CAM-PAR: Class Activation Map Guided Feature Disentanglement for Pedestrian Attribute Recognition

Hyojeong Lee

Graduate School of Artificial Intelligence
UNIST

Contents

Introduction: pedestrian attribute recognition(PAR) and its previous works

UNIST

Pedestrian Attribute Recognition (PAR)

Fig. 1: examples of pedestrian images
From PA100K.

Pedestrian Attribute Recognition(PAR) :

- Aims to predict multiple pedestrian attributes for a given image.
- **Subtask of multi-label classification.**

Applications:

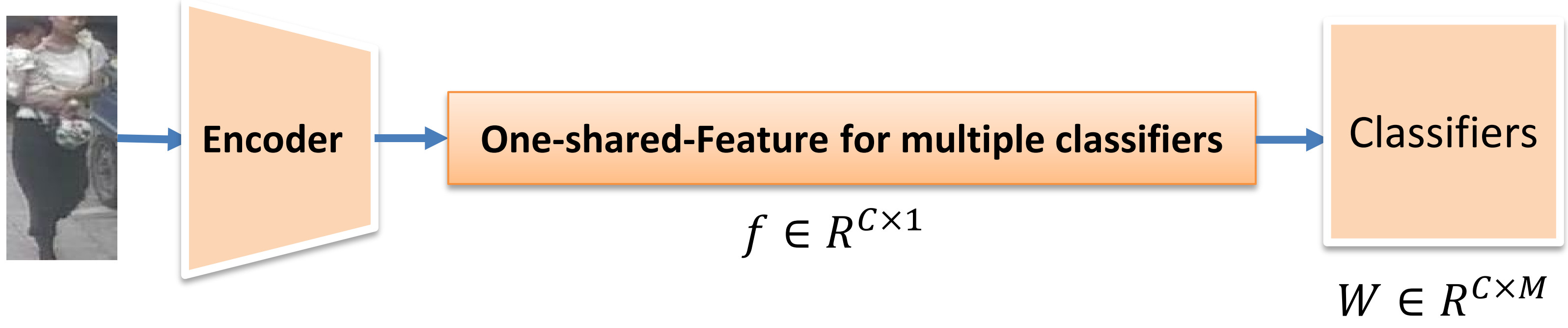
- scene-understanding.
- person re-identification and retrieval.

Common challenges:

1. Low resolution.
2. Unbalanced data distribution.
3. **Label dependency.**

UNIST

PAR as multi-label classification



Limitation of OFMA mechanism ([1] , Jian Jia et al)

Basic of multi-label classification:

w: classifier weights
f: encoder feature
p_t: threshold

$$\hat{y}_{i,j} = \begin{cases} 1 & if\ p_{i,j} \geq p_t \\ 0 & if\ p_{i,j} < p_t \end{cases}, \quad p_{i,j} = \sigma(logits_{i,j}) \qquad (1)$$

$$logits_{i,j} = w_j^T f = |w_j| \cdot |f| \cdot cos\theta \qquad (2)$$

$$\hat{y}_{i,j} = \begin{cases} 1 & if\ \ 0° \geq \theta \geq 90° \\ 0 & if\ \ 90° < \theta < 180° \end{cases} \qquad (3)$$

PAR as multi-label classification

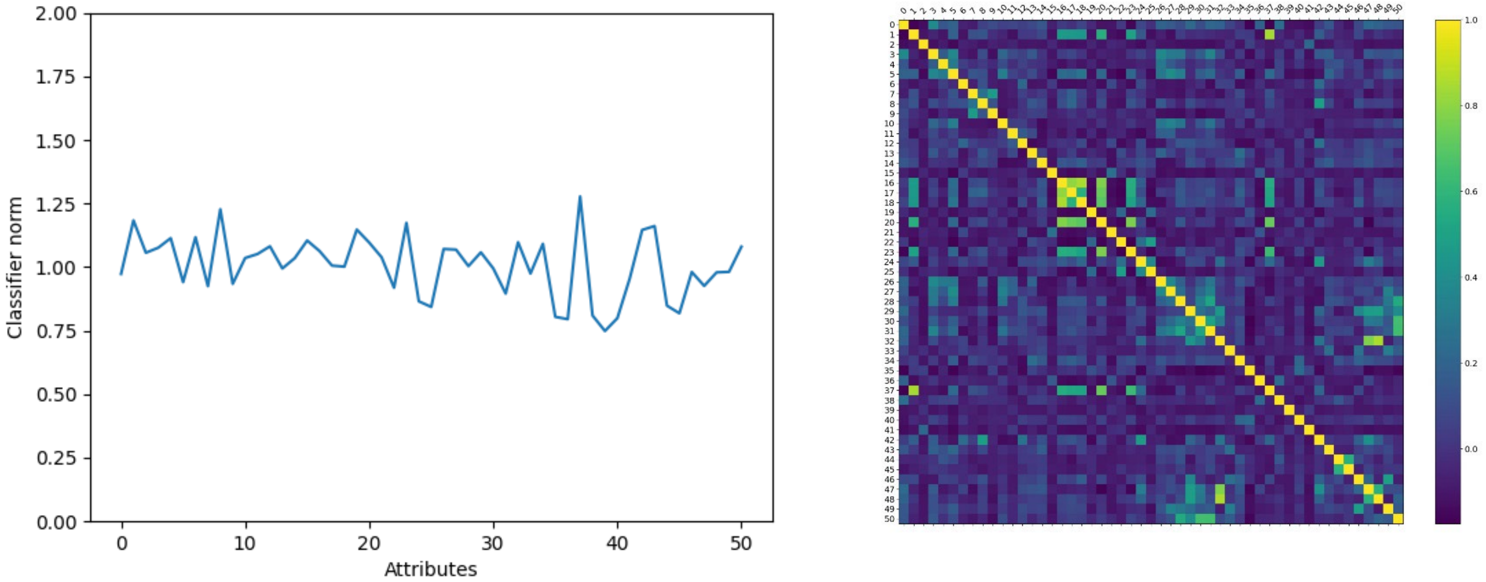


Fig. 2: classifier norms(left) and angles between classifier weights(right)

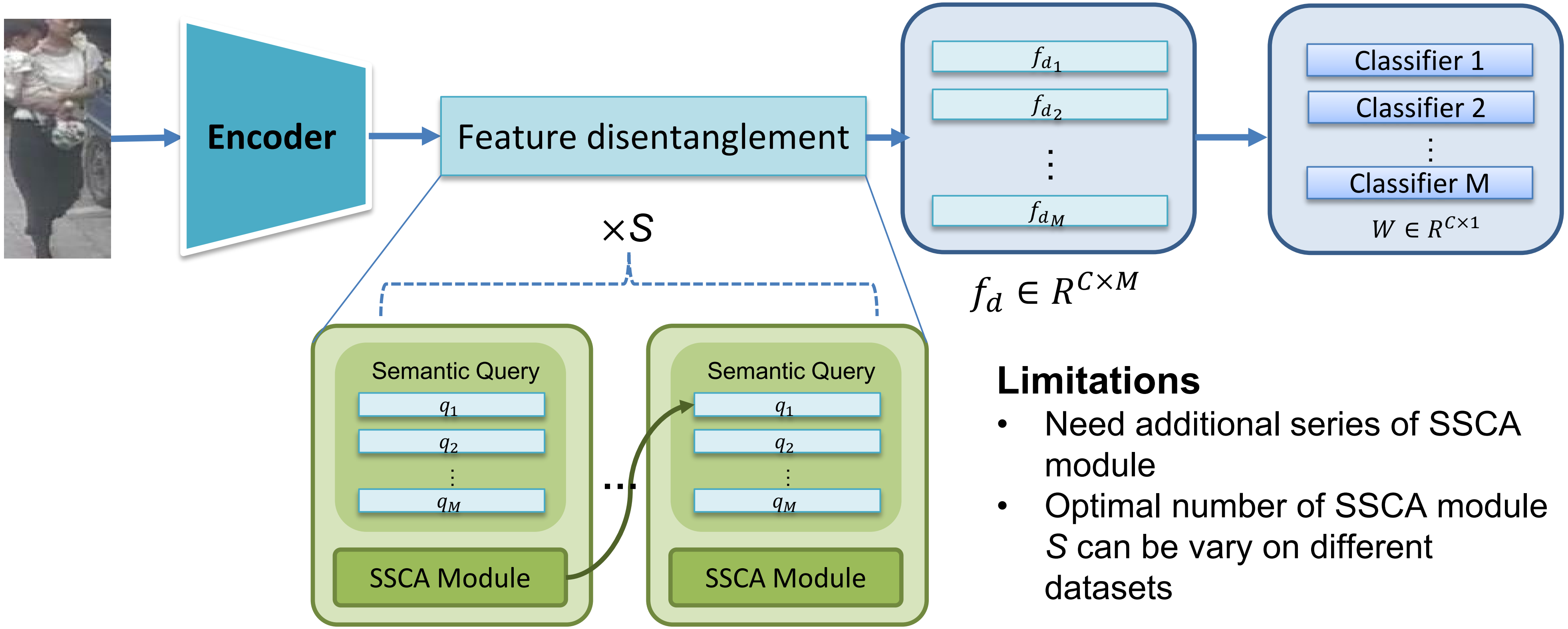
Two observations of well trained OFMA model

- Most classifier weights of attributes are orthogonal to each other
- Classifier norms are almost the same

UNIST

PAR as multi-label classification

Fig. 3: Structure of OFOA mechanism and SSCA module [1]



Limitations

- Need additional series of SSCA module
- Optimal number of SSCA module *S* can be vary on different datasets

*SSCA : cascaded semantic-spatial cross-attention

Attribute correlations of PAR

The main challenge in pedestrian attribute recognition is that different attributes are highly correlated.[5]

Pedestrian attribute recognition based on attribute correlation [2]:
⇒ Construct MXM matrix that model the relationships between any pair of attributes in the attribute set via self-attention.

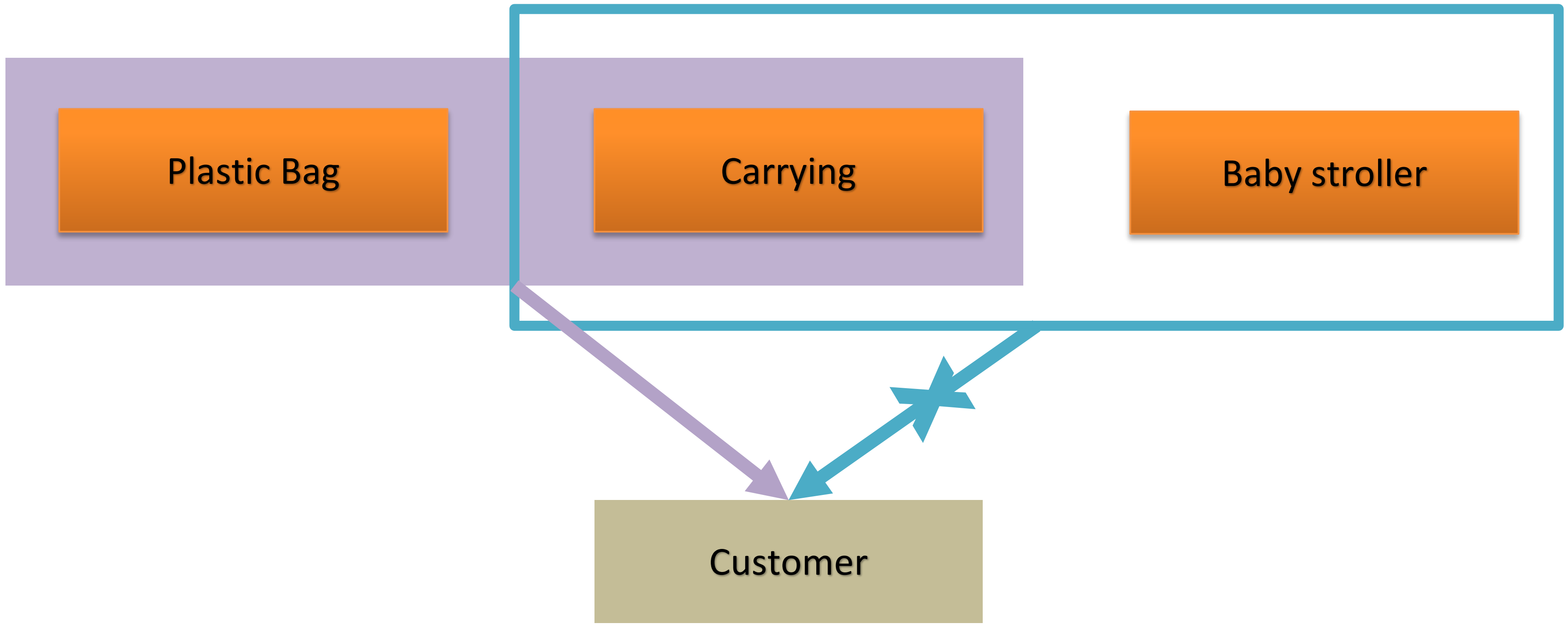
Multi-Label Image Recognition with graph convolutional Networks [4]:
⇒ Model the label correlation dependency in the form of conditional probability

Correlation Graph Convolutional Network for Pedestrian Attribute Recognition [3]:
⇒ Divide pedestrian attributes in three categories;inter, hierechical and positional, and mine relationships among attributes via self-attention.

Limitation:
All of above three previous works limited to attribute pairs, not attribute sets.

UNIST

Attribute correlations of PAR



Preliminary: collaborative filtering

Collaborative filtering (CF)

Recommender system:

- Content based : utilizes characteristics of item(e.g., genre, directors).
- Collaborative Filtering : utilizes past user behavior (e.g., clicks, purchases, ratings)

Types of user feedbacks:

- Explicit : indicate users' preference directly.
- Implicit : indirectly reflect opinion through user behavior, indicate frequency of actions named "confidence"

Collaborative filtering (CF)

Latent Factor Model using SVD for explicit feedback
⇒ Learn a latent factor that well explains a known user feedback.

Given user u, v and item i, j. User feedback of u over i is expressed as $r_{u,i}$.
Learning process:

$$\min(x, y) \sum_{u,i} \left(p_{u,i} - x_u^T y_i\right)^2 + \lambda(||x_u||^2 + ||y_i||^2) \tag{4}$$

where $x_u \in R^C$ is user-factor and $y_i \in R^C$ is item-factor.
For a known user u, predicted score for unknown item k for user u is as follows:

$$\hat{\boldsymbol{r}}_{\boldsymbol{u},\boldsymbol{k}} = \boldsymbol{x}_{\boldsymbol{u}}^{\boldsymbol{T}} \boldsymbol{y}_{\boldsymbol{k}} \tag{5}$$

Collaborative filtering (CF)

Collaborative filtering for implicit feedback datasets [6]
Preference from implicit feedback *confidence*

$$p_{u,i} = \begin{cases} 1 & if\ r_{u,i} > 0 \\ 0 & if\ r_{u,i} = 0 \end{cases} \tag{6}$$

Since implicit feedback does not directly indicate the user's preference, additional variable $c_{u,i}$ is introduced.

$$\boldsymbol{c_{u,i} = 1 + \alpha r_{u,i}} \tag{7}$$

Final cost function is,

$$\min(x, y) \sum_{u,i} c_{u,i}\left(p_{u,i} - x_u^T y_i\right)^2 + \lambda\left(\sum_u ||x_u||^2 + \sum_i ||\, y_i||^2\right) \tag{8}$$

UNIST

Proposed Methods

Methods: full framework

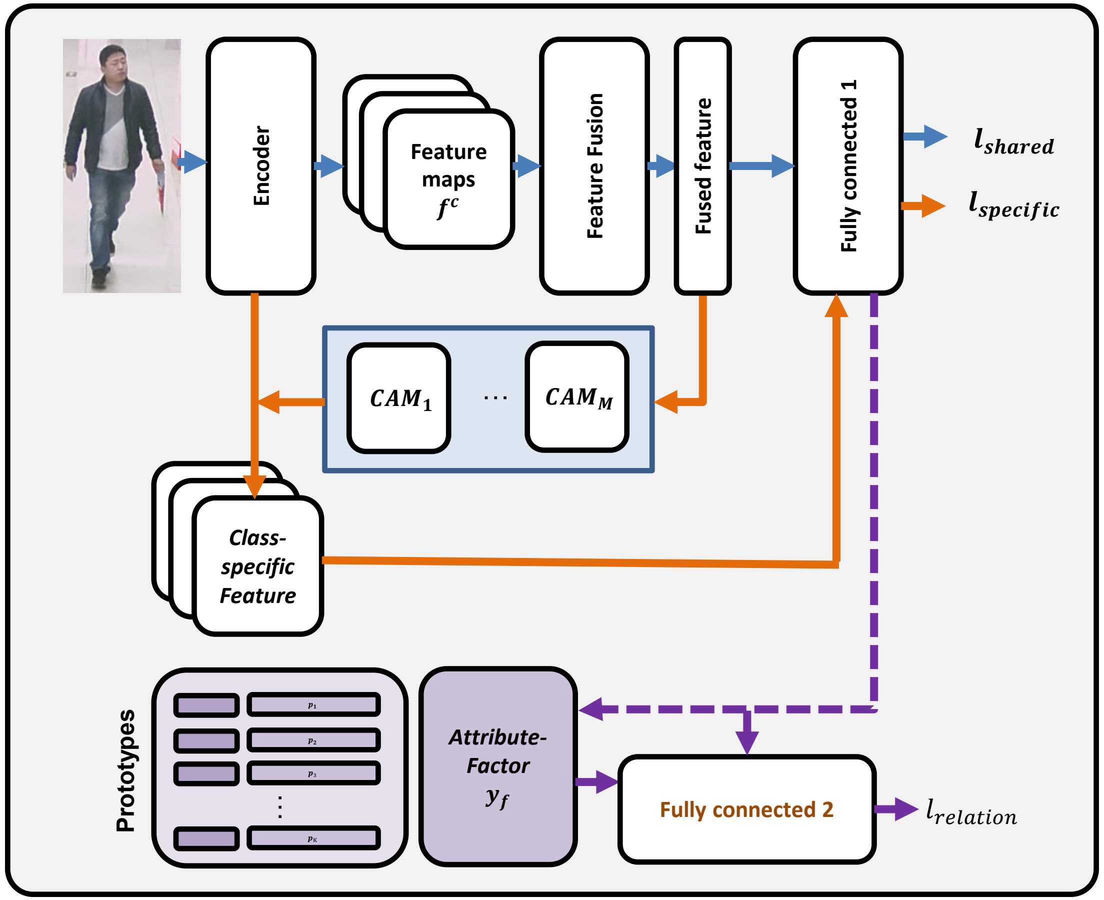


Fig. 4: Full framework of proposed methods

1. DAFL[1]: need extra learnable parameters for feature disentanglement.
⇒ **CAM-PAR** : use class activation map for feature disentanglement.

2. Attribute relation aware methods: only consider pairwise relationship between attributes
⇒ **CFAR** : Use collaborative filtering to model correlation of attribute sets

3. Adopted feature fusion strategy.

Methods: CF-PAR

Class activation map guided pedestrian attribute learning (CF-PAR)

Extract attribute-specific feature vectors using class activation map (CAM).

⇒Can achieve feature disentanglement with no need for extra module/parameters for OFOA mechanism.

1. **CAM generation for j-th attribute in i-th input image.**

$$A_{i,j} = w_j^T F_l(x_i) \quad (8)$$

$$CAM_{i,j}(x_i) = \frac{ReLU(A_{ij})}{max(ReLU(A_{ij}))} \quad (9)$$

where $F_l(\cdot)$ is last layer of the encoder and w is classifier weight.

UNIST

Methods: CF-PAR

2. Feature disentanglement using CAM

We can get attribute specific feature $f_{d_{i,j}}$ as follows:

$$f_{d_{i,j}} = CAM_{i,j}(x_i) \otimes f_{s_i} \tag{10}$$

where $f_i^c = F_l(x_i)$, denote encoder feature from the last layer before GAP.
The prediction process goes as:

$$l_{specific} = w^T \times GAP(f_d) \tag{11}$$

$$l_{shared} = w^T \times GAP(f^c) \tag{12}$$

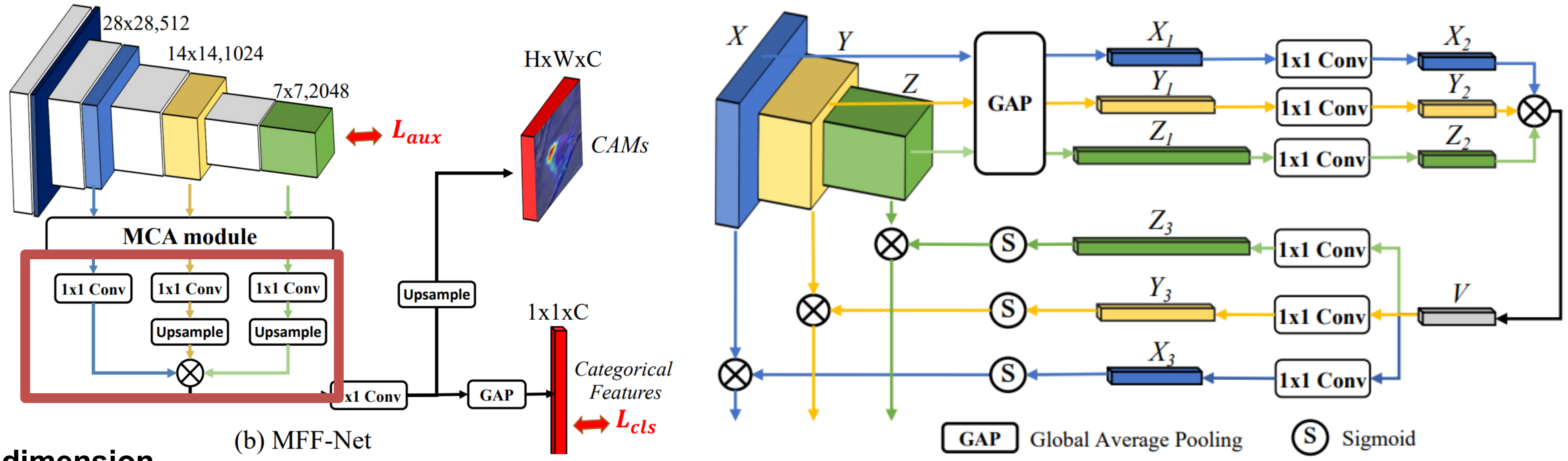
Since CAM itself is a result of multiplication between the encoder feature and classifier weight, we both minimize the loss for classification from $l_{specific}$ and l_{shared}.

$$Loss_1 = BCE(l_{specific}, y) + BCE(l_{shared}, y) \tag{13}$$

Where BCE is binary cross-entropy loss, y is ground truth label.

Methods: feature fusion

Adopt feature fusion strategy from [7] with minor modification to get
(1) more sophisticated CAM for feature disentanglement
(2) rich semantic information from different levels of features



Feature dimension after 1x1 Convolution:
Original paper : 128
In this work : 2048

Fig. 5: MFF-Net and MCA module from [7]

Methods: CFAR

Collaborative Filtering for Attribute Recognition (CFAR)

⇒ Effectively estimate the confidence of attributes based on a correlations between attribute sets.

⇒ Consider *pedestrian images* and *attributes* as *user* and *item* terms in collaborative filtering and aim to predict missing attributes.

CF for Attribute Recognition where $\boldsymbol{r_{i,j}} = \mathrm{l}_{specific_{\boldsymbol{i,j}}}$ from training phase**,**

$$p_{i,j} = \begin{cases} 1 & if\ \sigma(r_{i,j}) > p^t \\ 0 & if\ \sigma(r_{i,j}) \le p^t \end{cases}$$

$$c_{i,j} = 1 + \alpha\sigma(r_{i,j})$$

$$\min(x_f, y_f) \sum_{i,j} c_{i,j} \left(p_{i,j} - x_{f_i}^T y_{f_j}\right)^2 + \lambda\left(\sum_i ||x_{j_i}||^2 + \sum_j ||\, y_{j_j}\, ||^2\right)$$

Attribute confidence prediction,
$\hat{r} = \mathrm{x}_{f_i}^{\mathrm{T}} \mathrm{y}_{f_j}$
where $x_f \in R^{N \times C}$ is image-factor and $y_f \in R^{M \times C}$ is attribute-factor

Learned image and attribute factors later used in second step of training and inference.

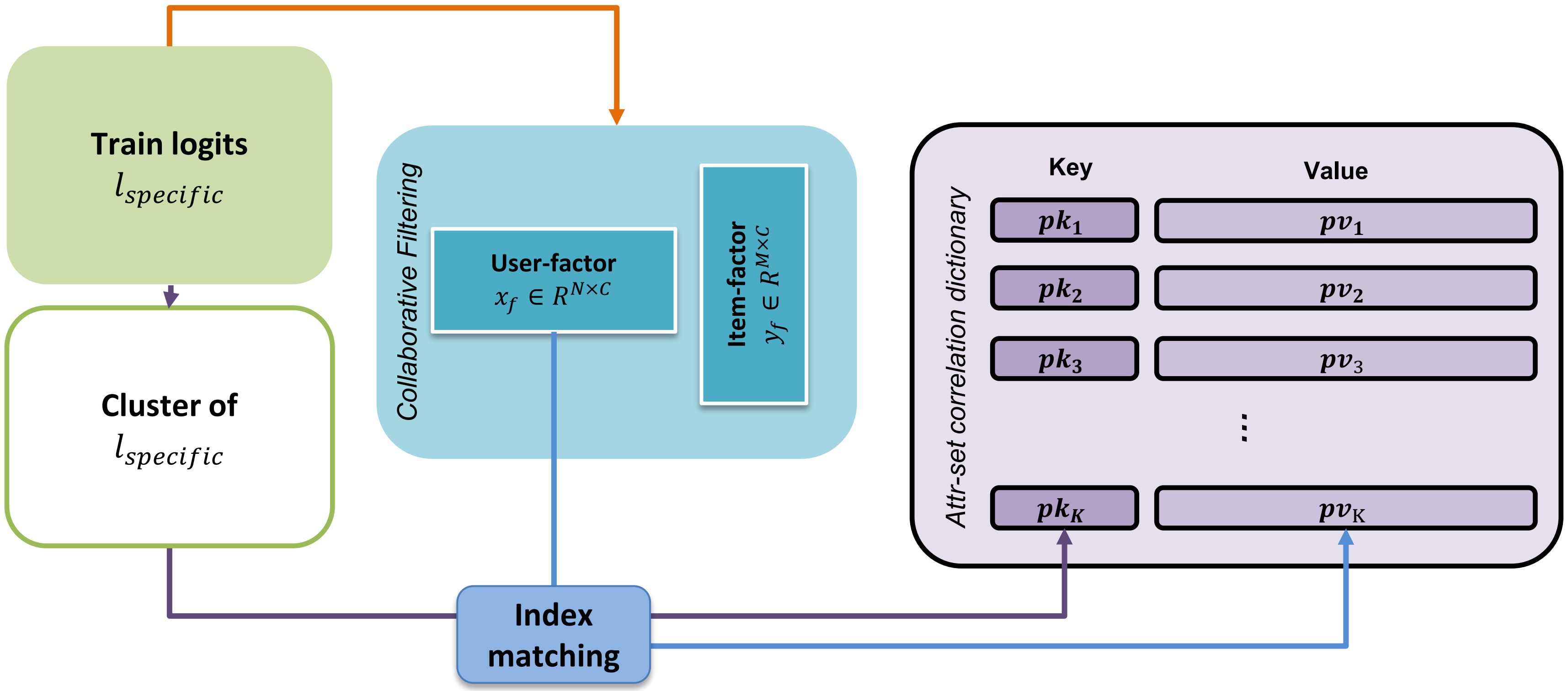
UNIST

Methods: CFAR

Attribute-set correlation dictionary

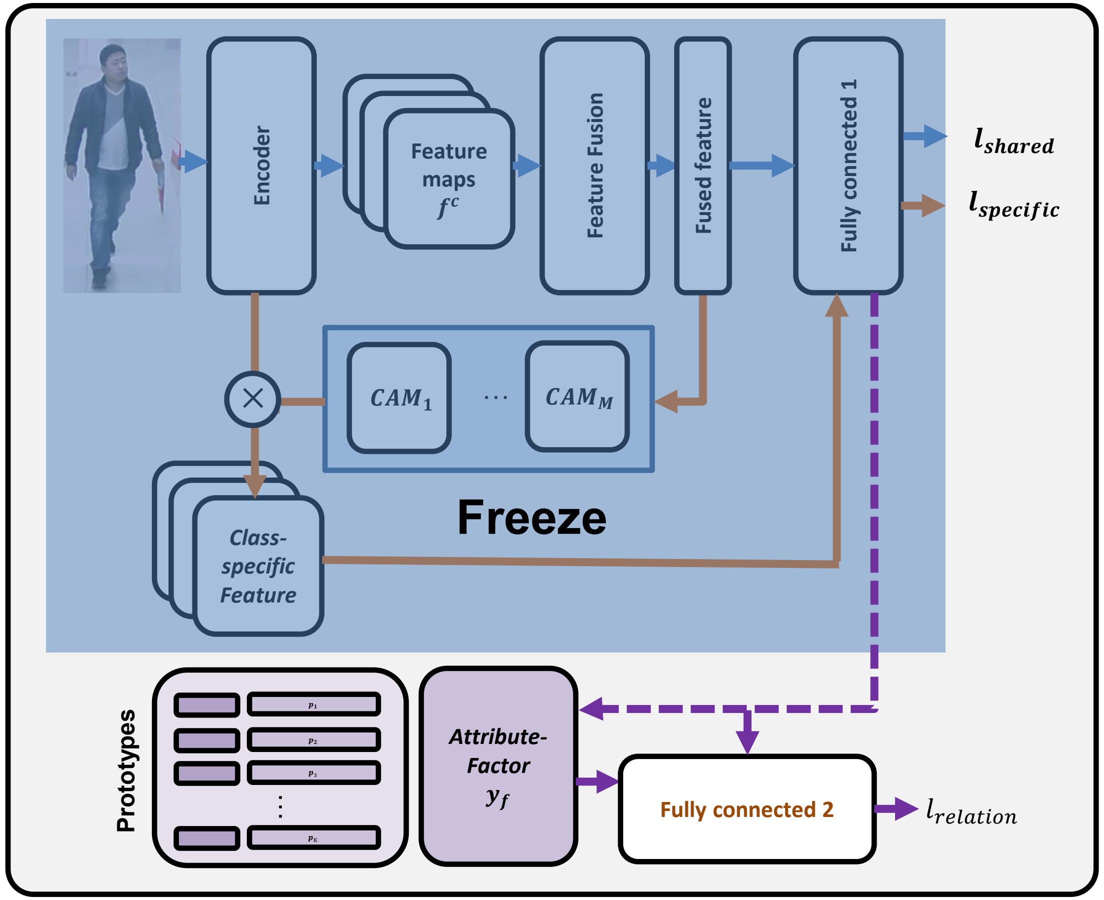
In the test phase, we should minimize the cost function for every test images.

⇒ To avoid this, **we construct attribute-set dictionary that can be used in test phase.**



Methods: CFAR

Training second fully connected layer with attribute set correlation dictionary



Second phase of training:
Freeze all other encoder and modules except second fully connected layer.

Train the second fully connected layer minimizing the loss below:

$$Loss = BCE\left(fc_2\left(concat\left(\hat{r}, l_{specific}\right)\right), y\right)$$

$$\hat{r} = pk^T y_f$$

where pk is the nearest key from $l_{specific}$

UNIST

Experiment Results

Dataset and metrics

Used Datasets : PA100K and RAPv1

Evaluation metrics

$$mA = \frac{1}{2M}\sum_{i=1}^{M}(\frac{TP_i}{P_i} + \frac{TN_i}{N_i})$$

$$Prec = \frac{1}{N}\sum_{i=1}^{N}(\frac{|Y_i \cap f(x_i)|}{|f(x_i)|})$$

$$Accu = \frac{1}{N}\sum_{i=1}^{N}(\frac{|Y_i \cap f(x_i)|}{|Y_i \cup f(x_i)|})$$

$$Recall = \frac{1}{N}\sum_{i=1}^{N}(\frac{|Y_i \cap f(x_i)|}{|Y_i|})$$

$$F1 = \frac{2 * Prec * Recall}{Prec + Recall}$$

UNIST

Comparison to previous works

Method	Backbone	PA100K			RAPv1		
		mA	Accu	F1	mA	Accu	F1
DeepMAR [13]	CaffeNet	72.70	70.39	81.32	73.79	62.02	75.56
HPNet [11]	InceptionNet	74.21	65.39	82.53	76.12	76.13	78.05
PGDM [26]	CaffeNet	82.97	73.08	85.76	74.31	64.57	77.35
LGNet [27]	Inception-V2	76.96	75.55	85.04	78.68	68.00	80.09
ALM [28]	BN-Inception	80.68	77.08	86.46	81.87	68.17	80.16
Baseline [10]	ResNet50	79.38	78.56	86.55	78.48	67.17	78.94
DAFL [4]	ResNet50	83.54	80.13	88.09	83.72	-	80.29
Our work	ResNet50	82.45	79.66	87.56	82.08	67.32	79.48

Table. 1: Comparison to previous works

Ablation Study

Method			RAPv1	
CAM-PAR	Fusion	CFR	mA	F1
-	-	-	78.48	78.94
-	✓	-	79.31	80.09
✓	-	-	79.54	79.04
✓	✓	-	81.18	79.18
✓	✓	✓	82.08	79.48

Table 2: Experiment on components of our proposed methods on RAPv1

Ablation Study

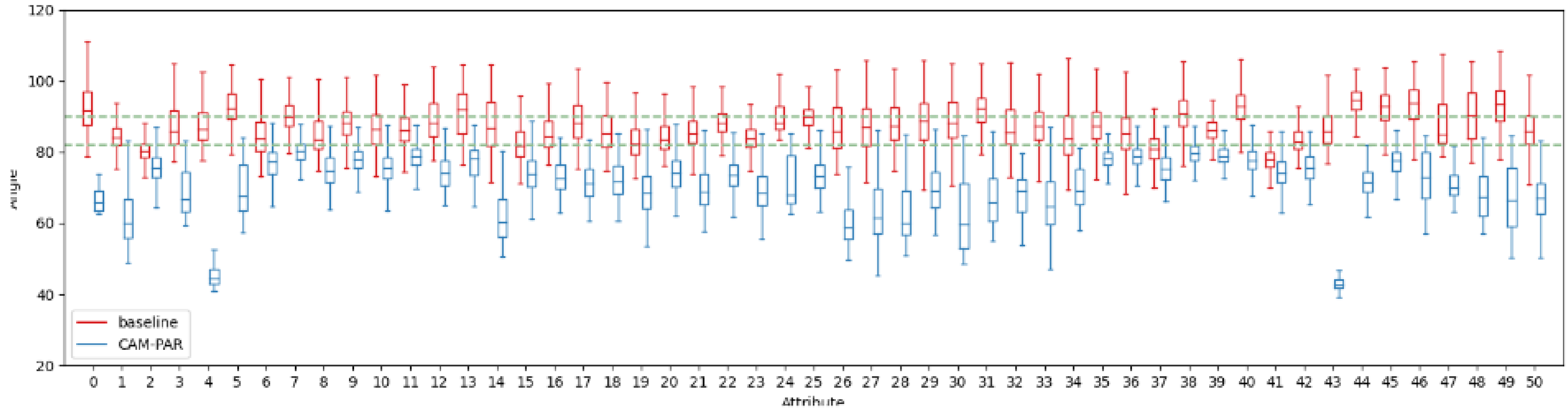


Figure 5: Box figure regarding the angle between feature map and classifier weights of baseline(blue) and our proposed methods(red) on RAPv1 dataset. Two dashed lines mark the decision boundary and theoretical optimal angle.

Conclusion

- Reviews the previous works regarding feature disentanglement for pedestrian attribute recognition tasks
- Proposes a novel approach that utilizes CAM-guided disentangled features for the PAR task.
- Propose a CFAR that model the correlation among the attribute-set and exploit them for attribute prediction.
- Our proposed method outperforms the baseline on the RAPv1 and PA100K but shows inferior performance than DAFL, which follows OFOA mechanism like ours do.

References

[1] J. Jia, N. Gao, F. He, X. Chen, and K. Huang, "Learning disentangled attribute representations for robust pedestrian attribute recognition," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, no. 1, pp. 1069–1077, Jun. 2022. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/19991

[2] Zhao, Ruijie & Lang, Congyan & Li, Zun & Liang, Liqian & Wei, Lili & Feng, Songhe & Wang, Tao. (2022). Pedestrian attribute recognition based on attribute correlation. Multimedia Systems. 28. 1-13. 10.1007/s00530-022-00893-y.

[3] Fan, Haonan, et al. "Correlation graph convolutional network for pedestrian attribute recognition." *IEEE Transactions on Multimedia* 24 (2020): 49-60.

[4] Chen, Zhao-Min, et al. "Multi-label image recognition with graph convolutional networks." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019.

[5] Dunfang Weng, Zichang Tan, Liwei Fang, and Guodong Guo. 2023. Exploring attribute localization and correlation for pedestrian attribute recognition. Neurocomput. 531, C (Apr 2023), 140–150. https://doi.org/10.1016/j.neucom.2023.02.019

References

[6] Hu, Yifan, Yehuda Koren, and Chris Volinsky. "Collaborative filtering for implicit feedback datasets." *2008 Eighth IEEE international conference on data mining*. Ieee, 2008.

[7] Wei, Jun, et al. "Shallow feature matters for weakly supervised object localization." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021.