 | 62.58 | 73.14 | 79.20 | 76.05 |
| Jia et al. [53] | - | ResNet50 | 58.19 | 73.09 | 70.33 | 71.68 | 63.61 | 76.88 | 76.62 | 76.75 |
| MCFL [64] | NCA 2022 | ResNet50 | 57.04 | 68.47 | 74.35 | 71.29 | 63.37 | 71.21 | **83.86** | 77.02 |
| VTB [18] | TCSVT 2022 | ViT-B/16 | 60.50 | 73.29 | 74.40 | 73.38 | 64.73 | 74.93 | 80.85 | 77.35 |
| VTB* [18] | TCSVT 2022 | ViT-L/14 | **63.12** | **74.77** | **77.24** | **75.50** | **68.34** | **76.81** | **84.51** | **80.07** |
| SequencePAR | - | ViT-L/14 | **66.70** | **78.75** | **78.52** | **78.40** | **70.28** | **82.13** | 80.55 | **81.14** |

# When Big Models Meet PAR



## ➢ SequencePAR

Experimental results on PETA and PA100K datasets

| Methods | Ref | Backbone | PETA | | | | PA100K | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Accuracy | Precision | Recall | F1 | Accuracy | Precision | Recall | F1 |
| DeepMAR [54] | ACPR 2015 | CaffeNet | 75.07 | 83.68 | 83.14 | 83.41 | 70.39 | 82.24 | 80.42 | 81.32 |
| HPNet [29] | ICCV 2017 | Inception | 76.13 | 84.92 | 83.24 | 84.07 | 72.19 | 82.97 | 82.09 | 82.53 |
| JRL [15] | ICCV 2017 | AlexNet | - | 82.55 | 82.12 | 82.02 | - | - | - | - |
| GRL [55] | IJCAI 2018 | Inception-V3 | - | 84.34 | 88.82 | 86.51 | - | - | - | - |
| MsVAA [56] | ECCV 2018 | ResNet101 | 78.56 | 86.79 | 86.12 | 86.46 | - | - | - | - |
| RA [57] | AAAI 2019 | Inception-V3 | - | 84.69 | 88.51 | 86.56 | - | - | - | - |
| VRKD [58] | IJCAI 2019 | ResNet50 | 80.95 | 88.37 | 87.47 | 87.91 | 78.49 | 88.42 | 86.08 | 87.24 |
| AAP [59] | IJCAI 2019 | ResNet50 | 79.95 | 87.58 | 87.73 | 87.65 | 78.30 | **89.49** | 84.36 | 86.85 |
| VAC [60] | CVPR 2019 | ResNet50 | - | - | - | - | 79.44 | 88.97 | 86.26 | 87.59 |
| ALM [61] | ICCV 2019 | BN-Inception | 79.52 | 85.65 | 88.09 | 86.85 | 77.08 | 84.24 | 88.84 | 86.46 |
| JLAC [14] | AAAI 2020 | ResNet50 | 80.38 | 87.81 | 87.09 | 87.50 | 79.47 | 87.45 | 87.77 | 87.61 |
| SSCsoft [62] | ICCV 2021 | ResNet50 | 78.95 | 86.02 | 87.12 | 86.99 | 78.89 | 85.98 | 89.10 | 86.87 |
| IAA-Caps [63] | PR 2022 | OSNet | 78.04 | 86.08 | 85.80 | 85.64 | 80.31 | 88.36 | 88.01 | 87.80 |
| MCFL [64] | NCA 2022 | ResNet50 | 78.89 | 84.57 | 88.84 | 86.65 | 77.80 | 85.11 | 88.20 | 86.62 |
| DRFormer [22] | NC 2022 | ViT-B/16 | 81.30 | 85.68 | **91.08** | 88.30 | 80.27 | 87.60 | 88.49 | 88.04 |
| VAC-Combine [65] | IJCV 2022 | ResNet50 | - | - | - | - | 80.66 | 88.72 | 88.10 | 88.41 |
| DAFL [66] | AAAI 2022 | ResNet50 | 78.88 | 85.78 | 87.03 | 86.40 | 80.13 | 87.01 | 89.19 | 88.09 |
| CGCN [13] | TMM 2022 | ResNet50 | 79.30 | 83.97 | 89.38 | 86.59 | - | - | - | - |
| CAS-SAL-FR [67] | IJCV 2022 | ResNet50 | 79.93 | 87.03 | 87.33 | 87.18 | 79.64 | 86.81 | 87.79 | 85.18 |
| VTB [18] | TCSVT 2022 | ViT-B/16 | 79.60 | 86.76 | 87.17 | 86.71 | 80.89 | 87.88 | 89.30 | 88.21 |
| FEMDAR [68] | SPIC 2023 | ResNet50 | 78.45 | 86.79 | 85.69 | 85.90 | 79.65 | 87.99 | 87.45 | 87.32 |
| EALC [69] | NC 2023 | EfficientNet-B4 | 81.71 | **88.58** | 88.23 | 88.40 | 80.27 | 87.32 | 88.98 | 88.14 |
| APTM [8] | ACM MM 2023 | Swin-B | - | - | - | - | 80.17 | 88.31 | 87.84 | 88.07 |
| PARFormer-L [21] | TCSVT 2023 | Swin-L | 82.86 | 88.06 | **91.98** | **89.06** | 81.13 | 88.09 | **91.67** | 88.52 |
| DFDT [70] | EAAI 2023 | Swin-B | 81.17 | 87.44 | 88.96 | 88.19 | 81.24 | 88.02 | 89.48 | 88.74 |
| OAGCN [71] | TMM 2023 | Swin-B | **82.95** | 88.26 | 89.10 | 88.68 | 80.38 | 84.55 | 90.42 | 87.39 |
| VTB* [18] | TCSVT 2022 | ViT-L/14 | 79.59 | 86.66 | 87.82 | 86.97 | **81.76** | 87.87 | **90.67** | **88.86** |
| SequencePAR | - | ViT-L/14 | **84.92** | **90.44** | 90.73 | **90.46** | **83.94** | **90.38** | 90.23 | **90.10** |

# When Big Models Meet PAR



## ➢ SequencePAR

Experimental results on RAPv1 and RAPv2 datasets

| Methods | Ref | Backbone | RAPv1 | | | | RAPv2 | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Accuracy | Precision | Recall | F1 | Accuracy | Precision | Recall | F1 |
| DeepMAR [54] | ACPR 2015 | CaffeNet | 62.02 | 74.92 | 76.21 | 75.56 | - | - | - | - |
| HPNet [29] | ICCV 2017 | Inception | 65.39 | 77.33 | 78.79 | 78.05 | - | - | - | - |
| JRL [15] | ICCV 2017 | AlexNet | - | 75.08 | 74.96 | 74.62 | - | - | - | - |
| GRL [55] | IJCAI 2018 | Inception-V3 | - | 77.70 | 80.90 | 79.29 | - | - | - | - |
| MsVAA [56] | ECCV 2018 | ResNet101 | - | - | - | - | 65.57 | 77.37 | 79.17 | 78.26 |
| RA [57] | AAAI 2019 | Inception-V3 | - | 79.45 | 79.23 | 79.34 | - | - | - | - |
| VRKD [58] | IJCAI 2019 | ResNet50 | 69.79 | 82.13 | 80.35 | 81.23 | - | - | - | - |
| AAP [59] | IJCAI 2019 | ResNet50 | 68.37 | 81.04 | 80.27 | 80.65 | - | - | - | - |
| VAC [60] | CVPR 2019 | ResNet50 | - | - | - | - | 64.51 | 75.77 | 79.43 | 77.10 |
| ALM [61] | ICCV 2019 | BN-Inception | 68.17 | 74.71 | 86.48 | 80.16 | 64.79 | 73.93 | 82.03 | 77.77 |
| JLAC [14] | AAAI 2020 | ResNet50 | 69.15 | 79.31 | 82.40 | 80.82 | 64.42 | 75.69 | 79.18 | 77.40 |
| SSCsoft [62] | ICCV 2021 | ResNet50 | 68.37 | 75.05 | 87.49 | 80.43 | - | - | - | - |
| IAA-Caps [63] | PR 2022 | OSNet | 68.47 | 79.56 | 82.06 | 80.37 | - | - | - | - |
| MCFL [64] | NCA 2022 | ResNet50 | 67.28 | 73.44 | **87.75** | 79.96 | - | - | - | - |
| DRFormer [22] | NC 2022 | ViT-B/16 | 70.60 | 80.12 | 82.77 | 81.42 | - | - | - | - |
| VAC-Combine [65] | IJCV 2022 | ResNet50 | 70.12 | **81.56** | 81.51 | 81.54 | - | - | - | - |
| DAFL [66] | AAAI 2022 | ResNet50 | 68.18 | 77.41 | 83.39 | 80.29 | 66.70 | 76.39 | 82.07 | 79.13 |
| CGCN [13] | TMM 2022 | ResNet50 | 54.40 | 60.03 | 83.68 | 70.49 | - | - | - | - |
| CAS-SAL-FR [67] | IJCV 2022 | ResNet50 | 68.59 | 77.56 | 83.81 | 80.56 | - | - | - | - |
| VTB [18] | TCSVT 2022 | ViT-B/16 | 69.44 | 78.28 | 84.39 | 80.84 | 67.48 | 76.41 | **83.32** | 79.35 |
| FEMDAR [68] | SPIC 2023 | ResNet50 | 66.88 | 79.11 | 79.24 | 78.76 | - | - | - | - |
| EALC [69] | NC 2023 | EfficientNet-B4 | 69.65 | 79.82 | 83.61 | 81.67 | - | - | - | - |
| PARFormer-L [21] | TCSVT 2023 | Swin-L | 69.94 | 79.63 | **88.19** | 81.35 | - | - | - | - |
| DFDT [70] | EAAI 2023 | Swin-B | **70.89** | 80.36 | 84.32 | **82.15** | 69.30 | **79.38** | 82.62 | **80.97** |
| OAGCN [71] | TMM 2023 | Swin-B | 69.32 | 78.32 | 87.29 | **82.56** | - | - | - | - |
| VTB* [18] | TCSVT 2022 | ViT-L/14 | 69.78 | 78.09 | 85.21 | 81.10 | **67.58** | 76.19 | **84.00** | 79.52 |
| SequencePAR | - | ViT-L/14 | **71.47** | **82.40** | 82.09 | 82.05 | **70.14** | **81.37** | 81.22 | **81.10** |

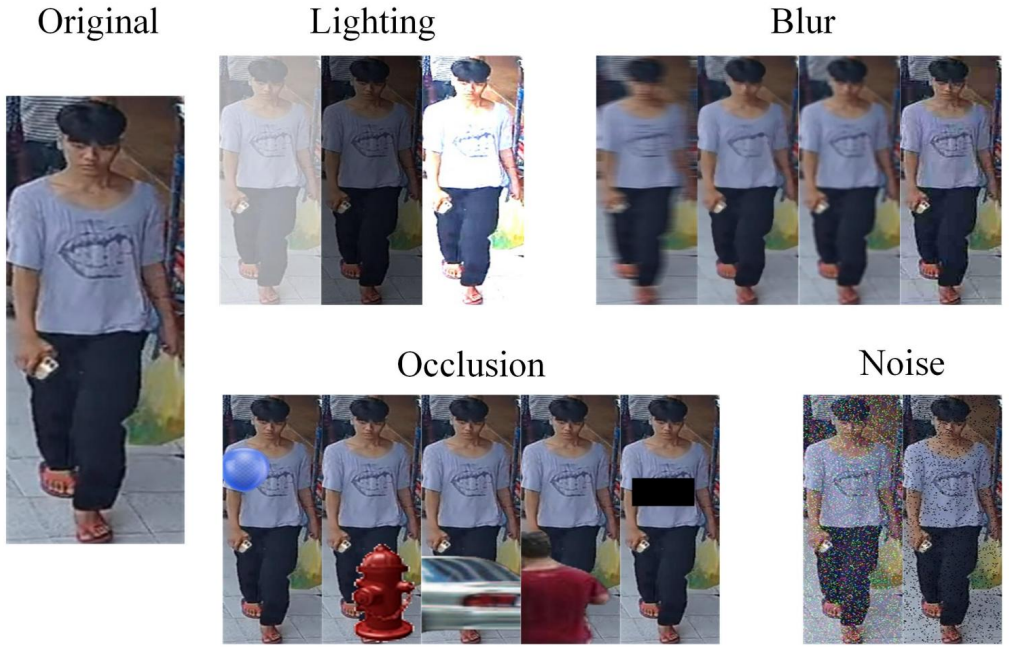
# When Big Models Meet PAR



## ➢ MSP60K Benchmark Dataset

- ***No large-scale PAR benchmark datasets are released in recent 5 years***
  - *Existing benchmark datasets are close to saturation;*
  - *Cross-domain (e.g., different environments, times, populations, and data sources) on the PAR is seldomly considered;*
  - *Data corruption during real-world application.*

| Dataset | Year | Attributes | Images | Scene Split |
|---|---|---|---|---|
| PETA [3] | 2014 | 61 | 19,000 | ✗ |
| WIDER [22] | 2016 | 14 | 57,524 | ✗ |
| RAPv1 [17] | 2016 | 69 | 41,585 | ✗ |
| PA100K [26] | 2017 | 26 | 100,000 | ✗ |
| RAPv2 [18] | 2019 | 76 | 84,928 | ✗ |
| Ours | 2024 | 57 | 60,015 | ✓ |




**Pedestrian Attribute Recognition: A New Benchmark Dataset and A Large Language Model Augmented Framework**
Jiandong Jin, Xiao Wang, Qian Zhu, Haiyang Wang, Chenglong Li, arXiv Pre-print arXiv:2408.09720, 2024

## ➢ MSP60K Benchmark Dataset

- ***Protocols for MSP60K Dataset***

**1). Large Scale**: We annotate 60,122 pedestrian images, each with 57 attributes, comprehensively analyzing pedestrian characteristics in various conditions.

**2). Multiple Distances and Viewpoints:** Images are captured from different angles and distances using various cameras and handheld devices, covering the front, back, and side views. The resolution of pedestrian images in our dataset is from 30×80 to 2005×3008.

**3). Complex and Varied Scenes:** Unlike existing datasets with uniform backgrounds, our dataset includes images from eight different environments with diverse backgrounds and attribute distributions, helping evaluate recognition methods in varied settings.

**4). Rich Source of Pedestrian Identity:** We gather data on pedestrians from different scenarios, nationalities, and seasonal variations, enhancing the dataset with diverse styles and characteristics.

**5). Simulated Complex Real-world Environments:** The dataset includes variations in lighting, motion blur, occlusions, and adverse weather conditions, simulating real-world challenges in pedestrian attribute recognition.

**6). Benchmark:** 17 open source PAR models are benchmarked on the MSP60K dataset.

# When Big Models Meet PAR



## ➢ MSP60K Benchmark Dataset

- *Attribute Groups*

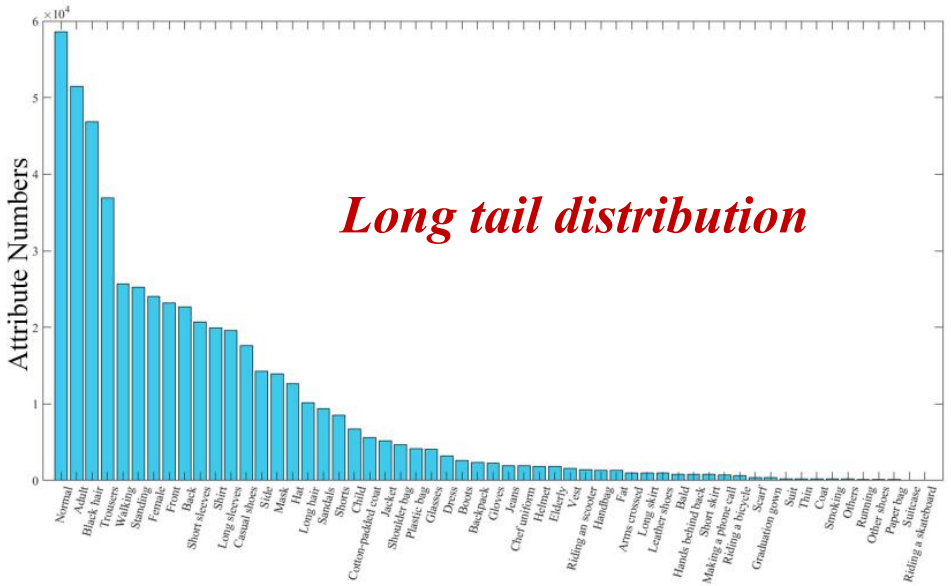
| Attribute Group | Details |
|---|---|
| Gender | Female |
| Age | Child, Adult, Elderly |
| Body Size | Fat, Normal, Thin |
| Viewpoint | Front, Back, Side |
| Head | Bald, Long Hair, Black Hair, Hat<br>Glasses, Mask, Helmet, Scarf, Gloves |
| Upper Body | Short Sleeves, Long Sleeves, Shirt, Jacket, Suit, Vest<br>Cotton Coat, Coat, Graduation Gown, Chef Uniform |
| Lower Body | Trousers, Shorts, Jeans, Long Skirt, Short Skirt, Dress |
| Shoes | Leather Shoes, Casual Shoes, Boots, Sandals, Other Shoes |
| Bag | Backpack, Shoulder Bag, Hand Bag<br>Plastic Bag, Paper Bag, Suitcase, Others |
| Activity | Calling, Smoking, Hands Back, Arms Crossed |
| Posture | Walking, Running, Standing, Bicycle, Scooter, Skateboard |

# When Big Models Meet PAR

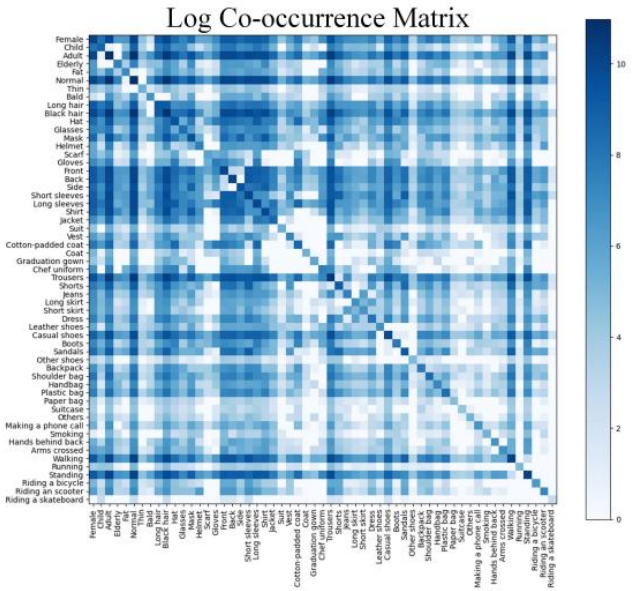


## ➢ MSP60K Benchmark Dataset

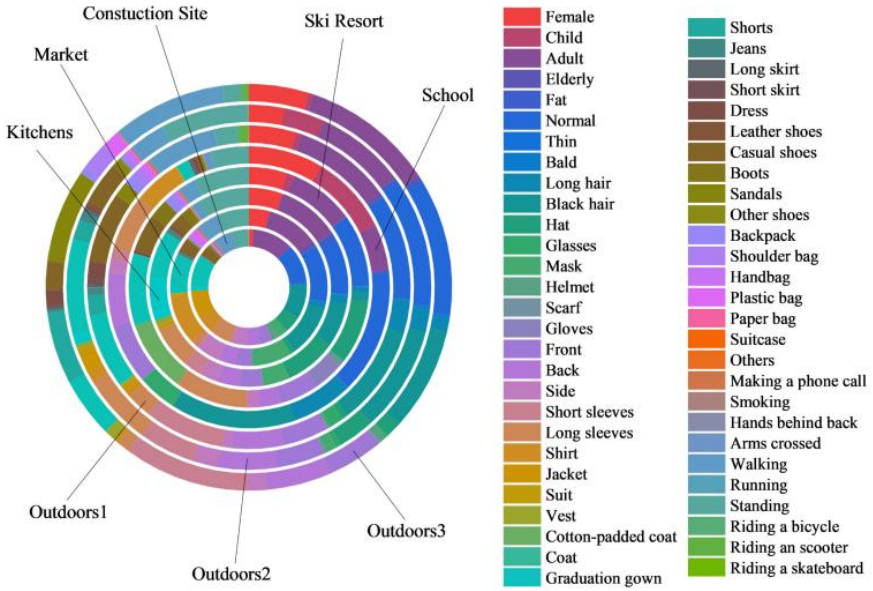
- *Statistical Analysis*




(a) Attributes Distribution




(b) Co-occurrence Matrix of Attributes

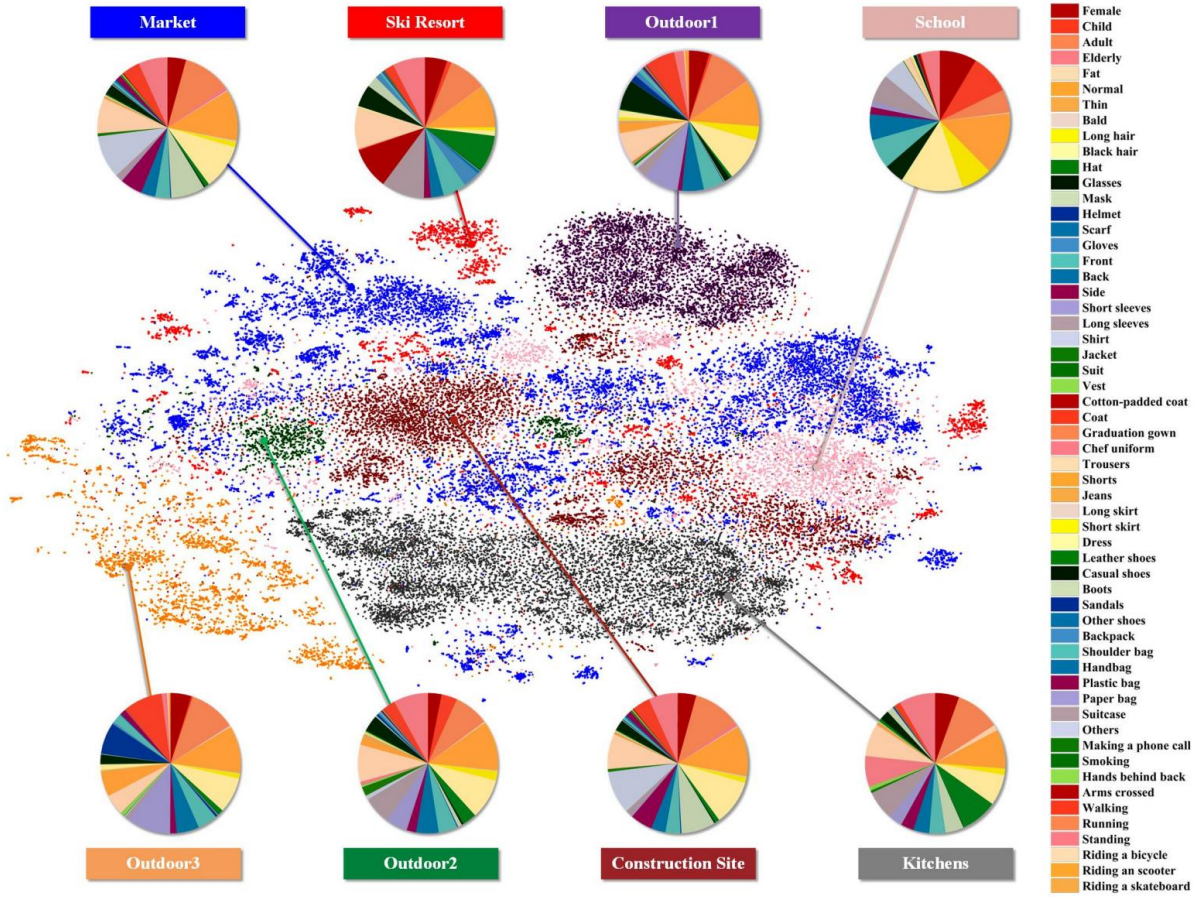



(c) Attributes Distribution of Different Sence

# When Big Models Meet PAR



## ➢ MSP60K Benchmark Dataset



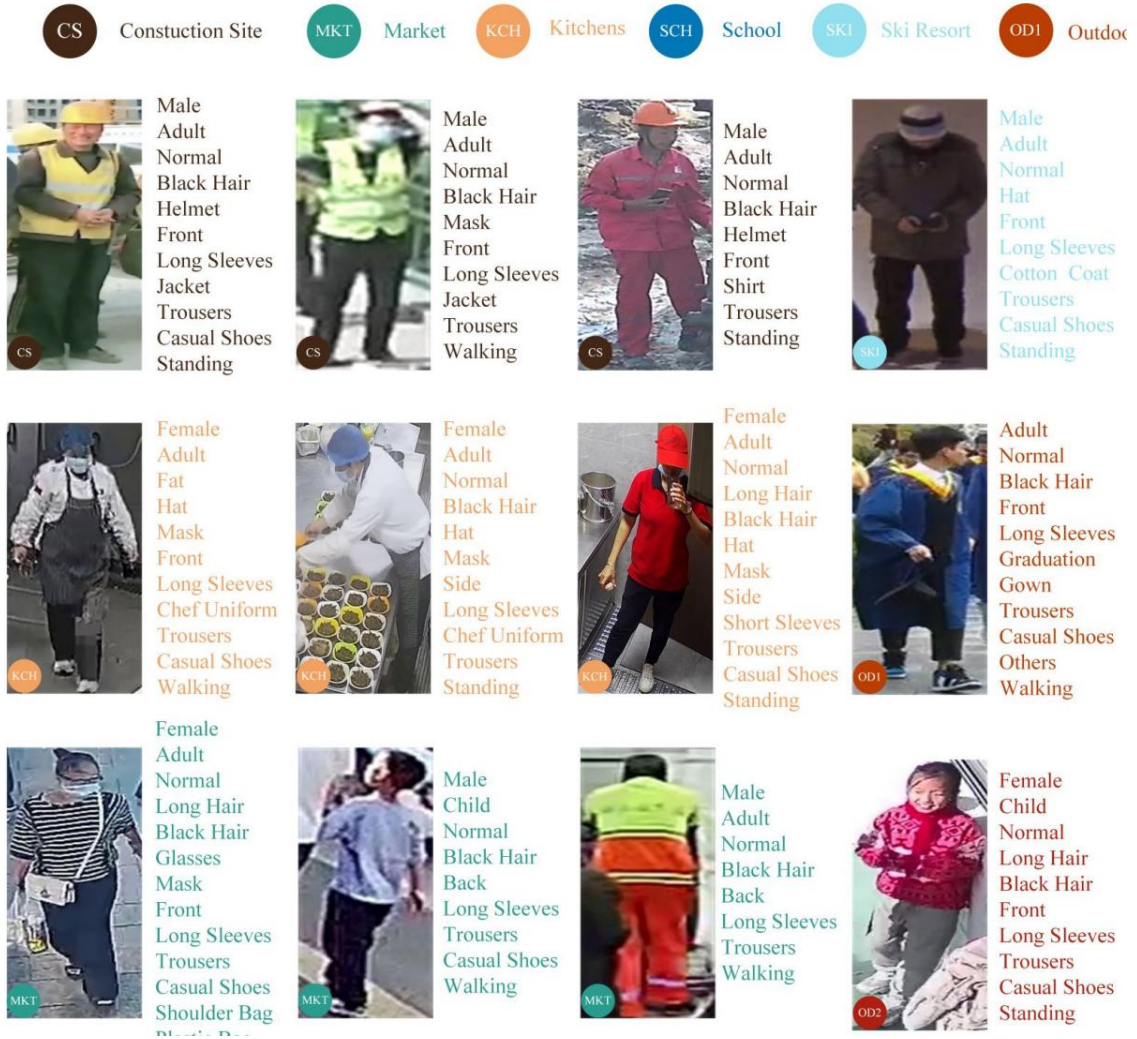



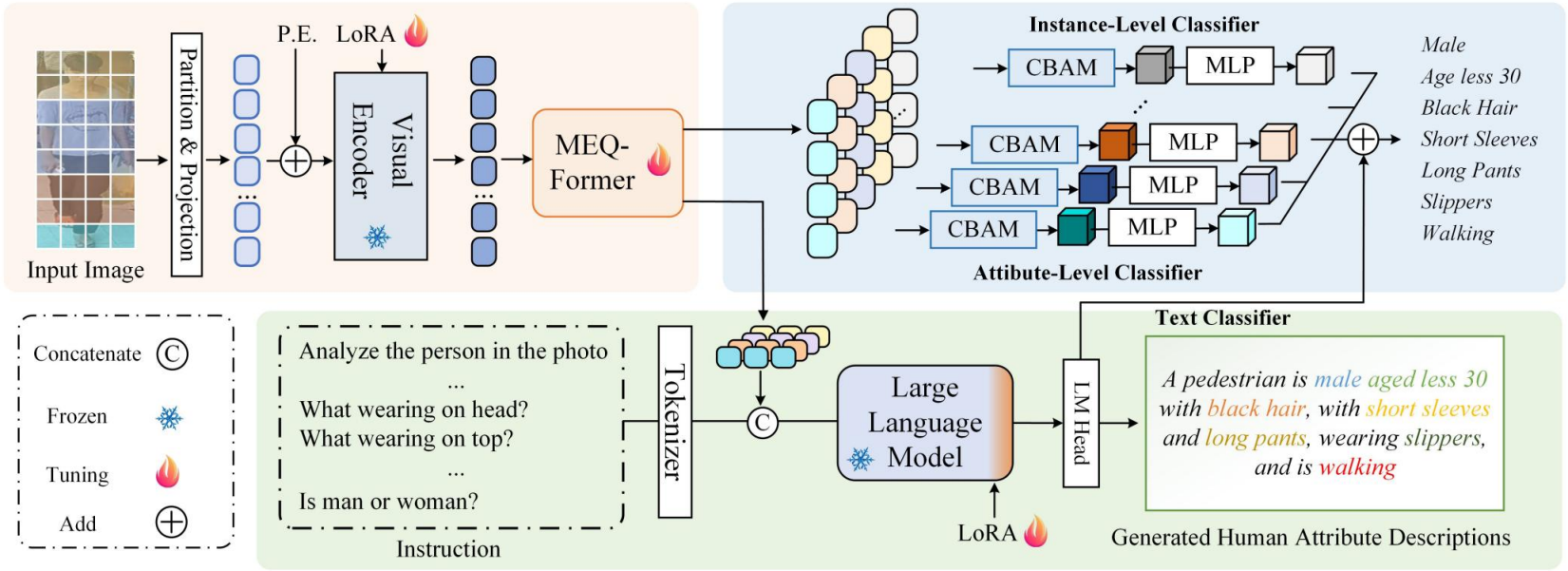

Figure 4. T-SNE visualization of scene samples in the MSP60K PAR dataset. Each colored cluster represents samples from different scenes, including “Market,” “Ski Resort,” “Outdoor1,” “School,” “Outdoor3,” “Outdoor2,” “Construction Site”, and “Kitchens”. For each scene, a pie chart is overlaid to illustrate the attribute distribution within that cluster. The legend on the right provides a detailed list of all attributes.

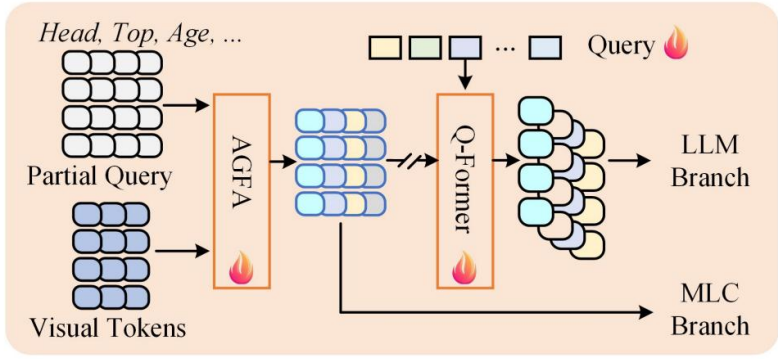
# When Big Models Meet PAR



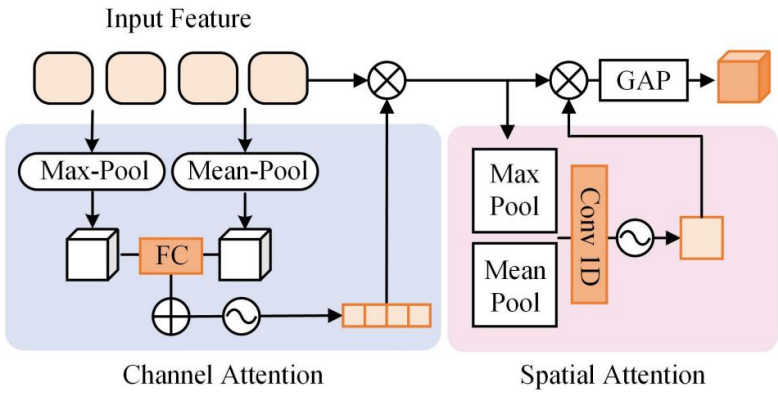
## ➢ MSP60K Benchmark Dataset

- ***LLM-PAR Framework***




(a). MEQ-Former




(b). CBAM module

An illustration of our proposed LLM-PAR framework.

# When Big Models Meet PAR



## ➢ MSP60K Benchmark Dataset

| Methods | Publish | Code | Random Split | | | | | Cross-domain Split | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | mA | Acc | Prec | Recall | F1 | mA | Acc | Prec | Recall | F1 |
| #01 DeepMAR [16] | ACPR15 | URL | 70.46 | 72.83 | 84.71 | 81.46 | 83.06 | 54.84 | 44.97 | 63.38 | 58.81 | 61.01 |
| #02 Strong Baseline [] | - | URL | 74.09 | 73.74 | 84.06 | 83.51 | 83.31 | 55.91 | 46.25 | 63.28 | 61.34 | 61.64 |
| #03 RethinkingPAR [11] | arXiv20 | URL | 74.01 | 74.20 | 84.17 | 83.94 | 84.06 | 55.98 | 46.52 | 62.85 | 62.09 | 62.47 |
| #04 SSCNet [10] | ICCV21 | URL | 69.71 | 69.31 | 79.22 | 82.47 | 80.82 | 52.84 | 40.88 | 56.26 | 58.64 | 57.43 |
| #05 VTB [2] | TCSVT22 | URL | 76.09 | 75.36 | 83.56 | 86.46 | 84.56 | 58.59 | 49.81 | 65.11 | 66.11 | 65.00 |
| #06 Label2Label [21] | ECCV22 | URL | 73.61 | 72.66 | 81.79 | 84.32 | 82.56 | 56.38 | 45.81 | 59.67 | 64.20 | 61.19 |
| #07 DFDT [46] | EAAI22 | URL | 74.19 | 76.35 | **85.03** | 86.35 | 85.69 | 57.85 | 49.97 | 65.34 | 66.18 | 65.76 |
| #08 Zhou et al. [48] | IJCAI23 | URL | 73.07 | 68.76 | 78.38 | 82.10 | 80.20 | 54.26 | 41.91 | 56.23 | 60.11 | 58.11 |
| #09 PARFormer [5] | TCSVT23 | URL | 76.14 | 76.67 | 84.77 | 86.93 | 85.44 | 57.96 | 50.63 | 62.28 | 71.04 | 65.82 |
| #10 SequencePAR [13] | arXiv23 | URL | 71.88 | 71.99 | 83.24 | 82.29 | 82.29 | 57.88 | 50.27 | 65.81 | 65.79 | 65.37 |
| #11 VTB-PLIP [51] | arXiv23 | URL | 73.90 | 73.16 | 82.01 | 84.82 | 82.93 | 56.30 | 46.77 | 61.20 | 64.47 | 62.18 |
| #12 Rethink-PLIP [51] | arXiv23 | URL | 69.44 | 68.90 | 79.82 | 81.15 | 80.48 | 57.18 | 46.98 | 63.57 | 62.16 | 62.86 |
| #13 PromptPAR [37] | arXiv23 | URL | **78.81** | **76.53** | **84.40** | **87.15** | 85.35 | **63.24** | **53.62** | **66.15** | **71.84** | **68.32** |
| #14 SSPNet [49] | PR24 | URL | 74.03 | 74.10 | 84.01 | 84.02 | 84.02 | 56.15 | 46.75 | 62.44 | 63.07 | 62.75 |
| #15 HAP [45] | NIPS24 | URL | 76.92 | 76.12 | 84.78 | 86.14 | **85.45** | 58.70 | 50.59 | 65.60 | 66.91 | 66.25 |
| #16 MambaPAR [39] | arXiv24 | URL | 73.85 | 73.64 | 83.19 | 84.29 | 83.28 | 56.75 | 47.34 | 61.92 | 64.98 | 62.80 |
| #17 MaHDFT [38] | arXiv24 | URL | 74.08 | 74.40 | 82.82 | 86.41 | 83.93 | 58.67 | 50.65 | 62.39 | 71.13 | 65.85 |
| Zero-shot | - | - | 56.93 | 52.97 | 72.26 | 64.69 | 67.46 | 52.19 | 39.26 | 60.12 | 52.09 | 55.15 |
| Ours | - | - | **80.13** | **78.71** | 84.39 | **90.52** | **86.94** | **66.29** | **58.11** | **65.68** | **81.21** | **72.05** |

# When Big Models Meet PAR



## ➢ MSP60K Benchmark Dataset

| Methods | Publish | PETA | | | | | PA100K | | | | | RAPv1 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | mA | Acc | Prec | Recall | F1 | mA | Acc | Prec | Recall | F1 | mA | Acc | Prec | Recall | F1 |
| SSCsoft [10] | ICCV21 | 86.52 | 78.95 | 86.02 | 87.12 | 86.99 | 81.87 | 78.89 | 85.98 | 89.10 | 86.87 | 82.77 | 68.37 | 75.05 | 87.49 | 80.43 |
| IAA [41] | PR22 | 85.27 | 78.04 | 86.08 | 85.80 | 85.64 | 81.94 | 80.31 | 88.36 | 88.01 | 87.80 | 81.72 | 68.47 | 79.56 | 82.06 | 80.37 |
| MCFL [1] | NCA22 | 86.83 | 78.89 | 84.57 | 88.84 | 86.65 | 81.53 | 77.80 | 85.11 | 88.20 | 86.62 | 84.04 | 67.28 | 73.44 | 87.75 | 79.96 |
| DRFormer [33] | NC22 | 89.96 | 81.30 | 85.68 | 91.08 | 88.30 | 82.47 | 80.27 | 87.60 | 88.49 | 88.04 | 81.81 | 70.60 | 80.12 | 82.77 | 81.42 |
| VAC [7] | IJCV22 | - | - | - | - | - | 82.19 | 80.66 | 88.72 | 88.10 | 88.41 | 81.30 | 70.12 | 81.56 | 81.51 | 81.54 |
| DAFL [12] | AAAI22 | 87.07 | 78.88 | 85.78 | 87.03 | 86.40 | 83.54 | 80.13 | 87.01 | 89.19 | 88.09 | 83.72 | 68.18 | 77.41 | 83.39 | 80.29 |
| CGCN [4] | TMM22 | 87.08 | 79.30 | 83.97 | 89.38 | 86.59 | - | - | - | - | - | 84.70 | 54.40 | 60.03 | 83.68 | 70.49 |
| CAS [44] | IJCV22 | 86.40 | 79.93 | 87.03 | 87.33 | 87.18 | 82.86 | 79.64 | 86.81 | 87.79 | 85.18 | 84.18 | 68.59 | 77.56 | 83.81 | 80.56 |
| VTB [2] | TCSVT22 | 85.31 | 79.60 | 86.76 | 87.17 | 86.71 | 83.72 | 80.89 | 87.88 | 89.30 | 88.21 | 82.67 | 69.44 | 78.28 | 84.39 | 80.84 |
| PromptPAR [37] | arXiv23 | 88.76 | 82.84 | **89.04** | 89.74 | 89.18 | 87.47 | **83.78** | **89.27** | **91.70** | **90.15** | 85.45 | **71.61** | 79.64 | 86.05 | 82.38 |
| PARformer [5] | TCSVT23 | 89.32 | 82.86 | 88.06 | **91.98** | 89.06 | 84.46 | 81.13 | 88.09 | 91.67 | 88.52 | 84.43 | 69.94 | 79.63 | **88.19** | 81.35 |
| OAGCN [28] | TMM23 | **89.91** | **82.95** | 88.26 | 89.10 | 88.68 | 83.74 | 80.38 | 84.55 | 90.42 | 87.39 | **87.83** | 69.32 | 78.32 | 87.29 | **82.56** |
| SSPNet [31] | PR24 | 88.73 | 82.80 | **88.48** | 90.55 | **89.50** | 83.58 | 80.63 | 87.79 | 89.32 | 88.55 | 83.24 | 70.21 | **80.14** | 82.90 | 81.50 |
| SOFA [42] | AAAI24 | 87.10 | 81.10 | 87.80 | 88.40 | 87.80 | 83.40 | 81.10 | **88.40** | 89.00 | 88.30 | 83.40 | 70.00 | **80.00** | 83.00 | 81.20 |
| FRDL [49] | ICML24 | 88.59 | - | - | - | 89.03 | **89.44** | - | - | - | 88.05 | 87.72 | - | - | - | 79.16 |
| Zero-shot | - | 61.32 | 50.75 | 68.57 | 64.00 | 65.52 | 65.26 | 56.99 | 79.21 | 65.20 | 70.75 | 65.46 | 50.90 | 64.48 | 65.20 | 66.06 |
| Ours | - | **92.25** | **84.59** | 88.41 | **92.94** | **90.39** | **91.09** | **84.12** | 87.73 | **94.09** | **90.41** | **87.80** | **71.86** | 78.36 | **88.20** | **82.64** |

# When Big Models Meet PAR



## ➢ MSP60K Benchmark Dataset

| # | CLS-Attr | FT Q-Former | LoRA | CLS-LLM | AGFA | CLS-IN | PETA Dataset | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | mA | Acc | F1 |
| 1 | ✓ | | | | | | 71.54 | 58.24 | 71.96 |
| 2 | ✓ | ✓ | | | | | 82.89 | 72.32 | 81.89 |
| 3 | ✓ | ✓ | ✓ | | | | 90.14 | 83.25 | 89.38 |
| 4 | ✓ | ✓ | ✓ | ✓ | | | 90.89 | 83.64 | 89.60 |
| 5 | ✓ | ✓ | ✓ | ✓ | ✓ | | 91.78 | 84.47 | 90.27 |
| 6 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | 92.25 | 84.59 | 90.39 |

Component Analysis on the PETA Dataset.

| Replacement | CLS-Mean | | CLS-LLM | |
|---|---|---|---|---|
| | mA | F1 | mA | F1 |
| #1 Ground Truth | 91.53 | 86.11 | 77.03 | 70.91 |
| #2 25% Mask(Padding) | 92.15 | 89.44 | 86.90 | 87.35 |
| #3 50% Mask(Padding) | 92.33 | 89.21 | 88.20 | 88.07 |
| #4 75% Mask(Padding) | 92.25 | 89.23 | 86.64 | 85.49 |
| #5 100% Mask(Padding) | 91.70 | 89.64 | 64.59 | 65.43 |
| #6 Random Sentence | 92.25 | 90.39 | 88.84 | 89.22 |

| LLMs | Vicuna-7B | OPT-6.7B |
|---|---|---|
| mA | 92.25 | 92.12 |
| F1 | 90.39 | 89.39 |

# When Big Models Meet PAR



## ➢ MSP60K Benchmark Dataset

**Question:** *Analyze the photo, What is on the head? What is the top wear? What is the bottom wear? What type of shoes? What type of bag? Age? Is the person male or female? Body shape? Facing to where? Activity?*

**GT:** Bald, Mask, Shirt, Trousers, Adult, Normal, Side, Standing

**MiniGPT-4:** The image shows a man wearing a black suit and a white shirt, standing in front of a counter with a sign that reads "Cash Only" on it. The man is looking at something on the counter. The man's face is obscured by a mask. The activity is unknown. The man's age is unknown. The man's body shape is unknown. The man is not carrying any bags.

**LLM Branch:** The adult man with normal body, and has bald head and mask, is facing to back, wearing shirt on tops, and trousers on bottoms, is standing.

**Vision Branch:** Hat, Gloves, Long sleeves, Shirt, Cotton coat, Trousers, Casual shoes, Female, Adult, Normal, Side, Standing

**GT:** Hat, Long sleeves, Cotton-padded coat, Trousers, Casual shoes, Adult, Normal, Front, Standing

**MiniGPT-4:** The image shows the person carrying the backpack is wearing a black hat. The activity appears to be skating. The person is likely male, and unknown age and gender. The body shape is slim. The person is facing forward.

**LLM Branch:** The adult man with normal body, and with hat, is facing to front, wearing long sleeves and cotton-padded coat on tops, and trousers on bottoms, with casual shoes, is standing.

**Vision Branch:** Hat, Gloves, Long sleeves, Cotton-padded coat, Trousers, Boots, Handbag, Adult, Normal, Front, Standing

**GT:** Long hair, Black hair, Mask, Long sleeves, Trousers, Female, Adult, Normal, Front, Standing

**MiniGPT-4:** The image shows a group of people wearing face masks and walking down a street. The people are wearing different types of clothing, including a white shirt, a blue jacket, and a black hoodie. Some of the people are carrying bags, while others are not. The people in the image are facing in different directions, and some are looking at something in the distance. The activity appears to be walking.

**LLM Branch:** The adult woman with normal body, and has long hair and black hair, is facing to front, wearing long sleeves on tops, and trousers on bottoms and dress on bottoms, and carrying backpack, is walking.

**Vision Branch:** Black hair, Short sleeves, Shorts, Casual shoes, Plastic bag, Female, Adult, Normal, Front, Walking

**GT:** Long hair, Black hair, Short sleeves, Shorts, Casual shoes, Female, Child, Normal, Front, Walking

**MiniGPT-4:** The image shows a woman wearing a purple shirt and white pants, walking down a street with a red umbrella in her hand. The woman is carrying a small bag on her shoulder. The woman is facing forward and appears to be walking towards the camera. The activity appears to be walking. The woman's body shape appears to be slender. The woman's age is difficult to determine from the image. The woman's face is not visible in the image. The woman's hair appears to be long and straight.

**LLM Branch:** The child woman with normal body, and has long hair and black hair, is facing to front, wearing short sleeves on tops, and shorts on bottoms, with casual shoes, is walking.

**Vision Branch:** Long hair, Black hair, Short sleeves, Shorts, Casual shoes, Female, Child, Normal, Front, Walking

# *Overview*

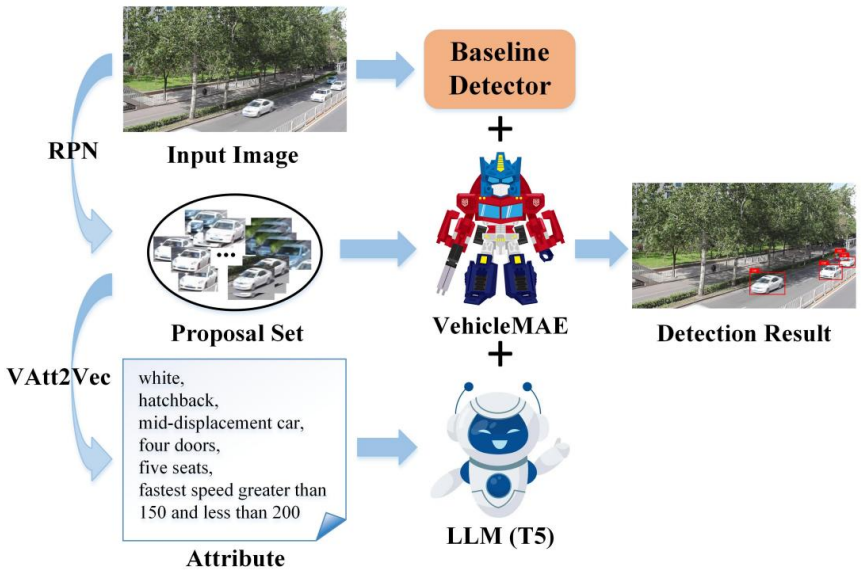


- **Background of Pedestrian Attribute Recognition (PAR)**
  - Task definition, Review of PAR (Major Challenges, Datasets, Mainstream Algorithms, … )
- **When Big Models Meet PAR**
  - CLIP, MAE, LLM, …
  - VTB, PromptPAR, SequencePAR, LLM-PAR, …
  - **Applications on Other Tasks**
    - Vehicle Detection,
    - MOT,
    - Retrieval,
    - Re-ID,
    - …
- **Conclusion & Discussion**

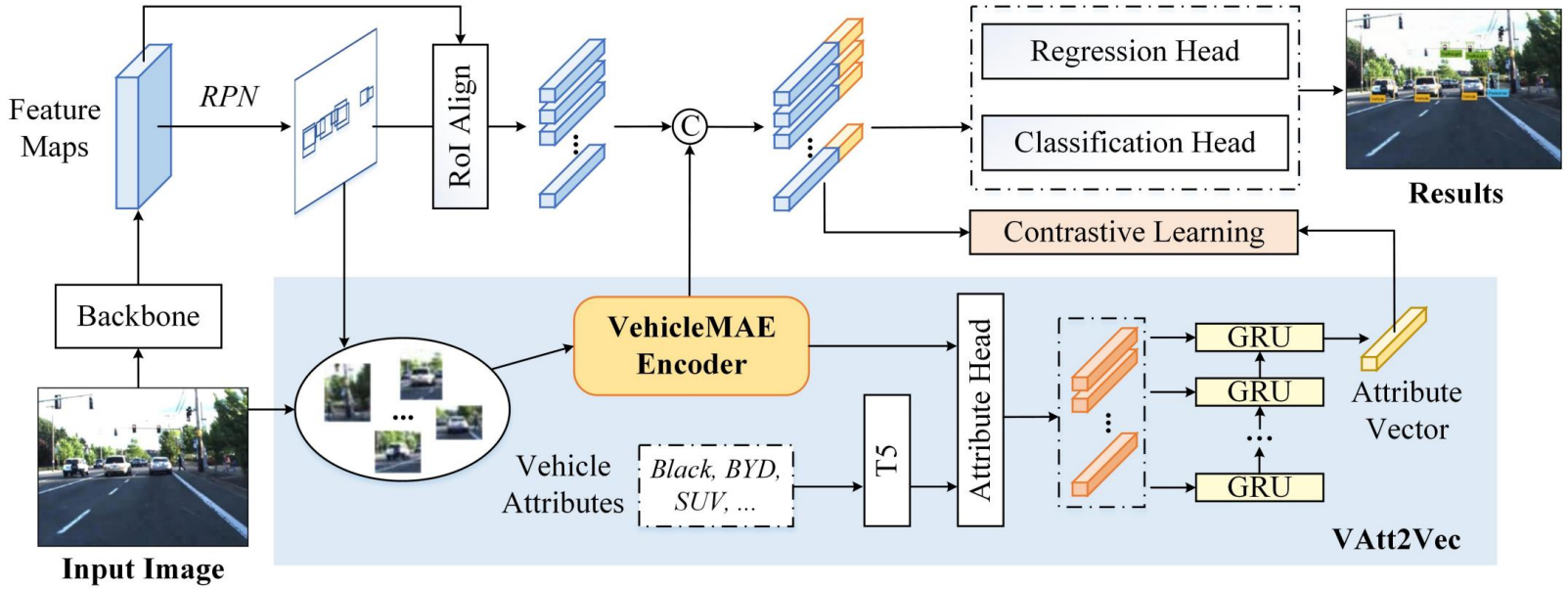
# When Big Models Meet PAR



**VFM-Det: Towards High-Performance Vehicle Detection via Large Foundation Models**,
Wentao Wu, et al. arXiv:2408.13031




Motivation

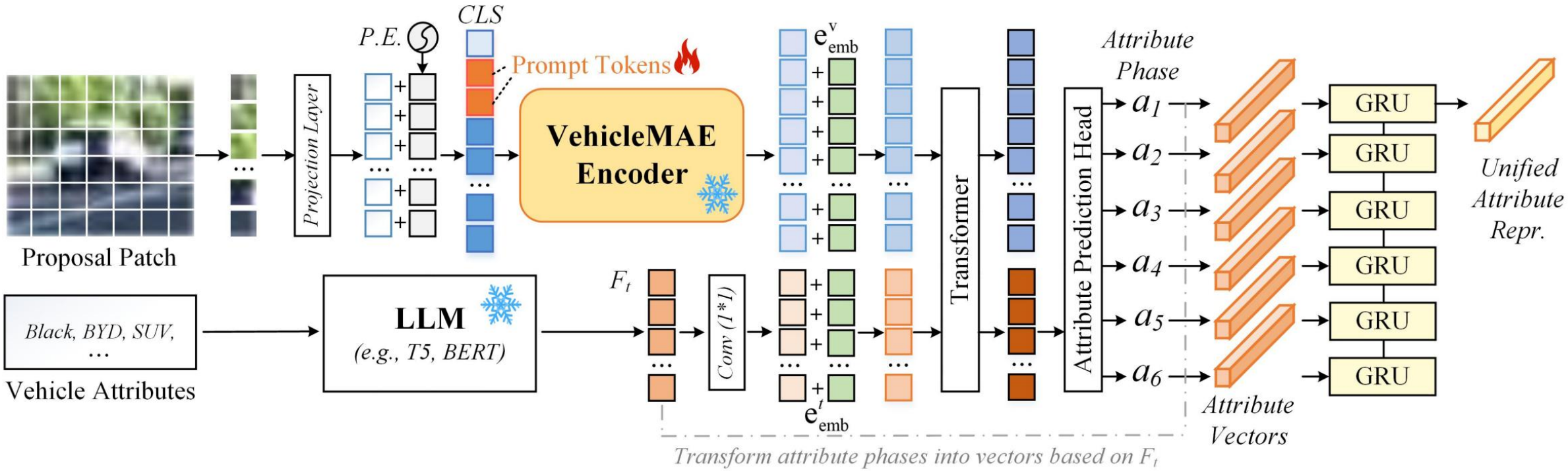



Framework

# When Big Models Meet PAR



**VFM-Det: Towards High-Performance Vehicle Detection via Large Foundation Models**,
Wentao Wu, et al. arXiv:2408.13031




The detailed architecture of our proposed **VAtt2Vec** module.

# When Big Models Meet PAR



**VFM-Det: Towards High-Performance Vehicle Detection via Large Foundation Models**,
Wentao Wu, et al. arXiv:2408.13031

| Method | Pre-train Modality | Cityscapes | | | COCO | | | UA-DETRAC | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $AP_{[0.5:0.95]}$ | $AP_{0.5}$ | $AP_{0.75}$ | $AP_{[0.5:0.95]}$ | $AP_{0.5}$ | $AP_{0.75}$ | $AP_{[0.5:0.95]}$ | $AP_{0.5}$ | $AP_{0.75}$ |
| DINO | Vision | 45.0 | 65.6 | 49.3 | 48.5 | 71.6 | 54.4 | 49.7 | 72.6 | 60.6 |
| IBOT | Vision | 44.1 | 64.8 | 48.3 | 50.0 | 73.5 | 55.2 | 51.4 | 73.1 | 63.0 |
| MoCov3 | Vision | 41.8 | 60.8 | 46.5 | 47.5 | 69.2 | 52.7 | 48.5 | 71.6 | 60.1 |
| MAE | Vision | 42.4 | 61.8 | 46.4 | 47.6 | 69.1 | 52.9 | 49.6 | 71.6 | 60.6 |
| MAE† | Vision | 43.7 | 63.0 | 47.8 | 47.1 | 69.2 | 52.1 | 49.8 | 71.9 | 60.9 |
| Ours | Vision and Unmatched Text | **46.9** | **66.5** | **51.6** | **51.5** | **75.3** | **56.6** | **51.7** | **73.7** | **63.1** |

➢ Comparison with other pre-trained models for vehicle detection.

| Method | Backbone | Cityscapes | | | COCO | | | UA-DETRAC | | | Params(M) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $AP_{[0.5:0.95]}$ | $AP_{0.5}$ | $AP_{0.75}$ | $AP_{[0.5:0.95]}$ | $AP_{0.5}$ | $AP_{0.75}$ | $AP_{[0.5:0.95]}$ | $AP_{0.5}$ | $AP_{0.75}$ | |
| Faster R-CNN | ResNet50 | 34.0 | 54.0 | 34.8 | 43.3 | 67.5 | 47.0 | 47.3 | 69.5 | 57.4 | 42 |
| Mask R-CNN | ResNet50 | 41.7 | 61.4 | 45.4 | 45.9 | 68.0 | 50.8 | 48.0 | 70.5 | 58.0 | 44 |
| RetinaNet | ResNet50 | 43.1 | 60.5 | 46.6 | 43.5 | 67.1 | 48.4 | 47.0 | 69.5 | 56.2 | 38 |
| DetectoRS | ResNet50 | 43.9 | 62.9 | 47.3 | 49.9 | 70.7 | 53.8 | **51.8** | 71.8 | 62.2 | 123 |
| Swin-T | Swin-Transformer | 44.0 | 63.9 | 48.9 | 46.6 | 69.8 | 51.7 | 49.7 | 71.0 | 60.1 | 47 |
| VitDet | ViT-Base | 45.2 | 64.1 | 50.1 | 50.4 | 72.4 | 55.8 | 51.0 | **75.1** | 61.1 | 141 |
| VFM-Det (Ours) | ResNet50/ViT-Base | **46.9** | **66.5** | **51.6** | **51.5** | **75.3** | **56.6** | 51.7 | 73.7 | **63.1** | 153 |

➢ Comparison with other object detection algorithms

# When Big Models Meet PAR



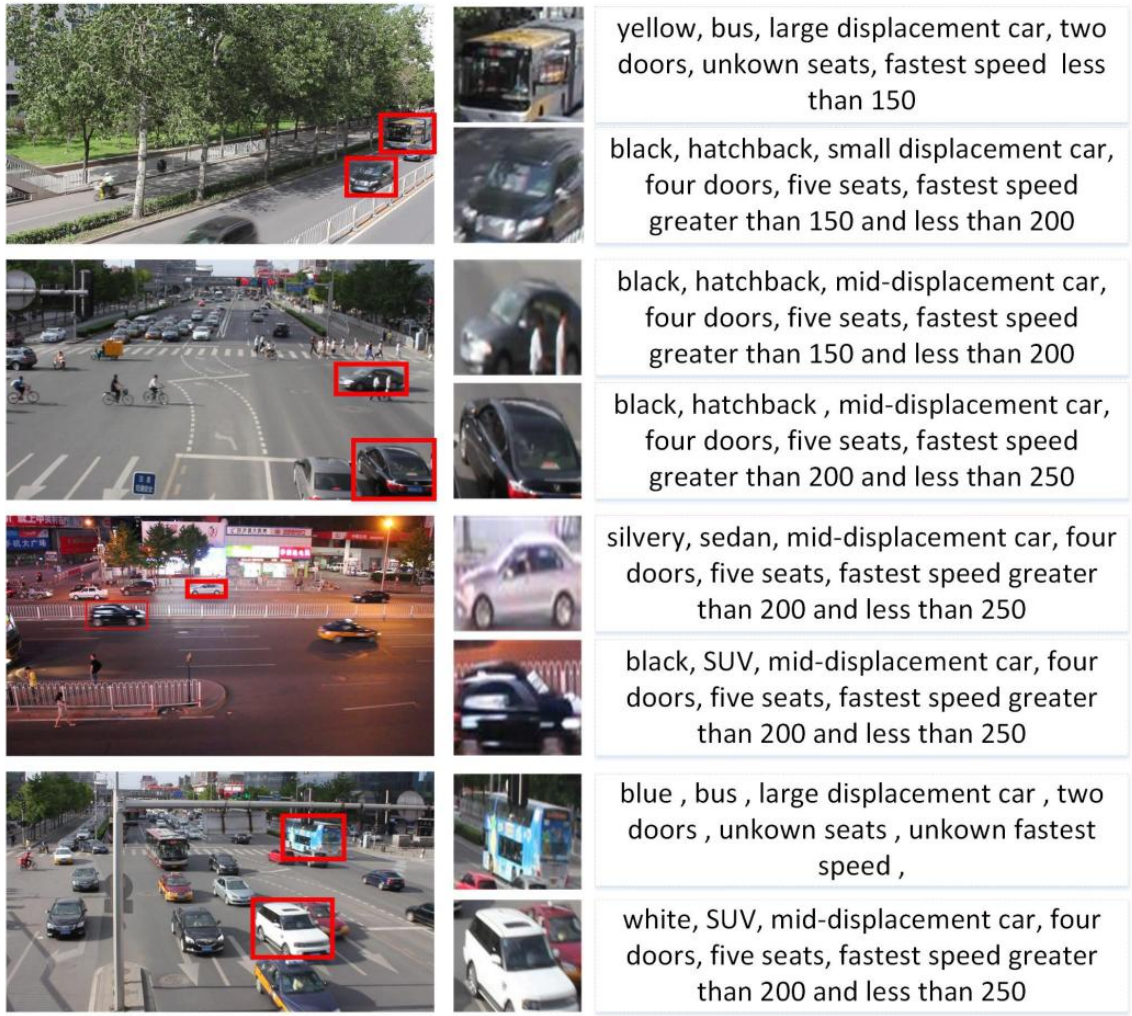
**VFM-Det: Towards High-Performance Vehicle Detection via Large Foundation Models**,
Wentao Wu, et al. arXiv:2408.13031

| Method | $AP_{[0.5:0.95]}$ | $AP_{0.5}$ | $AP_{0.75}$ |
|---|---|---|---|
| Baseline | 41.7 | 61.4 | 45.4 |
| +VehicleMAE Encoder | 45.0 | 65.4 | 48.9 |
| +VAtt2Vec | **46.9** | **66.5** | **51.6** |

➢ Component analysis

| LLM | $AP_{[0.5:0.95]}$ | $AP_{0.5}$ | $AP_{0.75}$ |
|---|---|---|---|
| BERT [19] | 45.2 | 63.1 | 51.2 |
| ALBERT [66] | 45.9 | 64.2 | 50.6 |
| MPNet [67] | 45.3 | 64.5 | 51.4 |
| CLIP [25] | 46.5 | 66.1 | **52.0** |
| T5 [18] | **46.9** | **66.5** | 51.6 |

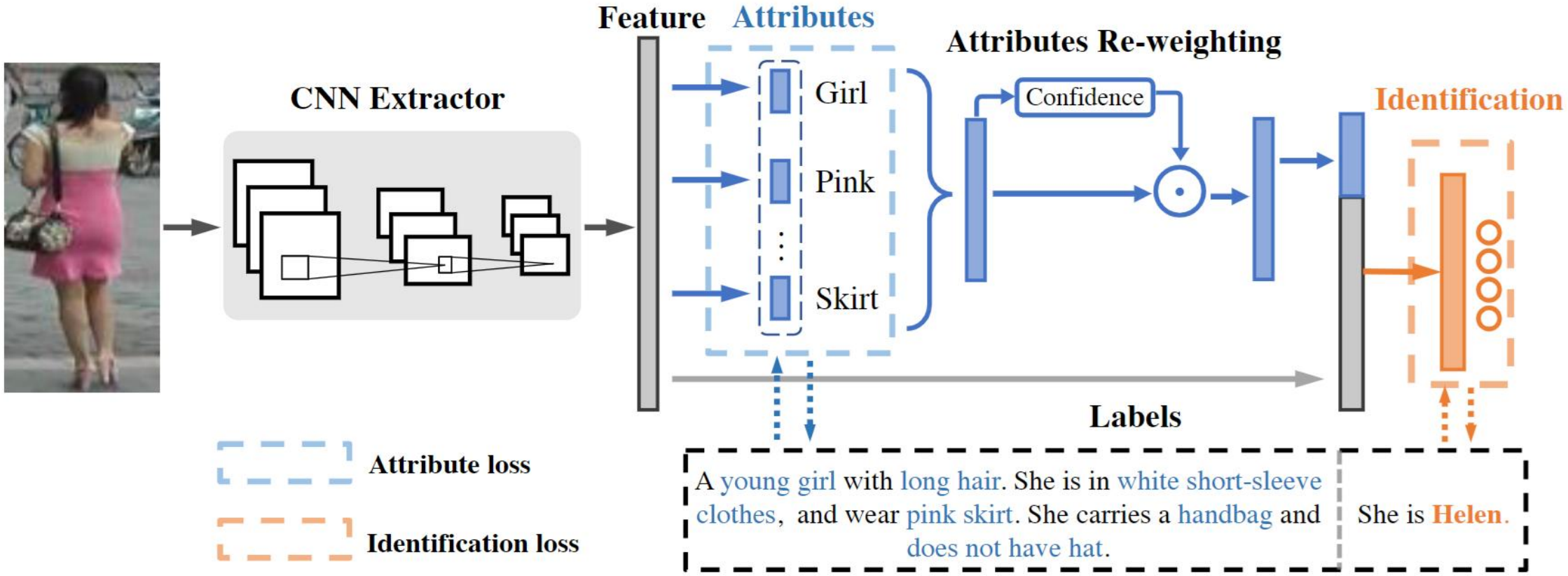
➢ Results of Different LLM



# When Big Models Meet PAR



[Pattern Recognition] Lin, et al. **Improving person re-identification by attribute and identity learning**. , 95, 151-161. (2019).

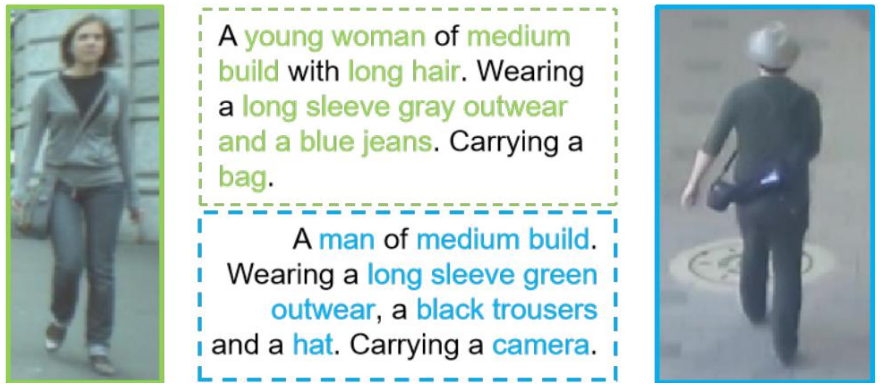



https://arxiv.org/pdf/1703.07220

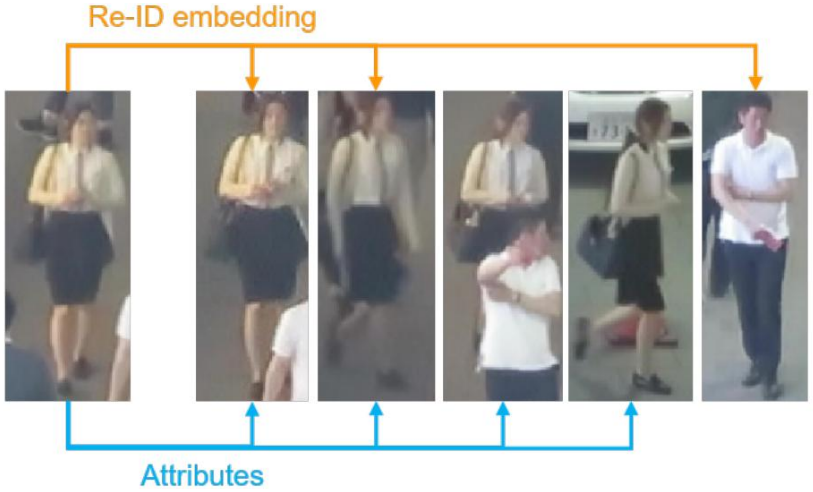
# When Big Models Meet PAR



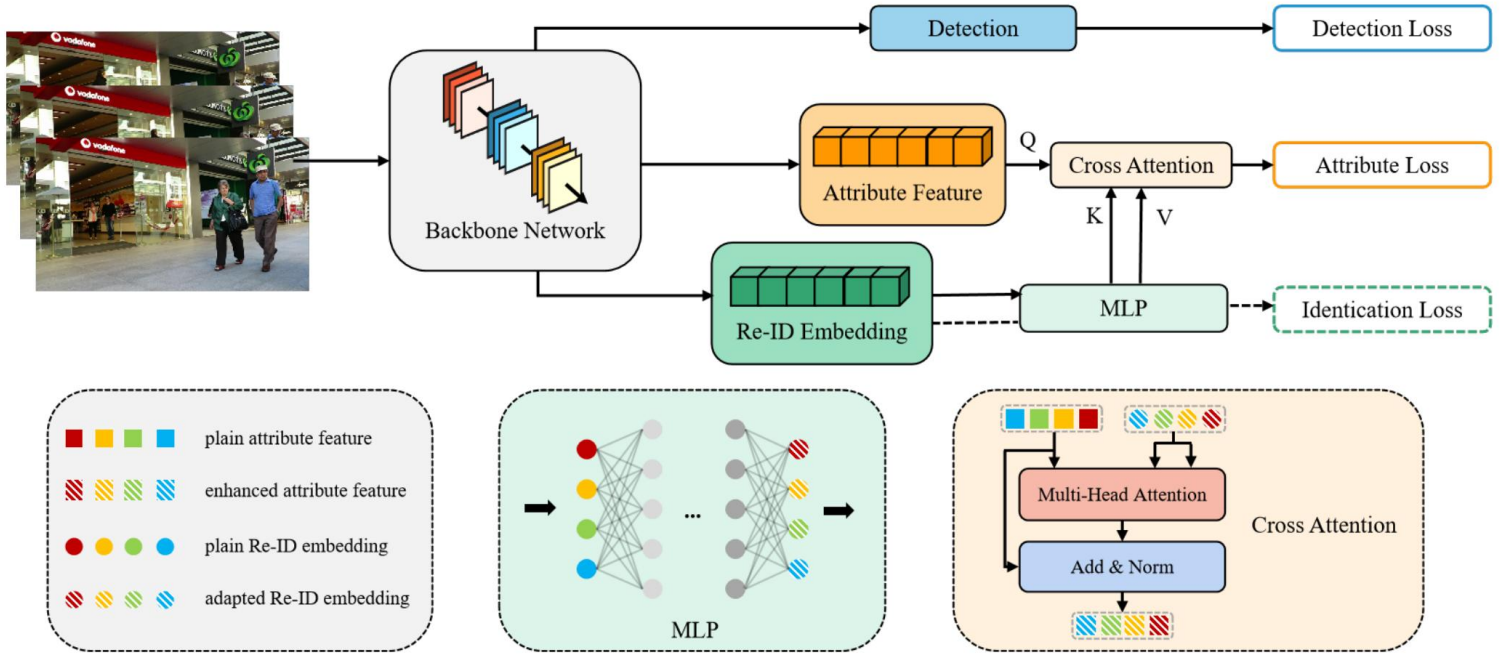
[IEEE TNNLS 2024] Li, Yunhao, et al. "**AttMOT: improving multiple-object tracking by introducing auxiliary pedestrian attributes.**"




(a) Examples of how attributes describe a person.




(b) Attributes are motion and occlusion irrelevant.

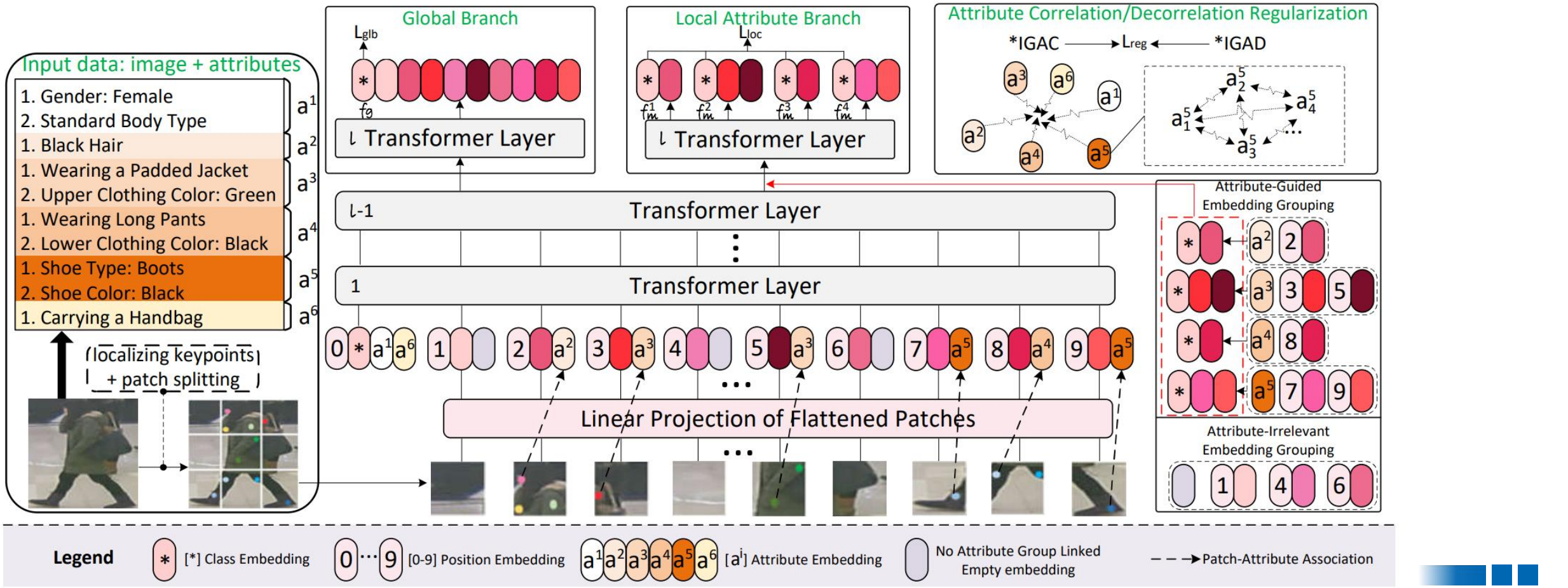



https://arxiv.org/pdf/2308.07537

# When Big Models Meet PAR



[CVPR-2024] Huang, Yan, et al. "**Attribute-Guided Pedestrian Retrieval: Bridging Person Re-ID with Internal Attribute Variability.**"



# *Overview*



- **Background of Pedestrian Attribute Recognition (PAR)**
  - Task definition, Review of PAR (Major Challenges, Datasets, Mainstream Algorithms, … )
- **When Big Models Meet PAR**
  - CLIP, MAE, LLM, …
  - VTB, PromptPAR, SequencePAR, LLM-PAR, …
  - Applications on Other Tasks
- **Conclusion & Discussion**

# Conclusion & Discussion



## ➢ Background of PAR

- Pedestrian Attribute Recognition: A Survey, Xiao Wang, Shaofei Zheng, Rui Yang, Aihua Zheng, Zhe Chen, Bin Luo, Jin Tang, Pattern Recognition, 2021.

## ➢ Big Model based PAR

- Cheng, Xinhua, et al. "**A simple visual-textual baseline for pedestrian attribute recognition.**" *IEEE TCSVT* - 2022
- Jun Zhu, et al. "**Learning clip guided visual-text fusion transformer for video-based pedestrian attribute recognition.**" CVPR-2023 Workshop,
- Xiao Wang, et al. "**Spatio-Temporal Side Tuning Pre-trained Foundation Models for Video-based Pedestrian Attribute Recognition.**" arXiv preprint arXiv:2404.17929 (2024).
- Jin, Jiandong, et al. "**Sequencepar: Understanding pedestrian attributes via a sequence generation paradigm.**" arXiv preprint arXiv:2312.01640 (2023).
- **Pedestrian Attribute Recognition: A New Benchmark Dataset and A Large Language Model Augmented Framework**, Jiandong Jin, Xiao Wang, Qian Zhu, Haiyang Wang, Chenglong Li, arXiv Pre-print arXiv:2408.09720, 2024

## ➢ Attribute based Applications

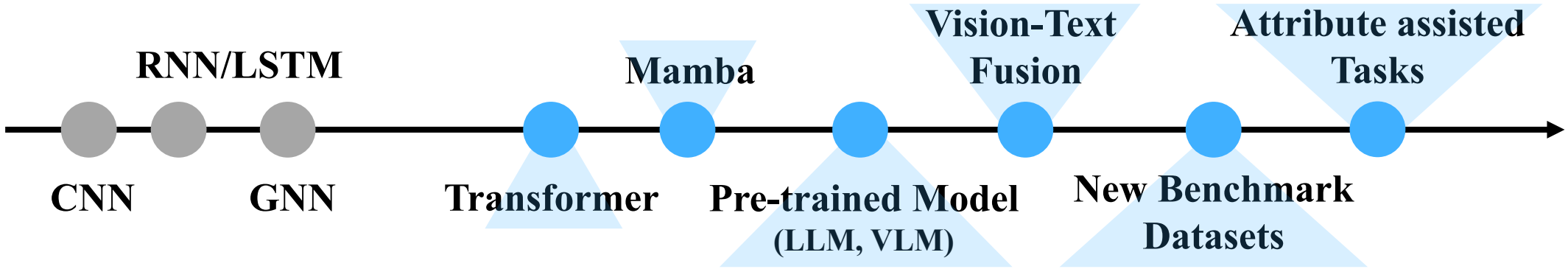
- Detection, MOT, Text-based Retrieval, Re-ID

## ➢ Other Works

- **An Empirical Study of Mamba-based Pedestrian Attribute Recognition**, Xiao Wang, Weizhe Kong, Jiandong Jin, Shiao Wang, Ruichong Gao, Qingchuan Ma, Chenglong Li, Jin Tang, arXiv Pre-print arXiv:2407.10374, 2024

# Conclusion & Discussion





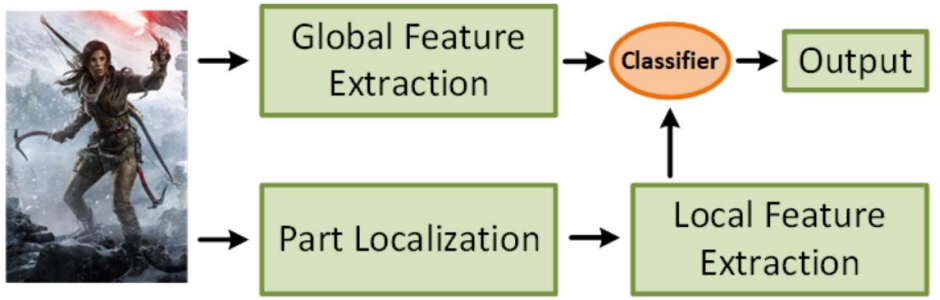


- **HumanBench: Towards General Human-centric Perception with Projector Assisted Pretraining**, Shixiang Tang, et al., CVPR-2023
- **"Unihcp: A unified model for human-centric perceptions."** Ci, Yuanzheng, et al. CVPR-2023.
- "**Beyond appearance: a semantic controllable self-supervised learning framework for human-centric visual tasks**." Chen, Weihua, et al. CVPR-2023.
- "**Plip: Language-image pre-training for person representation learning.**" Zuo, Jialong, et al. *arXiv preprint arXiv:2305.08386* (2023).

# Conclusion & Discussion



## Possible Research Directions:




① **More Accurate and Efficient Part Localization Algorithm**

② **Explore More Advanced Network Architecture, e.g., Mamba**

③ **Prior Knowledge guided Learning**
- We wear different clothes in various seasons, temperatures or occasions;
- History knowledge (e.g., Wikipedia)

④ **Multi-modal Pedestrian Attribute Recognition**
- Thermal, Depth, Event, Radar, …


⑤ **Video-based Pedestrian Attribute Recognition**

⑥ **Joint Learning of Attribute and Other Tasks**

⑦ **Pre-training for PAR**

⑧ **LLM/VLM for PAR**

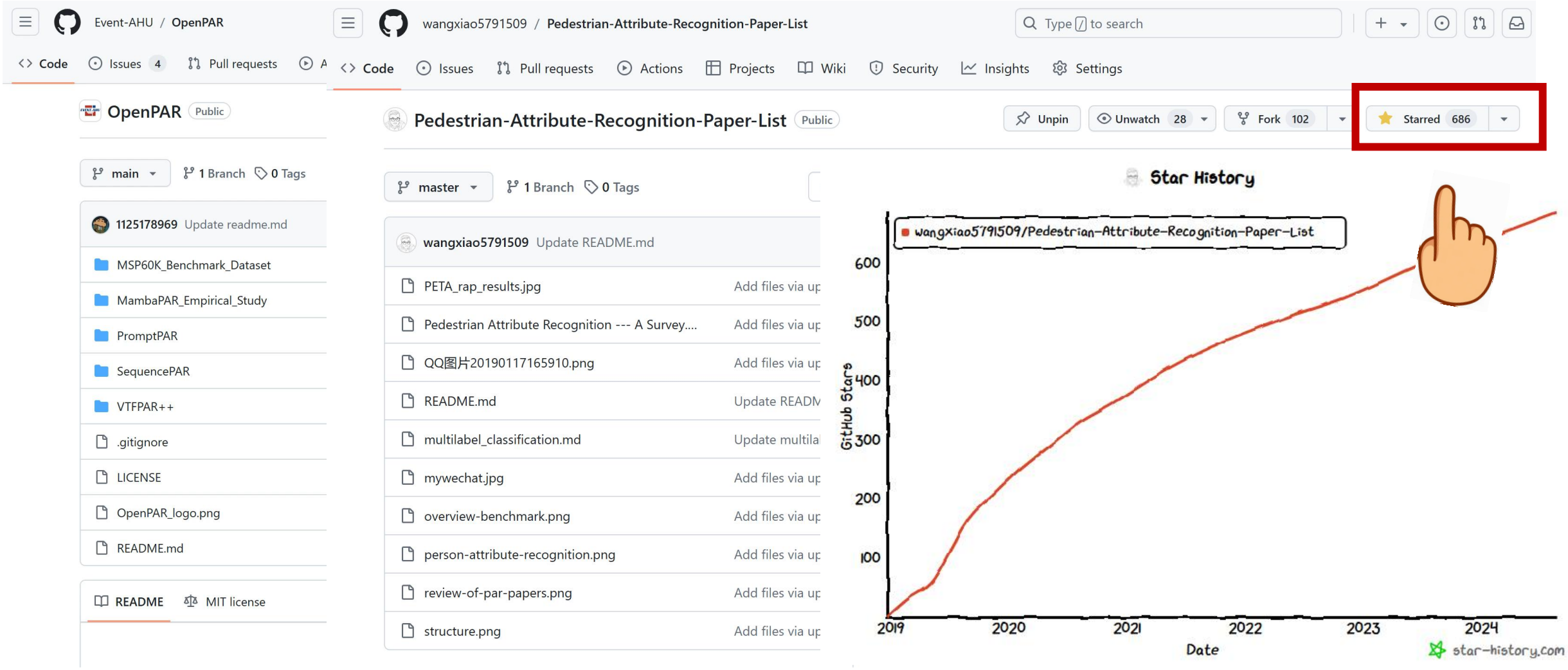
⑨ **New Benchmark Datasets for PAR**

CSIG · 云讲堂

OpenPAR

EVENT-AHU RESEARCH

- https://github.com/Event-AHU/OpenPAR
- https://github.com/wangxiao5791509/Pedestrian-Attribute-Recognition-Paper-List



CSIG · 云讲堂

EVENT-AHU RESEARCH

# *Thanks for your attention!*

## *Q & A*


安徽大学—结构模式与视觉学习研究组

Anhui University-Structural Patterns and Visual Learning (AHU-SPVL) Group

个人主页

https://wangxiao5791509.github.io/

Email: xiaowang@ahu.edu.cn