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

Hyo Jeong Lee

Graduate School of Artificial Intelligence

Ulsan National Institute of Science and Technology

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

A thesis/dissertation submitted to
Ulsan National Institute of Science and Technology
in partial fulfillment of the
requirements for the degree of
Master of Science

Hyo Jeong Lee

06.08.2023 of submission
Approved by

백 승렬

Advisor
Seungryul Baek

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

Hyo Jeong Lee

This certifies that the thesis/dissertation of Hyo Jeong Lee is approved.

06.08.2023 of submission

Signature

백 승렬

Advisor: Seungryul Baek

Signature

Kwang In Kim

Signature

Taehwan Kim

Signature

Name of Thesis Committee member #3

Signature

Name of Thesis Committee member #4

*three signatures total in case of masters

Abstract

As a sub-task of multi-label classification, a pedestrian attribute recognition (PAR) task aims to train a model to detect various attributes for a given image. To achieve better model performance, It is necessary to understand the characteristic of the pedestrian image. Inevitably, most of the pedestrian images have a low resolution because their source is from surveillance cameras, and it is known that some of the pedestrian attributes are highly correlated with each other. To reflect these characteristics, a number of previous methods are proposed. J.Jia *et al.,* propose disentangled attribute feature learning (DAFL) framework for robust training against noisy pedestrian images. DAFL disentangles one-shared encoder feature to attribute specific features using multi-head attention and achieves significant improvements in model performance. But as additional modules are used for disentanglement, the model becomes more complicated. To address this, we propose **Class Activation Map guided Pedestrian Attribute Recognition (CAM-PAR)** that disentangle features with no need for additional parameters and explore the use of class activation map in multi-label classification domain. On the other hand, other works focus on relations in pedestrian attributes and propose methods that utilize this prior to predicting attributes. But these previous works are limited to modeling pairwise correlation of pedestrian attributes. We propose a **Collaborative Filtering for Attribute Recognition (CFAR)** module that models correlation of *attribute sets* using collaborative filtering and utilizes it for attribute prediction. Experiments on PA100K and RAPv1 datasets show that our proposed model surpasses the baseline method and has achieved competitive results against previous state-of-the-art methods.

Contents

List of Figures

List of Symbols

N	number of input images
M	number of attributes
p_t	threshold value for classification
w	classifier weights
f	encoder feature
f_d	disentangled feature vector
r	user feedback
$\hat{r}$	estimated user feedback
p	known preferences of users
x	input image
y	image labels
c	confidence variable
$F_l(\cdot)$	last layer of encoder
K	number of clusters
pk	key of the attribute-set correlation dictionary
pv	value of the attribute-set correlation dictionary

I Introduction

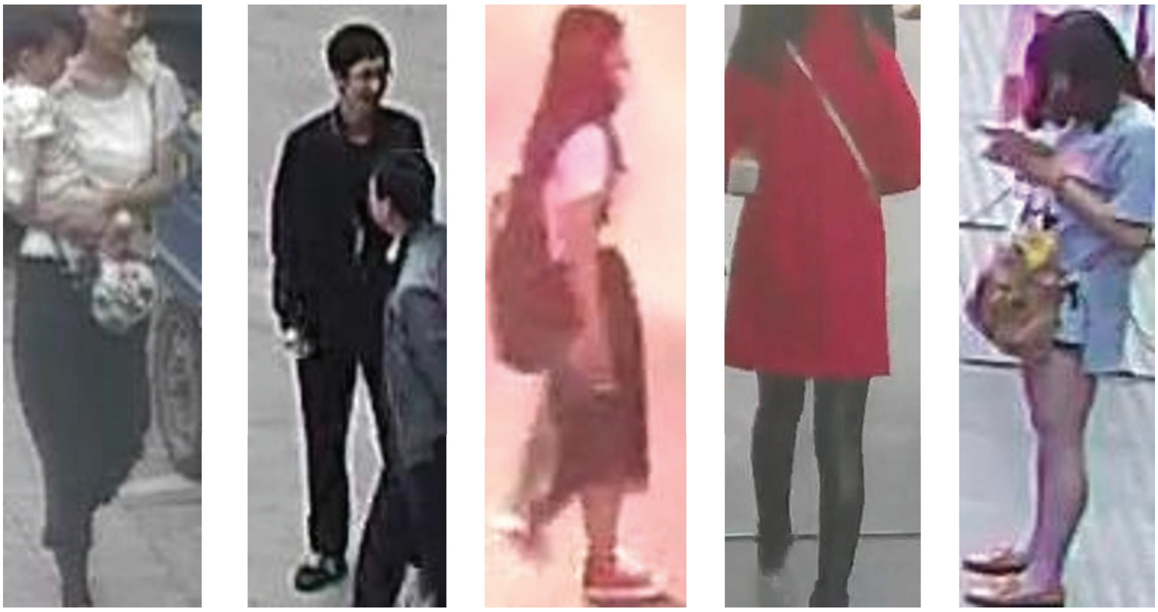


Figure 1: : some examples of pedestrian images in PA100K [1] dataset. Attributes of the first sample images can be *female, hand-Carry* and *short-sleeves.*

The pedestrian attribute recognition (PAR) task is a sub-task of multi-label classification that aims to detect various attributes for a given image. Type of pedestrian attributes can be appearances of pedestrians (e.g., gender, hairstyle, and age) and actions (e.g., holding, talking) that pedestrians take. PAR has drawn attention from the vision community and industry due to its wide application in scene understanding, person re-identification, and surveillance system.

The characteristics of pedestrian images can be summarized by two categories: (1) Most of the pedestrian datasets have a low resolution because their source is from surveillance cameras. (2) Some of the pedestrian attributes have a strong correlation. We'll review the previous works and their limitations that reflect these two characteristics in the PAR model and introduce our methods. As a sub-task of multi-label classification, most methods proposed for PAR task follow a similar mechanism of what multi-label classification takes [2], [3], [4], which is to predict multiple labels using a shared vector from an encoder. Despite this mechanism being adopted frequently in PAR tasks, [5] argue that this mechanism can harm the model robustness and mathematically proved its limitation. [5] further named this mechanism as One-shared-Feature-for-Multiple-Attribute (OFMA) and showed that it is impossible to get the optimal angle between classifier weights and one shared feature vector, which is a key factor in predicting logits when using the OFMA mechanism for multi-label classification. To tackle this problem, authors of [5] suggest Disentangled Attribute Feature Learning (DAFL) framework that classifies labels using attribute-specific feature vectors and corresponding multiple classifiers. Disentanglement is executed on a cascaded semantic-spatial cross-attention (SSCA) module, which learns semantic queries using multi-head attention. DAFL shows competitive performance compared to other methods on frequently used pedestrian datasets like RAP [6], PA100K [7] and PETA [8]. However, the series of additional SSCA module of DAFL makes the model more complicated and require a careful choice of a number of SSCA modules. To alleviate this, we used a class activation map (CAM) to disentangle features for pedestrian attribute learning without needing extra parameters to be learned. CAM-guided feature disentanglement procedure goes as follows: (1) Generate CAM using backbone encoder and classifier. (2) Disentangle features by multiplying CAM and encoder features. This particular procedure has been used in weakly supervised segmentation task domains [9], but utilizing CAM to disentangle features for multi-label classification or PAR tasks has not been investigated as far as we know. In addition, We employ a modified version of the feature fusion method introduced by [10] for better performance. Details of modifications we made will be described in a later section. On the other hand, previous works [11], [12], [13], [14] suggested models that consider pairwise relationships when predicting labels, but the correlation between *attribute-sets* per image has not been addressed. In our work, we emphasize the effect of attribute-set correlation on PAR task and propose a method for effectively modeling the attribute-set correlations using collaborative filtering.

The main contribution of this paper is as follows :

- We show CAM generated by the backbone can aid the model in learning discriminative attribute-specific features for a pedestrian attribute recognition task in a self-supervised manner.
- We propose a novel approach to model a correlation of attribute-sets using collaborative filtering for pedestrian attribute recognition task.
- Extensive experiments show that our model shows improvements compare to baseline on PA100K and RAP datasets.

The rest of this paper is organized as follows. Section 2 introduces the related works of pedestrian attribute recognition, weakly supervised semantic segmentation, and collaborative filtering methods. Section 3 revisit the limitations of the OFMA mechanism and its alternative for robust prediction. Section 4 proposes our approaches and conducted experiments in Section 5. Section 6 for the conclusion of this paper and discussing future directions.

II Related Works

2.1 Pedestrian Attribute Recognition

Various methods are proposed to learn the discriminative attribute feature of pedestrians. As the pedestrian attribute recognition task can be considered as a subtask of general multi-label classification, most PAR model [15], [7] follows a multi-label classification pipeline. Global image-based methods like [16], [17], [18] uses the whole pedestrian image as an input to predict attributes. [16] adopt a pre-trained AlexNet as encoder and multiple KL-loss (Kullback-Leibler divergence-based loss function) losses for each attribute. [18] propose multi-task CNN that shares visual knowledge via matrix decomposition based on the assumption that prior statistical feature information will aid the classifiers. These global image-based methods show decent performance with simple model structures but are limited to recognizing fine-grained attributes. To alleviate this limitation, part-based [19], [20], [21] and attention based [7], [22] methods are proposed. [20] conduct a body part division to capture the local characteristics of an attribute (e.g., jeans appear in the lower part of the body). Instead of using a body part detector, [20] divided a pedestrian image into 15 overlapping patches instead of using a body part detector to estimate pedestrian attributes because it is a challenging problem on its own. Instead of using fixed image patches, [21] proposes DeepCAMP, which learns discriminative patch groups via clustering. [7] exploited attention mechanism to localize multi-scale features from various levels of CNN encoder. Our proposed model is similar to [7] with multi-scaling but can be trained in an end-to-end manner compared to [7], which requires the model to be trained sequentially. And given the characteristic of pedestrian attributes that there are correlations between them, [11], [13] and [23] focus on modeling attribute relationships and use it as prior for pedestrian attribute recognition. [11] construct matrix that models the relationships between any pair of attributes in the attribute set via self-attention. [23] also construct a static label dependency matrix by calculating conditional probability among attributes. But most of them are limited to model correlations between attribute pairs, not attribute sets. To address this, we propose collaborative filtering for the attribute recognition model (CFAR) that effectively estimates the confidence of attributes based on latent correlations between attribute sets.

2.2 Weakly Supervised Semantic Segmentation/Object Localization

Weakly supervised semantic segmentation/object localization (WSSS/WSOL) tasks aim to segment an image using only weak supervision like image-level labels. Compared to full supervision, weak supervision can be obtained at a lower annotation cost. We review below a class activation map (CAM) based on WSSS/WSOL which is the most prominent method recently. Previous work [24] show that well-trained CNN on classification task can also localize objects in a single forward pass. The localization can be done by multiplying classifier weights and global average pooled features of the last layer of CNN. This procedure is called Class Activation Mapping (CAM). Generated CAM can be used as a pseudo ground truth map for the segmentation/localization model. Still, it may give the model incomplete and sparse pixel-level ground truth due to its characteristic highlighting the most discriminative region of the input image. To alleviate this, several methods [10], [9] are proposed to make CAM more complete. [10] propose the Shallow feature-aware Pseudo supervised Object Localization (SPOL) model for WSOL. [10] emphasize that shallow features of CNN have rich information about the detailed object boundaries, helpful to get the accurate initial CAM seed for WSOL. On the other hand, previous methods did not take full advantage of the shallow features due to their abundant background noise. [10] suggested MCA Module that fusion model feature from shallow to deep and effectively reduce the background noise. [9] observe cases that contain a considerable number of false positive and false negative pixels in CAM results generated by existing models. To this end, [9] proposed Class Re-Activation Maps (ReCAM) that employ softmax cross entropy (SCE) losses to converge CAM with binary cross entropy(BCE) loss. This way, the contrastive nature of SCE reduces the ambiguity within present class pixels, leading to more accurate CAM for WSSS. Our model borrows the concept of disentangling class-specific feature maps using CAM from [9] but is different in task objective. Using CAM for the pedestrian attribute recognition task has not been explored as far as we know.

2.3 Collaborative Filtering

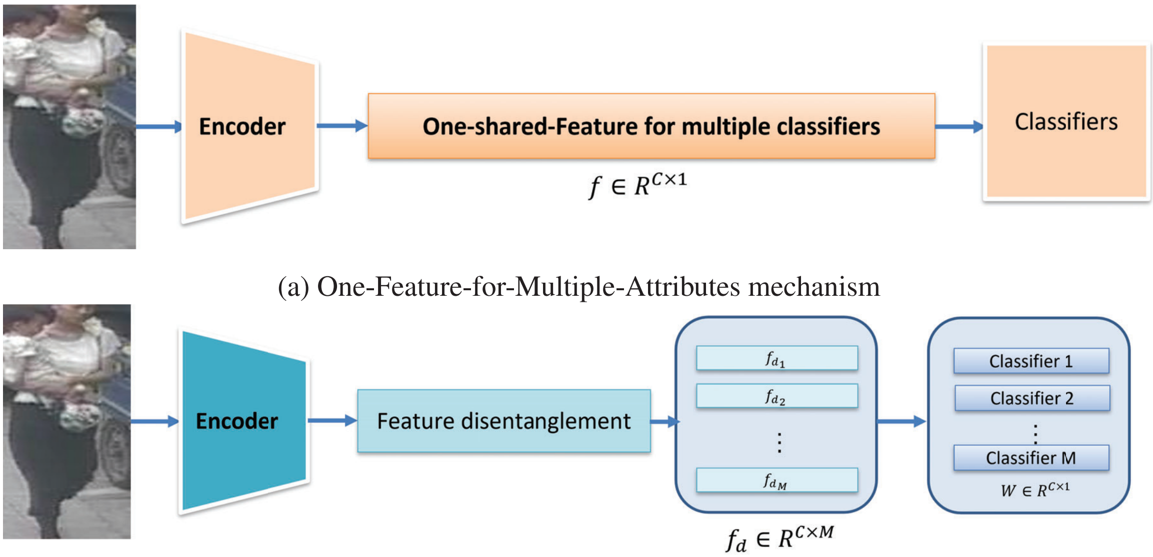
Recommender systems aim to provide users with predictions and recommendations of items [25]. To achieve this, various approaches like collaborative filtering, content-based filtering, and hybrid models have been proposed. In this section, we'll focus on collaborative filtering approaches. Collaborative Filtering (CF) is among the most popular approaches in the RS field and industry. The key assumption of CF is that if some user groups A and B have similar behaviors or rate items similarly, and hence will act or rate on other items similarly [26]. CF methods can be roughly divided into three categories: memory-based [27] [28], model-based [29] [30], and hybrid recommender systems that combine memory-based and model-based approaches. Memory-based CF methods utilize the entire user-item pairs database to predict recommendations. [31] uses singular value decomposition (SVD) to learn latent factors that well explain explicit user feedback(e.g., ratings, stars). [27] further expand SVD-based collaborative filtering methods to take implicit feedback (e.g., clicks, purchases) [27] introduce the concept of confidence level for implicit feedback datasets and alternatively optimize user-factors and item-factors matrix. In our approach, we model the correlations of attribute sets by using the CF method proposed by [27], considering the image sample and existing attributes as conventional user and item in CF.

III Preliminary

3.1 Multi-Label Classification

In this section, we review the limitation of the current multi-label classification scheme and stress the need for feature disentanglement. Previous works [32] and [5] named the feature learning mechanism that learns a shared vector to classify multiple labels as One-shared-Featrue-for-Multiple-Attributes (OFMA) mechanism. This OFMA mechanism is the most generic scheme for multi-label classification tasks. Still, authors of [32] and [5] argue that this mechanism harms the model performance in terms of robustness. This non-robustness degrades the overall model performance and can be severe in the case of PAR task given the characteristic of pedestrian images that have low-resolution.

OFMA Mechanism



(a) One-Feature-for-Multiple-Attributes mechanism

(b) One-Feature-for-One-Attributes mechanism

Figure 2: : general pipline of OFMA and OFOA mechanism for attribute recognition.

Given a dataset $D = \{x_i, y_j | i = 1, 2, ..., N\}$ and $y_j \in \{0,1\}^M$, where N and M indicate the number of train images and labels, x_i denotes i-th sample among input images. Multi-label classification aims to predict a binary label $\hat{y}_i$ when input image x_i is given. The binary values in the label indicate the presence and absence of the corresponding labels in the image.

For the j-th label in i-th sample image x_i, Prediction $\hat{y}_{i,j}$ is decided as follows with threshold p_t:

$$\hat{y}_{i,j} = \begin{cases} 1 & \text{if } p_{i,j} \geq p_t \\ 0 & \text{if } p_{i,j} < p_t \end{cases} \tag{1}$$

$$p_{i,j} = \sigma(logits_{i,j}) \tag{2}$$

where $\sigma(\cdot)$ is the sigmoid function and *logits* is the output value of the classifier. The *logits* for j-th attribute of i-th image is computed as:

$$logits_{i,j} = w_j^T f = |w_j| \cdot |f| \cdot cos\theta, \tag{3}$$

where $f \in R^D$ is one shared feature vector from encoder and $w \in R^{D \times M}$ is classifier weights. By taking Eq. 2 and Eq. 3 into Eq. 1, [32] and [5] state that prediction only depends on the angle θ between the feature vector and the classifier weight.

$$\hat{y}_{i,j} = \begin{cases} 1 & \text{if } 0^\circ \leq \theta \leq 90^\circ \\ 0 & \text{if } 90^\circ < \theta < 180^\circ \end{cases} \tag{4}$$

when $p_t = 0.5$. Therefore, the robust model should make the angle *theta* between the classifier and the feature vector small as possible if the target attribute is present in the input image.

Limitations of OFMA mechanism

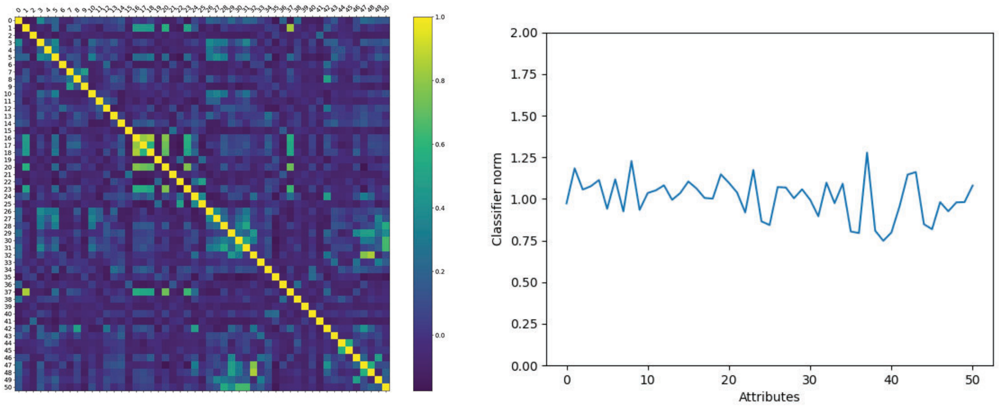


Figure 3: : The left plot shows the angle between classifiers, and the right plot shows the norm of classifier weights. The model is trained on RAPv1.

Previous works [32], [5] observe two crucial characteristics of a well-trained model that follows the OFMA mechanism: (1) Most classifier weights of attributes are orthogonal to each other, (2) Classifier norms are almost the same with little fluctuation. Therefore, in the case that all labels are present, it is impossible to make all of the angles between classifiers and one shared vector to 0°, which are expected to be small as possible for robust classification refer to Eq. 4.

3.2 OFOA mechanism

To alleviate this limitation of the OFMA mechanism, [5] propose One-specific-feature-for-One-attribute (OFOA) mechanism, which disentangles one shared feature to attribute specific features. By doing this, we can make the prediction with a larger margin from the decision boundary. [5] propose semantic-spatial cross-attention modules (SSCA) that disentangle one shared feature vector to attribute specific

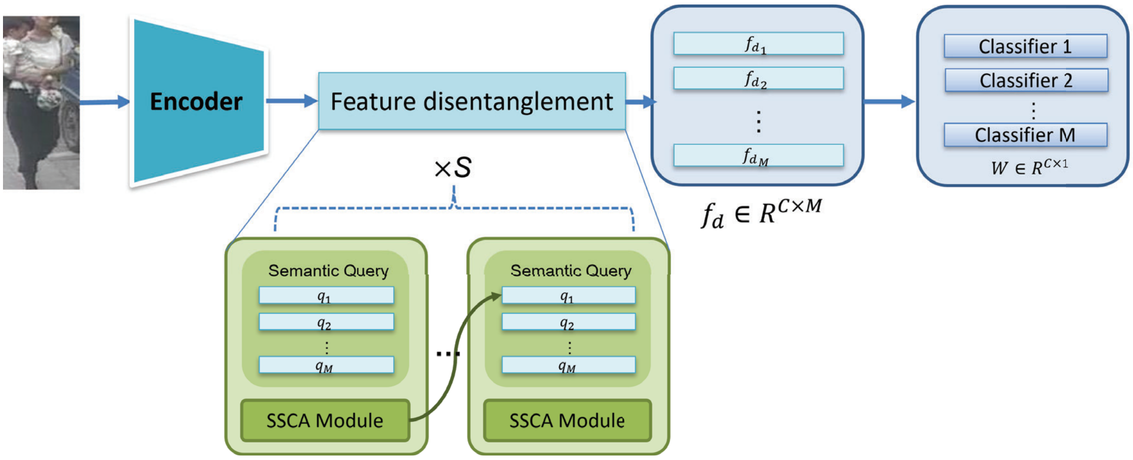


Figure 4: : above figure shows the OFOA mechanism and SSCA module for feature disentanglement proposed by DAFL. S denotes the number of SSCA modules that are used.

features. SSCA learns the semantic query for each attribute using multi-head attention and uses the semantic queries to classify attributes. The limitation of this method is that it needs additional series of SSCA modules for disentanglement and the performance of the model depends on the choice of S.

3.3 Collaborative Filtering

Collaborative filtering is one of the approaches to building recommender systems. Different from the content-based approach that uses user or product profiles (e.g., genre, actors), collaborative filtering (CF) utilizes past user behaviors (e.g., clicks, purchases, ratings). CF's main idea is to recommend items to a user u based on ratings or behavior of other users that acted similarly to u. Collaborative filtering achieved remarkable progress due to its scalability but often suffers from cold start problems.

There are two types of feedback from users, explicit and implicit feedback. The explicit feedback includes explicit ratings from users (e.g., star ratings, thumbs-up/down) and the implicit feedback that captures user behaviors (e.g., clicks, purchases).

Latent Factor Model for Explicit Datasets

One of the frequently used approaches for collaborative filtering is the latent factor model. The aim of the latent factor model is to learn a latent feature that well explains known user feedback. A number of works employ singular value decomposition (SVD) for extracting the latent factor of the user-item matrix that contains user feedback. Given user u, v and item i, j, known user feedback of u over i is expressed as $r_{u,i}$. For the stem of works that are based on SVD, the general process of learning latent features goes by minimizing the cost function below where $x_u \in R^D$ for the user-factor and $y_i \in R^D$ is the item factor.

$$min(x,y)\sum(r_{u,i} - x_u^T y_i)^2 + \lambda(||x_u||^2 + ||y_i||^2) \tag{5}$$

Latent Factor Model for Implicit Datasets

Authors of [27] identify the characteristics of implicit feedback as follows: (1) No negative feedback. The zero value of $r_{u,i}$ does not necessarily mean that user u does not prefer the item i. (2) Implicit feedbacks are noisy. value of r does not directly indicate the user's preference. For example, we cannot be certain that some individual likes the item he/she purchased if only information we know is the implicit behavior. (3) The numerical value of implicit feedback indicates confidence. As stated before, the value of implicit feedback is not directly connected to user preference, but it can give us insight into certain observations regarding confidence. To reflect this characteristic in Eq.5, [27] propose a model for implicit feedback with confidence variable.

First, we assume implicit feedback value $r_{u,i}$ as preference $p_{u,i}$. The $p_{u,i}$ is binarized $r_{u,i}$ and possible value 1, 0 indicates user u likes i or not correspondingly.

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

But the belief of the assumption can vary based on confidence levels. Therefore, an additional variable $c_{u,i}$ is introduced which measure our *confidence* in observing $p_{u,i}$.

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

where α is a hyper-parameter. With this confidence variable $c_{u,i}$, [27] expend Eq. 5 to account for the confidence levels. Therefore, the expended cost function for implicit user feedback is:

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

where, $\lambda(\sum_u ||x_u||^2 + \sum_i ||y_i||^2)$ is regularizing term for preventing overfitting and λ for hyper-parameter.

IV Methods

The full framework of our proposed method consists of two stages of learning: the first step for feature disentanglement and second step for utilizing correlation information of attribute sets. The second step of training is complementary, and can be omitted depending on the situation in which this proposed method is to be applied.

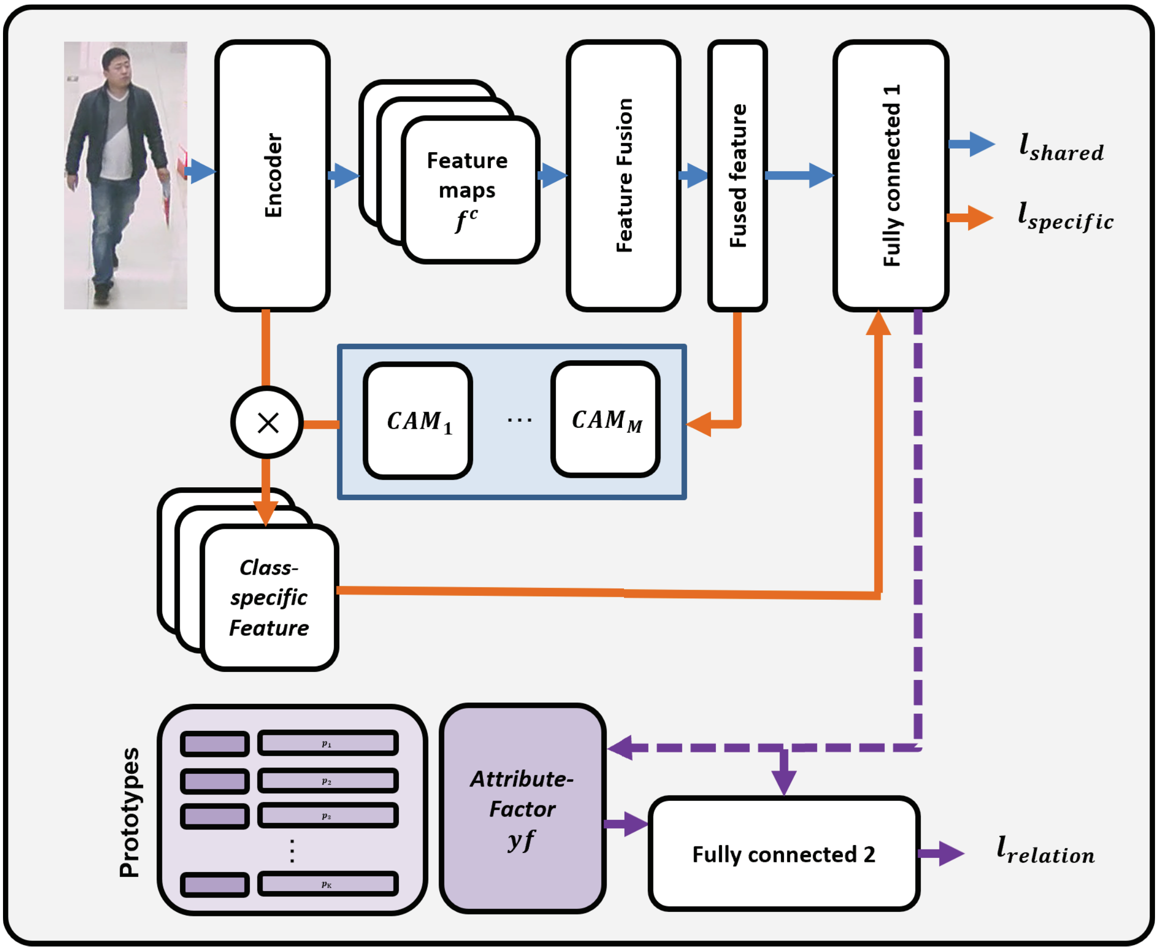


Figure 5: : Full framework of our proposed methods. The orange line shows the disentanglement process, and the purple dashed line is the second training phase that utilizes attribute set correlation information. Prototypes are generated in the training phase and stored for the second phase of training and final inference.

4.1 CAM guided PAR

As shown in [32], classifying multiple labels using one shared vector can harm the model's robustness. Inspired by these findings, we introduce CAM-guided disentanglement for pedestrian attribute recognition. To achieve this, we first generate the class activation map. Given i-th input image x_i, classifier weight $w \in R^{M \times C}$ and the last layer of encoder before global average pooling $F_l(\cdot)$, we obtain class activation map for j-th attribute as follows:

$$A_{i,j} = w_j^T F_l(x_i) \tag{9}$$

$$CAM_{i,j}(x_i) = \frac{ReLU(A_{i,j})}{max(ReLU(A_{i,j}))} \tag{10}$$

where $j = \{1, 2, ..., M\}$ and M for number of attributes and $CAM_{i,j}(x_i) \in R^{W \times H}$. Then we conduct element-wise multiplication between $CAM_{i,j}(x_i)$ and encoder feature $f^c = F_l(x)$ where $f_i^c \in R^{C \times W \times H}$.

$$f_{d_{i,j}} = CAM_{i,j}(x_i) \otimes f_i^c \tag{11}$$

Therefore, we obtain attribute-specific feature vectors $fd_{i,j}$ that contain disentangled information for each attribute. The prediction j-th attribute for i-th image using attribute-specific features calculated as:

$$l_{specific_{i,j}} = w_j^T \times GAP(f_{d_{i,j}}) \tag{12}$$

where $GAP(\cdot)$ denotes global average pooling. However, since CAM itself is a result of multiplication between the encoder feature and classifier weight, the model may struggles to converge. To resolve this, we both minimize the loss for classification using f_i^c and $f_{d_{i,j}}$. For the loss function, Binary Cross-Entropy (BCE) loss is adopted:

$$L(p, y) = -\sum_{J=1}^{M} (y_{i,j} log(p_{i,j}) + (1 - y_{i,j}) log(1 - p_{i,j})) \tag{13}$$

The loss for the first training phase is formulated as follows:

$$l_{specific_{i,j}} = w_{i,j}^T \times GAP(f_{d_{i,j}}) \tag{14}$$

$$l_{shared_{i,j}} = w_{i,j}^T \times GAP(f^c) \tag{15}$$

$$Loss_1 = L(l_{shared_{i,j}}, y) + L(l_{specific_{i,j}}, y) \tag{16}$$

With this approach, we can train the model via disentangled features with no additional parameter added. Compared to [32], which requires a series of spatial cross-attention modules, our model can achieve feature disentanglement more simply.

4.2 Feature Fusion

In this subsection, we introduce the feature fusion module from previous work [10] and its slight modification for the PAR task. As our model is compromised of methods from two main vision domain; WSSS and PAR, the proposed model can benefit from adopting a feature fusion strategy in both ways: (1) More sophisticated CAM for feature disentanglement. (2) Rich semantic information from different levels of features.

Authors of [10] propose the multiplicative feature fusion network (MFF-Net) to reduce the background noise of features derived from the shallow layers of CNN when generating CAM. The MCA module is denoted as Feature Fusion in Fig. 5. The MCA module combines feature maps after GAP from different levels of CNN to one shared latent vector V to couple various feature representations. Then, the shared latent vector V is transferred back to the original dimension, combined again by element-wise multiplication to reduce the background noise after 1×1 convolution that matches the dimension of feature vectors.

In the original MFF-Net, 1×1 convolution after the MCA module reduces the channel dimension to less than the original dimension of all participating features. This dimensionality reduction does not significantly affect the WSOL performance in the original paper since [10] employs a standalone classifier to predict the label. But in the PAR scenario, too small a number of feature dimensions can cause underfitting. Therefore, we modified the 1×1 convolution after the MCA module to match the dimension of feature vectors to the deepest feature vector that participates (e.g., 2048 in ResNet50). This fusion module can be located before CAM generation.

4.3 CF as Auxiliary Information

It is well known that some pedestrian attributes are correlated with each other. For example, the attribute "woman" and "skirt" are likely appear together but "man" and "skirt" are rarely together. Utilizing these correlations among attributes can further improve the performance of the PAR model. To address this, we propose a novel approach to the model distribution of attribute sets and use it as auxiliary information via collaborative filtering. In this approach, we consider pedestrian images and attributes as user and item terms in collaborative filtering and aim to predict missing attributes.

Given the input image x and attribute set y, the confidence of having j-th attributes in i-th input image x_i denoted by $r_{i,j} = \sigma(logit_{i,j})$ where $logit_{i,j}$ is the output value of the classifier from the first step of training. Since $r_{i,j}$ is an implicit feedback, we follow the collaborative filtering approach proposed by [27] to learn the latent factors, To learn the latent factors, we minimize the cost function below:

$$p_{u,i} = \begin{cases} 1 & \text{if } r_{i,j} \geq p^t \\ 0 & \text{if } r_{i,j} < p^t \end{cases}, \qquad c_{ui} = 1 + \alpha r_{i,j} \tag{17}$$

$$min(x,y) \sum_{i,j} c_{i,j}(p_{i,j} - x_{f_i}^T y_{f_j})^2 + \lambda (\sum_i \|x_{f_i}\|^2 + \sum_j \|y_{f_j}\|^2) \tag{18}$$

Where, $x_f \in R^{N \times D}$ is image-factors and $y_f \in R^{M \times D}$ is attribute-factors and N, M as number of users and attributes correspondingly. The learned latent factors are used in the second step of training and final inference.

Attribute-set Correlation Dictionary

Since minimizing the cost function for every new image in the test set is inefficient, we introduce a novel approach that constructs an Attribute-set correlation dictionary to incorporate the learned latent factor in the test phase. Among $r \in R^{N \times M}$, we employ k-means clustering and consider each centroid as prototypes and set as key pk in a dictionary and $pk \in R^{K \times D}$. And find the user factors in the same

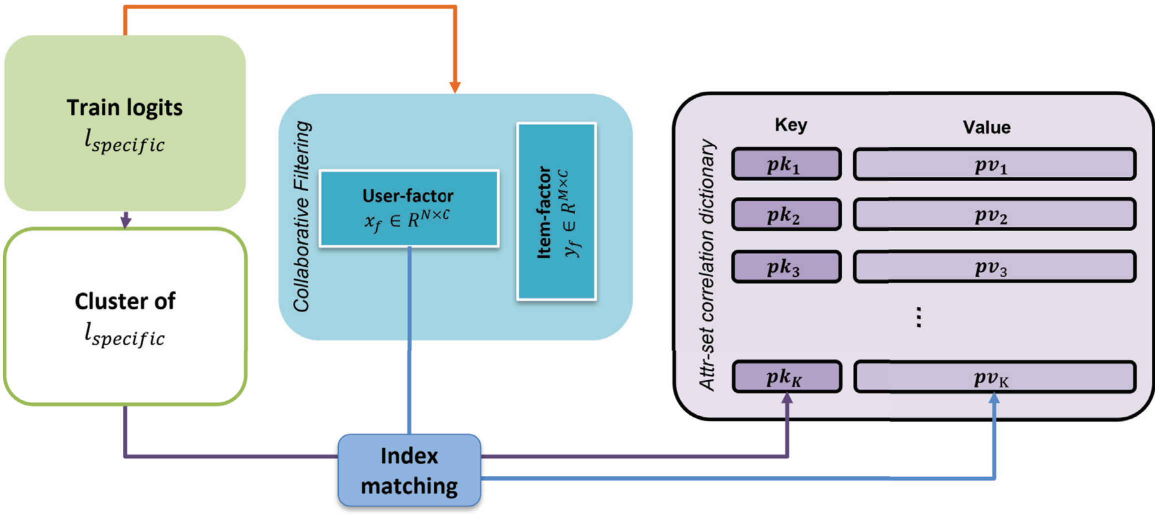


Figure 6: : pipeline of constructing attribute-set correlation dictionary. The $logits_{specific}$ is from the converged network from the first step of training.

cluster as the target prototype and set their mean as the value pv of the corresponding key. In the second training phase, We search the nearest pk from each $l_{specific}$ and use the prototype to infer confidence $\hat{r}$. For the i-th image x_i and j-th attribute y_j, confidence $\hat{r_{i,j}}$ is calculated as follows:

$$\hat{r}_{k,j} = p_k^T y_{f_j} \tag{19}$$

where pv_k is the value of the nearest prototype from $l_{specific_{i,j}}$. To utilize this confidence as auxiliary information for the PAR task, we concatenate $\hat{r}$ and $l_{specific}$ from the converged network from the first phase of training and feed them to another fully connected layer for the final prediction. We train the second fully connected layer by minimizing the loss function below:

$$Loss_2 = L(fc_2(concat(\hat{r}, l_{specific})), y) \tag{20}$$

where fc_2 is a fully connected layer. Using this method, we can model the correlations between attribute sets and use them for classification. Different from only modeling correlations of attribute pairs, reflecting correlations of attribute sets can help predict abstract attributes like a role (e.g., customer, clerk) because abstract attributes can be inferenced by context from multiple other attributes are giving. For example, if attributes "carrying" and "plastic bag" are present, "customer" attributes are likely to appear together, but if "carrying" and "baby stroller" are present, there is no strong correlation to the attribute "customer".

V Experiments

5.1 Evaluation Metrics and Datasets

In this section, we introduce datasets used to validate our method and criteria for evaluation. We conducted experiments in two popular pedestrian datasets;PA100K, RAPv1 and followed evaluation criteria widely used in PAR and multi-label classification problems. The criteria include label-based and instance-based metrics. For the instance-based metrics, accuracy, precision, recall, and F1 score are used. For label-based metrics, mean accuracy (mA) is used. The criteria are calculated as follows:

$$mA = \frac{1}{2M}\sum_{i=1}^{M}(\frac{TP_i}{P_i} + \frac{TN_i}{N_i}) \tag{21}$$

$$Accu = \frac{1}{N}\sum_{i=1}^{N}(\frac{|Y_i \cap f(x_i)|}{|Y_i \cup f(x_i)|}) \tag{22}$$

$$Prec = \frac{1}{N}\sum_{i=1}^{N}(\frac{|Y_i \cap f(x_i)|}{|f(x_i)|}) \tag{23}$$

$$Recall = \frac{1}{N}\sum_{i=1}^{N}(\frac{|Y_i \cap f(x_i)|}{|Y_i|}) \tag{24}$$

$$F1 = \frac{2 * Prec * Recall}{Prec + Recall} \tag{25}$$

where N, M denotes the number of input images, and M is the number of attributes. TP_i and TN_i are the number of correctly predicted positive and negative examples. P_i and N_i are the numbers of positive and negative examples. Y_i is the ground truth label of the i-th instance.

5.2 Implementation Details

In our work, we used ResNet50 as an encoder that extracts the pedestrian features. The encoder is firstly pre-trained on ImageNet and fine-tuned for pedestrian datasets. Hyperparameter for the training set to same as [15]. For feature fusion, we used the last three layers of ResNet50; the dimension for the fused feature vector is 2024.

5.3 Results

In this subsection, we compare the performance between our proposed approach and previous works and conduct an ablation study to validate the effect of the proposed methods; CAM-PAR, Feature fusion, and CFAR.

Comparision to the Earlier Works

Tab. 1 shows the experimental result of our proposed method and previous state-of-the-art methods on two popular benchmarks for the pedestrian attribute recognition task. Our methods surpass most of the previous state-of-the-art methods and outperform the baseline method by 3.07% and 3.6% mA on

Method	Backbone	PA100K			RAPv1		
		mA	Accu	F1	mA	Accu	F1
DeepMAR [17]	CaffeNet	72.70	70.39	81.32	73.79	62.02	75.56
HPNet [7]	InceptionNet	74.21	65.39	82.53	76.12	76.13	78.05
PGDM [33]	CaffeNet	82.97	73.08	85.76	74.31	64.57	77.35
LGNet [34]	Inception-V2	76.96	75.55	85.04	78.68	68.00	80.09
ALM [35]	BN-Inception	80.68	77.08	86.46	81.87	68.17	80.16
Baseline [15]	ResNet50	79.38	78.56	86.55	78.48	67.17	78.94
DAFL [5]	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: : experiment result of our proposed method and performance comparison with previous state-of-the-art methods on two commonly used datasets, PA100K [1] and RAPv1 [6]. Five metrics (mA, accuracy, precision, recall, and F1) are calculated for the evaluation.

Method			RAPv1	
CAM-PAR	Fusion	CFAR	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

RAPv1 and PA100K datasets. But our method shows inferior model performance in all aspects than DAFL, which adopts the same OFOA mechanism as ours. Although our approach is more lightweight, the performance gap is quite large, Therefore, it seems necessary to exploit various losses adopted by the DAFL [5] method in addition to conducting disentanglement over features.

Ablation Study

In this section, we discuss the effect of the proposed modules for the mA and F1, CAM-PAR, feature fusion strategy, and CFAR. As we can see in Tab. 2, applying CAM-PAR to baseline achieved 1.06% and 0.56% performance increase. And feature fusion alone for 0.83% and 1.61%. Adopting both methods to baseline further increases the performance by 2.7% and 0.7%. In the end, adding CFAR to CAM-PAR that adopted a feature fusion strategy made improvements by 3.6% and 1.0% from the baseline via experiments. Meanwhile, the F1 score is the highest among all experiments shown in Tab. 2 when we only apply the feature fusion strategy on the baseline. This is because the F1 score is a metric for instances. The fused encoder feature contains rich entangled information from different layers; thus, it gives the classifier the semantic information of the given input sample. The experiments in which all modules are combined show higher mA than only the feature fusion strategy used. This is because disentanglement by CAM-PAR and correlation prior given by CFAR provide information centered on attributes, which can perform better at metrics for labels like mA.

Then we validate the argument that our model using CAM-guided disentangled feature for the PAR task is more robust than the baseline, which follows the general OFMA mechanism. According to Eq. 4, we can tell the model is robust for the positive attributes when the angle θ between the feature vector and the classifier weight is close to 0°. In Fig. 7, we plot the θ of baseline and our proposed model, decision boundary, and optimal angle stated in [5] for the RAPv1. We can see that most θ values from our proposed method are closer to 0° than the value from the baseline.

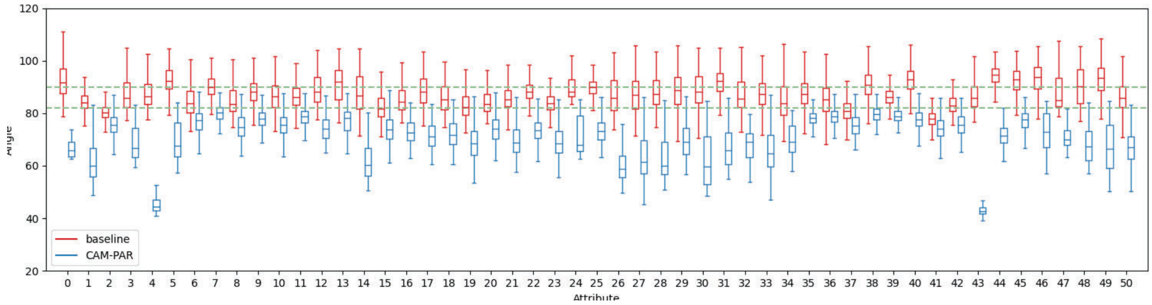


Figure 7: Box figure regarding the angle between feature map and classifier weights of baseline(blue) and our proposed methods(red) on RAPv1 dataset. Two dashed lines mark the decision boundary and theoretical optimal angle.

VI Conclusion

This paper reviews the previous works regarding feature disentanglement for pedestrian attribute recognition tasks and proposes a novel approach CAM-PAR that utilizes a class activation map to disentangle one shared encoder feature to attribute features for the PAR task. Different from previous works, CAM-PAR can disentangle the encoder feature with no need for extra parameters to be learned and it is the first approach that exploits CAM for PAR task. And we also propose CFAR that models correlation of attribute sets using collaborative filtering. Our proposed method outperforms the baseline on the RAPv1 and PA100K.

References

[1] X. Liu, H. Zhao, M. Tian, L. Sheng, J. Shao, J. Yan, and X. Wang, "Hydraplus-net: Attentive deep features for pedestrian analysis," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 1–9.

[2] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, and W. Xu, "CNN-RNN: A unified framework for multi-label image classification," *CoRR*, vol. abs/1604.04573, 2016. [Online]. Available: http://arxiv.org/abs/1604.04573

[3] T. Chen, M. Xu, X. Hui, H. Wu, and L. Lin, "Learning semantic-specific graph representation for multi-label image recognition," 2019.

[4] D. Ungarbayev, O. Demirel, and M. T. Akhtar, "Automatic data augmentation method with improved interpretability for image classification in computer vision applications," in *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2022, pp. 1356–1361.

[5] 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

[6] D. Li, Z. Zhang, X. Chen, H. Ling, and K. Huang, "A richly annotated dataset for pedestrian attribute recognition," *CoRR*, vol. abs/1603.07054, 2016. [Online]. Available: http://arxiv.org/abs/1603.07054

[7] X. Liu, H. Zhao, M. Tian, L. Sheng, J. Shao, S. Yi, J. Yan, and X. Wang, "Hydraplus-net: Attentive deep features for pedestrian analysis," in *2017 IEEE International Conference on Computer Vision (ICCV)*. Los Alamitos, CA, USA: IEEE Computer Society, oct 2017, pp. 350–359. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/ICCV.2017.46

[8] Y. DENG, P. Luo, C. C. Loy, and X. Tang, "Pedestrian attribute recognition at far distance," in *Proceedings of the 22nd ACM International Conference on Multimedia*, ser. MM '14. New York, NY, USA: Association for Computing Machinery, 2014, p. 789–792. [Online]. Available: https://doi.org/10.1145/2647868.2654966

[9] Z. Chen, T. Wang, X. Wu, X.-S. Hua, H. Zhang, and Q. Sun, "Class re-activation maps for weakly-supervised semantic segmentation," 2022.

[10] J. Wei, Q. Wang, Z. Li, S. Wang, S. K. Zhou, and S. Cui, "Shallow feature matters for weakly supervised object localization," 2021.

[11] R. Zhao, C. Lang, Z. Li, L. Liang, L. Wei, S. Feng, and T. Wang, "Pedestrian attribute recognition based on attribute correlation," *Multimedia Systems*, vol. 28, pp. 1–13, 06 2022.

[12] Z. Tan, Y. Yang, J. Wan, G. Guo, and S. Z. Li, "Relation-aware pedestrian attribute recognition with graph convolutional networks," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, pp. 12 055–12 062, Apr. 2020. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/6883

[13] H. Fan, H.-M. Hu, S. Liu, W. Lu, and S. Pu, "Correlation graph convolutional network for pedestrian attribute recognition," *IEEE Transactions on Multimedia*, vol. 24, pp. 49–60, 2022.

[14] D. Weng, Z. Tan, L. Fang, and G. Guo, "Exploring attribute localization and correlation for pedestrian attribute recognition," *Neurocomputing*, vol. 531, pp. 140–150, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231223001583

[15] J. Jia, H. Huang, W. Yang, X. Chen, and K. Huang, "Rethinking of pedestrian attribute recognition: Realistic datasets with efficient method," 2020.

[16] P. Sudowe, H. Spitzer, and B. Leibe, "Person attribute recognition with a jointly-trained holistic cnn model," *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, pp. 329–337, 2015.

[17] D. Li, X. Chen, and K. Huang, "Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios," in *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, 2015, pp. 111–115.

[18] A. H. Abdulnabi, G. Wang, J. Lu, and K. Jia, "Multi-task cnn model for attribute prediction," *Trans. Multi.*, vol. 17, no. 11, p. 1949–1959, nov 2015. [Online]. Available: https://doi.org/10.1109/TMM.2015.2477680

[19] L. Bourdev and J. Malik, "Poselets: Body part detectors trained using 3d human pose annotations," in *International Conference on Computer Vision (ICCV)*, 2009. [Online]. Available: http://www.eecs.berkeley.edu/~lbourdev/poselets

[20] J. Zhu, S. Liao, D. Yi, Z. Lei, and S. Z. Li, "Multi-label cnn based pedestrian attribute learning for soft biometrics," in *2015 International Conference on Biometrics (ICB)*, 2015, pp. 535–540.

[21] L. Yang, L. Zhu, Y. Wei, S. Liang, and P. Tan, "Attribute recognition from adaptive parts," 2016.

[22] N. Sarafianos, X. Xu, and I. A. Kakadiaris, “Deep imbalanced attribute classification using visual attention aggregation,” 2018.

[23] Z.-M. Chen, X.-S. Wei, P. Wang, and Y. Guo, “Multi-label image recognition with graph convolutional networks,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 5177–5186.

[24] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, “Learning deep features for discriminative localization,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, jun 2016, pp. 2921–2929. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/CVPR.2016.319

[25] J. Bobadilla, F. Ortega, A. Hernando, and A. Gutiérrez, “Recommender systems survey,” *Knowledge-Based Systems*, vol. 46, pp. 109–132, 2013. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950705113001044

[26] K. Goldberg, T. Roeder, D. Gupta, and C. Perkins, “Eigentaste: A constant time collaborative filtering algorithm,” *Information Retrieval*, vol. 4, pp. 133–151, 07 2001.

[27] Y. Hu, Y. Koren, and C. Volinsky, “Collaborative filtering for implicit feedback datasets,” in *2008 Eighth IEEE International Conference on Data Mining*, 2008, pp. 263–272.

[28] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme, “Bpr: Bayesian personalized ranking from implicit feedback,” in *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, ser. UAI ’09. Arlington, Virginia, USA: AUAI Press, 2009, p. 452–461.

[29] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, and T.-S. Chua, “Neural collaborative filtering,” 2017.

[30] Q.-T. Truong, A. Salah, and H. W. Lauw, “Bilateral variational autoencoder for collaborative filtering,” in *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, ser. WSDM ’21. New York, NY, USA: Association for Computing Machinery, 2021, p. 292–300. [Online]. Available: https://doi.org/10.1145/3437963.3441759

[31] R. Bell, Y. Koren, and C. Volinsky, “Modeling relationships at multiple scales to improve accuracy of large recommender systems,” in *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD ’07. New York, NY, USA: Association for Computing Machinery, 2007, p. 95–104. [Online]. Available: https://doi.org/10.1145/1281192.1281206

[32] M. Hassanin, I. Radwan, S. Khan, and M. Tahtali, “Learning discriminative representations for multi-label image recognition,” *Journal of Visual Communication and Image Representation*, vol. 83, p. 103448, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1047320322000116

[33] D. Li, X. Chen, Z. Zhang, and K. Huang, "Pose guided deep model for pedestrian attribute recognition in surveillance scenarios," in *2018 IEEE International Conference on Multimedia and Expo (ICME)*, 2018, pp. 1–6.

[34] P. Liu, X. Liu, J. Yan, and J. Shao, "Localization guided learning for pedestrian attribute recognition," 2018.

[35] C. Tang, L. Sheng, Z. Zhang, and X. Hu, "Improving pedestrian attribute recognition with weakly-supervised multi-scale attribute-specific localization," 2019.

Acknowledgements

First of all, I would like to express my deep gratitude to Prof. Kwang-in Kim for being very considerate during the master's course so that I can concentrate on my research. In addition, I really appreciate the insightful guidance Prof. Kwang-in Kim gave me. Without this, there would have been a lot of difficulty in finishing this paper. I am also grateful to my lab colleagues for giving me a lot of advice from their experiences. Their comments were a lot of help and encouragement to me, who was writing an academic paper for the first time.