# A Human Retrieval System based on Human Attribute Ontology and Deep Multi-task Neural Network

Trang T.T. Phung[1],[2], Ngoc Q. Ly[1], Fukuzawa Masayuki[3]
[1] Faculty of Information Technology, HCM University of Science –VNUHCM, Vietnam.
[2] Saigon University, HCM City, Vietnam
[3] Kyoto of Institue of Technology (KIT), Japan
Correspondence: Trang T.T. Phung, trangphung@sgu.edu.vn
Communication: received 31 Jan 2024, revised 12 Mar 2024, accepted 05 Apr 2024
Digital Object Identifier: 10.32913/mic-ict-research.v2024.n2.1255

**Abstract: The goal of this research is to enhance the capability of image retrieval systems to understand images more effectively. We present a model designed for searching human objects (such as pedestrians or persons) within expansive image datasets. Our unique approach involves developing an image retrieval system that incorporates attribute learning and the Human Attribute Ontology (HAO). This research offers several key contributions: (1) The development of the Human Attribute Ontology (HAO) which serves as a repository for storing prior knowledge about images. Thanks to its hierarchical structure, this ontology facilitates the reuse of prior knowledge, optimizing the subsequent stages of attribute learning and image retrieval; (2) The implementation of a Convolutional Neural Network (CNN) to spearhead attribute learning, leveraging the HAO to enhance accuracy; (3) The creation of a Human Image Retrieval system that utilizes both attribute learning and the HAO. Our system delves deeper by understanding images at the attribute level, highlighting the advantages of harnessing the ontology to reuse existing knowledge. The efficacy of our methodology is validated through experiments on benchmark datasets like PETA and Pa100k achieving state-of-the-art results.**

**Keywords:** *Image Retrieval, Object Retrieval, Human Attribute Ontology, Attribute Learning, Deep Learning, Deep feature learning.*

## I. INTRODUCTION

Image retrieval (IR) is a useful application for people in many aspects from daily life to special business. Studies about IR have been developed in methods for retrieval systems with more and more variety and popularity. Human image retrieval (HR) has an important role in surveillance and security [1, 2]. Recently, the approaches for building person retrieval systems are popular such as part-based [3, 4], graph-based [5–7], attribute-based [8–13], and text-based [14, 15], and used deep learning models [16] for feature extraction or fine-tuning models from famous CNNs such as ResNet [17], VGG [14], and EfficientNet [18]. In the semantic image approach, models have used ontology [19] for building retrieval systems, such as facial attribute retrieval with FAO (face attribute ontology) [20], and Bio-FAO (Bio-inspired Facial Aesthetic Ontology)[21].

Human image retrieval systems have met some challenges in images such as motion blur, occlusion, pose, camera angles, lighting conditions, and other kinds of image degradation.

Recently, with The Industry Revolution 4.0, intelligent products have been developed for society's demand. It is also a big challenge. How to solve the challenges is a difficult question.

Using attribute recognition for image retrieval will help the retrieval be more intelligent because it could find out attributes of objects in images instead of looking at objects only. For example, finding a person with the attribute "red shirt", the recognition will help the retrieval can detect the attribute "red shirt" instead of the shirt only.

Ontology is a tool to represent knowledge for sharing and reusing [17]. Using ontology for organizing and storing prior knowledge of images will help image retrieval systems at the semantic level. With the information on images in ontology, we can gain relevant information on images such as attributes of objects, relations between objects and objects as intra-relations, and attributes and objects as inter-relations. On the other hand, ontology supports the reuse of prior knowledge without learning (training and testing) or re-learning by relations in ontology.

In this paper, we try to build an HR by using three main tasks: (1) Designing the HAO for organizing and storing prior knowledge of images, (2) proposing a new Attribute Learning method based on HAO, and (3) building a human image retrieval based on attribute learning and HAO. The model has solved some challenges in image understanding and closed the gap in image semantics as follows: First, improving attribute learning to increase accuracy and more intelligence for the retrieval system. Second, building the HAO to support attribute learning and reusing prior knowledge, more understanding of images at the semantic level, and effective performance for the retrieval system.

In short, our contribution to this topic is as follows:

- Building a new ontology named: Human Attribute Ontology (HAO) for organizing and storing information of human object attributes.
- Proposing an Attribute Learning method based on deep neural network and HAO.
- Building a new human attribute retrieval system based on HAO and Attribute Learning.
- The results of the proposal gained good results on the benchmark datasets PETA, and Pa100k.

## II. RELATED WORKS

### 1. Deep Attribute Learning

Visual attributes are generally defined as mid-level semantic visual concepts [22] between low-level features and high-level semantics of images. Attribute learning has an important role in supporting image understanding in a fine-grained approach. In the traditional approach, studies used hand-crafted feature SIFT, HOG for feature extraction and SVM for attribute learning. For example, instead of searching for zebras, it is easier to find a black and white striped pattern [23] or to find a spotted dog is replaced by looking for a black dot pattern [24]. Popular works about attribute learning are in [5, 6, 25, 26], attribute learning for face retrieval of [27], Relative Attributes learning [28]. Recently, most studies have used deep learning to improve the accuracy of attribute learning. Many works about object attributes are studied and developed for building effective applications such as person attributes, Face attributes [29–31], vehicle attributes [32], and fashion attributes [27].

In human attributes learning, most studies use transfer learning of famous neural networks, such as VGG, ResNet, etc for feature extraction. After the feature extraction step, the feature vector will be input for the attribute learning method, such as SVM, Bayesian, or other Neural Networks. At the end of the attribute learning step, an attribute vector is returned with a one-hot type vector or list of scores on each attribute. Some popular studies in person attribute learning [33–43].

In [44], Yuxuan Shi et al. (2020) proposed an appearance attribute model named “Attribute Mining and Reasoning” (AMR) for person retrieval by using CNN and Graph Convolution to build two components: ALE (Attribute Attribute Localization Ensemble) and AR (Attribute Reasoning).

In [15], Chiat-Pin Tay et al. proposed an Attribute Attention Network (AANet) by using ResNet-50 architecture. The AANet consists of a global person ID task, a part detection task, and a crucial attribute detection task. Focused attributes are the six attributes such as hair, upper clothing color, lower clothing color, etc.

In [45], Jingya Wang et al. (2018) used encoder-decoder to build a Transferable Joint Attribute-Identity Deep Learning (TJ-AIDL) for simultaneously learning an attribute-semantic for reID and gained the state-of-the-art methods on four challenging benchmarks including VIPeR, PRID, Market-1501, and DukeMTMC.

In [46], Yifang Men et al. (2020) used GAN for person attribute learning and reID with attribute pose, head, upper clothes, and pants.

In [10], the authors proposed a Person re-identification (re-ID) and attribute recognition to share a common target for learning pedestrian descriptions. They used two steps in the model: (1) feature extraction step and (2) classifier step by building a multi-task attribute-person recognition (ARN)network using Resnet50 for feature extraction and classifier block with linear transform and sigmoid. The result of attribute learning is to return a one-hot attribute vector for each image.

In [40], the authors used EfficientNet for attribute learning based on the human part and feature pyramid structure in video surveillance to gain good results in datasets PETA, RAP, and PA-100k.

In [47], authors Yiqiang Chen et al. proposed a pedestrian attribute recognition such as gender, clothing, or hair by combining deep features (feature extraction from CNN) and handcrafted features Local Maximal Occurrence (LOMO) and part-based learning to model attribute patterns related to different body parts. Their experiments achieved state-of-the-art results on PETA, ApiS, and VIPeR datasets.

In [48], authors Xiangpeng Song et al. used GCN to build a pedestrian attribute recognition model and gained the state-of-the-art on the RAP dataset.

In [42], Zichang Tan et al. used GCN to build a new end-to-end network, named Joint Learning of Attribute and Contextual Relations (JLAC) for pedestrian attribute recognition with two modules (1) an attribute graph Attribute Relation Module (ARM) and (2) Contextual Relation Module (CRM) to explore the contextual relations among those regions. the experiments on PA-100K, RAP, and PETA attribute datasets gained good results.

In [5], to exploit the inter-local relation and the intra-local relation among parts from different images for person re-identification (Re-ID), they proposed a novel deep graph model named Heterogeneous Local Graph Attention Networks (HLGAT). The model demonstrated effectiveness and gained state-of-the-art on Market-1501, CUHK03, DukeMTMC-reID, and MSMT17.

In [49], Qi Dong et al. proposed an Attribute-Image Hierarchical Matching (AIHM) model for attribute learning and retrieval with text query input. They used zero-shot learning and deep Multitask learning (MTL) for the system and gained good results on benchmark datasets PA100K, Market1501, and DukeMTMC.

Sarafianos et al. [50] focused on visual attributes of people such as backpack, hat, gender, height, and clothing to propose a method by combining multi-task learning and curriculum learning for attribute classification to try to transfer knowledge from a strong group to a weakly correlated group. Their experiments gained state-of-the-art on PETA and WIDER datasets.

In [51], Shengyu Pei et al. suggested a multitask model for person re-identification based on attributes. They used ResNet for the feature extraction performance. After using generalized mean pooling for feature aggregation, the identity-based and attribute-based double-stream network modules pay attention to the relationship between identities and pedestrian features, and the relationship between attributes and pedestrian features, to fully activate the relationship between pedestrian attributes and identities. Their experiments are on datasets Market-1501 and DukeMTMC-reID and gained good results to prove the effectiveness of the method.

In [52], authors Ngoc Q. Ly et al. suggested a deep attribute learning for Object retrieval based on object ontology, and deep neural network. With improving imbalanced data and multitask learning, the model gained state-of-the-art on the DeepFashion dataset.

In this paper, we followed [52] to build human attribute learning based on HAO and deep learning. We tried to use transfer learning for feature extraction by using a CNN EfficientNet [18] for the feature extraction step and with the support of HAO and a linear classifier, our attribute learning model gained some state-of-the-art on PETA and PA100k datasets.

## 2. Image retrieval based on attribute learning

The image retrieval system tries to find candidates that are similar to the image input (query). In the traditional model, most studies used the similarity of feature vectors of candidate images and query images to collect results. Recently, to improve the accuracy of the image retrieval system, studies built image retrieval systems based on attribute learning of objects in images. It is the fine-grained approach to make the retrieval more intelligent than the old system. Because it could recognize attributes of objects and retrieve the attributes, instead of comparing feature vectors in low-level with blind semantics of images. Using deep learning for building attribute learning in retrieval systems is a robust step.

Recently, Person Re-Identification based on attribute learning has developed progress. Some popular works that used deep learning for attribute learning in Person Re-ID are [17, 44, 45]. In [12], Yutian Lin et al. (2019) built a person re-identification and attribute recognition share a common target for learning pedestrian descriptions. They proposed an attribute-person recognition (APR) network and achieved the state-of-the-art on Market-1501.

The authors in [53] proposed SAL (Symbiotic Adversarial Learn) by using two GANs to build the model of person search and gained high results in the PETA and Market1501 datasets. In [35], the authors proposed the UPAR to unique the pedestrian attribute recognition and person search. They used ConvNeXt [54] to build Attribute recognition and build person search based on attributes, and gained state-of-the-art on four datasets Market1501, Peta, Pa100k, and Rapv2.

- Text-based query for Image Retrieval

Most image retrieval uses image query. Some previous works admit using text query for image retrieval without the caption of images by annotation.

In [55], Zhe Wang et al. built a person search named ViTAA(Visual-Textual Attributes Alignment) by natural language. They tried to match specific attribute phrases to the corresponding visual regions with attribute learning support. The model achieved state-of-the-art performances.

In [49] Qi Dong et al. proposed a person search method using a text attribute query. They built an Attribute-Image Hierarchical Matching (AIHM) model to match text attribute descriptions with noisy surveillance person images by jointly learning global category-level and local attribute-level textual-visual embedding as well as matching. The AIHM model gained state-of-the-art on Market-1501, DukeMTMC, and PA100K datasets.

In [56], Tinghuai Ma et al. proposed a novel Dual-path CNN with MaxGatedblock(DCMG) for text-based person re-identification. The model is based on two deep residual CNNs and gained state-of-the-art on the CUHK-PEDES dataset.

In [57], Sijin Wang et al. built an Image-text retrieval of natural scenes by using graph matching. Focusing on two kinds of visual concepts, objects, and their relationships, they proposed to represent image and text with two kinds of scene graphs: visual scene graph (VSG) and textual scene graph (TSG). The model achieved state-of-the-art results on Flickr30k and MS COCO.

In [58], the authors proposed ASMR (Adaptive Semantic Margin Regularizer) for Attribute-based person search by building a network with a special loss function in embedding space of images and semantics of images as attributes or person categories. The study gained high results on datasets PETA, PA100k, and Market-1501.

### 3. Image retrieval based on attribute learning and ontology

Ontology is a tool to represent knowledge for sharing and reusing [19]. Especially. It is useful to support the semantics of images in building applications for image retrieval image interpretation [52, 59–62]. In attribute learning, ontology supports building the hierarchical semantic tree for facial attributes [20, 63, 64]. In [63], Garcıa-Rojas et al. built an ontology for facial representation for emotional face expression profiles. Combining deep imbalanced data learning and ontology, Hung M Nguyen et al. [20] proposed a facial attribute ontology (FAO) for attribute learning and achieved high accuracy with 40 facial attributes on CelebA and LFWA datasets. Focusing on aesthetics, Mingliang Xu et al. [21] suggested Bio-FAO (Bio-inspired Facial Aesthetic Ontology) and combined it with CNN to predict facial aesthetics. Using fuzzy reasoning, Renan Contreras et al. [63] proposed an emotion recognition method based on ontology. In [52], authors Ngoc Q. Ly et al. proposed an attribute learning model based on ontology and attribute learning for person re-identification on the Market-1501 dataset and object retrieval on the DeepFashion dataset.

In our paper, we used deep learning and ontology to build an image retrieval based on attributes. By combining deep learning and ontology, our image retrieval has increased accuracy and effectiveness. In the next part, we will show the design of Human Attribute Ontology in section III, the attribute learning in section IV, the human image retrieval in section V, the experiments in section VI and results in section VII, and the conclusion in section VIII.

## III. HUMAN ATTRIBUTE ONTOLOGY

In this paper, we propose a novel ontology named Human Attribute Ontology (HAO) that facilitates the organization and storage of knowledge related to human objects in the real world.

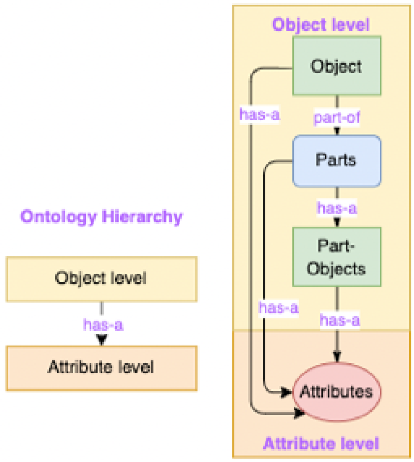



Figure 1: HAO with Hierarchical concepts in general

We utilize attribute learning and ontology to construct a human retrieval system operating at the attribute level. The ontology plays a crucial role in facilitating both the attribute learning stage and the retrieval stage. The reason for using ontology in the retrieval system is to (1) Reuse prior knowledge by rules and hierarchy in the ontology will support the attribute learning step; (2) Decrease the gap between the semantics of images from low-level features and high-level semantics of images, the ontology is to connect concrete images in the real world to the semantics of images in the ontology.

### 1. Ontology definition

Our HAO has five components with (C, P, Re, Ru, I):

- C (Concepts): They include names of objects, parts of objects, and attributes of objects. An example of the hierarchy of concepts is represented in the Figure 1 and Figure 2
- P (Properties): Each concept has a set of properties for describing itself.
- Re (Relations): A set of relationships of the concepts.
- Ru (Rules): A set of rules for reasoning concepts in the system.
- I (Instances): A set of Images with objects represented knowledge in HAO.

There are some definitions for constraining between concepts in the HAO:

- Definition 1: Each object has one part or some parts that are made in the object. The whole part is the biggest part of the object. One object has one whole part. For example, a human object has head part, top part, bottom part, foot part and whole part.
- Definition 2: Each part has one or many objects inside the part and is called PartObject. Each part may be

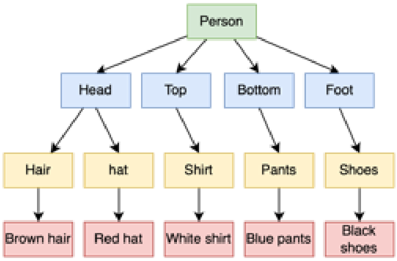



Figure 2: An example of hierarchical concepts in HAO

included null PartObject. For example, the top part of the person has a shirt object; the head part has a hat object.
- Definition 3: Each part has some attributes from itself or attributes of PartObjects in the part. And then each object has some attributes from its Parts and PartObjects.
- Definition 4: Each attribute belongs to one attribute group. For example, Asian is in the race group, and yellow skin is in the physical appearance group.
- Definition 5: There are two types of attributes, they are (1) Common type (also called visual concept type or general type) such as color, shape, texture, and size; (2) Specific type for character of each object. For example, a person object has a specific type: emotion, race, age, and gender. Fashion objects include material, office clothing, and party clothing.
- Definition 6: Visual concepts such as color, shape, and texture have a role in describing objects and attributes, they are in properties P of HAO.

## 2. An example of HAO

In this paper, we focused on human objects and human attributes to build HAO. We collect concepts about humans and relevant concepts about humans to build the ontology.

*a) Concepts*

Example of concepts of human objects and its hierarchy in Figure 1, Figure 2, Table I, Table II

*b) Relations*

There are relations between concepts of human object in the Table III.

*c) Rules*

A Rule has the following form:

$$Input \xrightarrow{\text{Condition}} Output$$

Each rule is defined by a tuple:

$$(Inputs, Outputs, Conditions).$$

TABLE I: CONCEPTS: OBJECTS AND PARTS.

| Concepts | Type/Description | Context |
|---|---|---|
| Person | Object | Person |
| Fashion | Object | Fashion |
| Face | Object | Person |
| Shirt | Object | Person, Fashion |
| Pant | Object | Person, Fashion |
| Hat | Object | Person, Fashion |
| Head | Part of Person | Person |
| Top | Part of Person | Person |
| bottom | Part of Person | Person |
| Foot | Part of Person | Person |
| body | Part of Person | Person |
| wholepart | Part of Person | Person |

TABLE II: CONCEPTS: ATTRIBUTES.

| Attributes | Objects | Parts | Visual concepts | Groups |
|---|---|---|---|---|
| Blue jean | Person | bottom | Color | Fashion |
| Tall | Person | Whole | Size | Height |
| Short | Person | Whole | Size | Height |
| Thin | Person | Whole | Size | Body type |
| Fat | Person | Whole | Size | Body type |
| Bright skin | Person | Whole | Color | Complex |
| Straight hair | Person | Head | Texture | Hair style |

Example rule: ”smile”$\rightarrow$ ”happy” with null condition with a tuple (”smile”, ”happy”, null), ̈long hair” $\rightarrow$ ”female” with rate=90% with a tuple ”long hair”, ”female”, rate=0.9). default of rate=1.0. Part= ”top” and color= ”red” $\rightarrow$ ”upred”.

There are some rules in the HAO as follows in the Table IV. Our HAO gains some characteristics in the standard of ontology: systematic, flexible, and extensible. With the knowledge in HAO, it can support some functions as follows:

- Support semantics of images by concepts hierarchy in HAO
- Support reasoning by rules in HAO to gain new knowledge without learning.
- Reuse prior knowledge: How many attributes are for training?

In this paper, we use HAO for support attribute learning in the following functions: (1) Support attribute training and matching for decreasing time and data space search Attribute for fine-grained learning, representation, and searching. (3) in the retrieval stage, HAO helps with preprocessing queries and reranking results of the retrieval system.

In the next section, we present the Attribute Learning

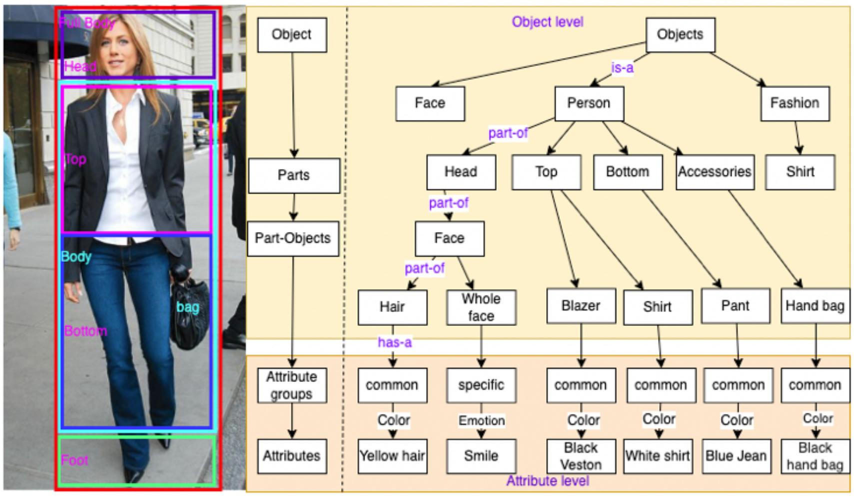



Figure 3: An example of HAO

TABLE III: Relations.

| Relations | Examples |
|---|---|
| Is-a | Person ”is-a” object (sub-object and parent-object) |
| Has-a | Person “has-a” Face |
| Part-of | Head “part-of” Person (component) |
| Group-of | Black hair “group-of” black color, physical appearance, Asian race, Black hair “belong-to” physical appearance |
| Same-as | black skin “same-as” Dark skin, boy same-as male |
| Belong-to | Red “belong-to” Color, Color “has-a” red, Attribute “belong-to” part, Part ”belong-to” object, Attribute “belong-to” attribute-group |

TABLE IV: RULES.

| Rule ID | Input | Output | Condition |
|---|---|---|---|
| R1 | Smile | Happy | Null |
| R2 | Long hair | female | Rate=0.9 |
| R3 | Mustache | Male | |
| R4 | Bushy Beard | Male | |
| R5 | Red dress | Young girl | Rate=0.9 |
| R6 | Veston | Formal | |
| R7 | backpack | young style | |

method that used deep learning and HAO support to improve accuracy of the human attribute recognition.

## IV. ATTRIBUTE LEARNING

Attributes are a mid-level between the low-level feature to the high-level semantics of images [9]. In our system, Attribute Learning (AL) is a very important module for the representation of images. It is a middle step for the retrieval system to transform from a raw image to an attribute vector

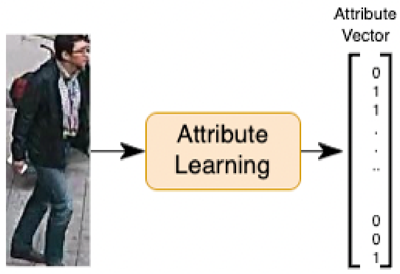



Figure 4: Attribute learning model in general

(shown in Fig. 4), that is conveying important information about the image at the semantic level to improve the image understanding capacity of the retrieval system.

The results of attribute learning help our retrieval to recognize images based on attributes. It shows some advantages: (1) the system is a fine-grained approach, can understand the image in detail and then increase accuracy, and (2) the system is more intelligent because it can retrieve on attributes of the object instead of based on low-level features in traditional ways or the whole object only.

In formal statements of attribute learning as follows:
Let $X = \{x_1, x_2, \ldots, x_N\}$, a set of images of the dataset, N is the number of the images.
Let $A = \{a^1, a^2, \ldots, a^M\}$, a set of M attribute labels of the system. Let $Y = \{y_1, y_2, \ldots, y_N\}$, a set of attribute labels of images, $y_i$ is a binary vector that indicates the presence with 1 and absence with 0 of a particular attribute in the image (img), $y_i \in \{0, 1\}^m$, $y_i = \{y_{i,1}, y_{i,2}, \ldots, y_{i,M}\}$, with:

$$y_{i,j} = \begin{cases} 1 \text{ if attribute } j^{th} \left(a^j\right) \text{ presents in img } i^{th} \\ 0 \text{ otherwise.} \end{cases} \quad (1)$$

Our method consisted of three main steps. They are (1)

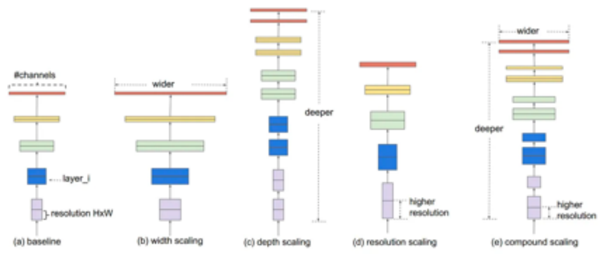



Figure 5: EfficientNet [65]

feature extraction, (2) Partition, and (3) classification step:

- Feature Extraction (FE)

In the FE step, we used EfficientNetB5 [65] for extracting a feature vector length of 2048. With HAO knowledge, we separate the feature vector into parts of the object that are declared in the ontology and collect the attribute list of each part, meaning which attributes are in which part, and then input them to the next step which is the attribute classification step.

EfficientNet [65] is the CNN, that used a technique of scaling up width, and depth to build the model. The authors of EfficientNet developed seven models, named B1 to B7. In the topic, we used EfficientNet-B5 for attribute learning in the feature extraction step with a 2048-length of the feature vectors.

- Part-based attribute learning: Connecting HAO and AL

We build AL based on the part objects with the support of ontology HAO. With the support of HAO, our attribute learning approach is based on parts of the object. From the part information of object in HAO, we get the part information and divide the image object into a number of parts corresponding to the part of the object in HAO. For example, with human object, there are five parts of the human object, they are head part, top part, bottom part, foot part, and whole part.

With each part, we get the list of attributes of the part and input them for the training. For example, with the top part, there are attributes in the top part such as shirt and attributes of shirt: red shirt, black shirt, ... With part and attribute in part gained from the HAO, they are input for the training in the next step.

Part-based attribute learning will be effective: (1) to decrease the cost for the training instead of training on the whole part; (2) Each attribute will be determined by a part of an object to avoid training with attributes and parts are not relevant; (3) attributes are grouped on part and hierarchy, then it is not the same level difference from attribute 1 to attribute m make a list for input training.

After the partition step, the next step is the training step for attribute classification.

- Attribute classification

In the classification step, we adopted the attribute learning model [12] as a multi-label classification task also multiple binary classifiers with Binary Cross-Entropy loss, SGD optimization, and sigmoid function for the probability of the result attribute vector are used. The loss function of the step is as follows:

$$L_{CLS} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M}\omega_j (y_{i,j}\log(\sigma(p_{i,j})) + (1-y_{i,j})\log(1-\sigma(p_{i,j}))) \tag{2}$$

With

$$\sigma(z) = \frac{1}{1+e^{-z}}$$

and $p_{i,j}$ is the output of the classifier layer, and $\omega_j$ adopted from [12] for imbalanced data between attributes with the formula (3):

$$\omega_j = \begin{cases} e^{1-r_j}, y_{i,j} = 1 \\ e^{r_j}, y_{i,j} = 0 \end{cases} \tag{3}$$

With $r_j$ is the positive sample ratio of $j^{th}$ attribute in the training set.

- Multitask learning: Connecting HAO and AL

Multitask Learning (MTL) is an advanced machine learning technique to improve performance [66] by sharing common data and combining related tasks to train together. There are two stages in the MTL, the shared stage and the separate stage.

In multi-task learning, We build MTL for attribute learning with two stages: a shared stage and a separate stage. In the shared stage, we used a CNN (EfficientNet) for feature extraction of images. After that, the feature map of the previous step will be partitioned into parts (Head, Top, Bottom, Foot) corresponding to parts of human information in HAO (shown in Figure 6). The feature vector of each part will be input into the next step. In the separate stage, we build attribute classifiers of attributes following attribute groups (gender, race, ...) corresponding to attribute groups and attribute lists on each part of human information in HAO. A summary of MTL of attribute learning on the Head part is shown in Figure 7.

- Loss function

Loss function in the attribute learning stage includes loss of FE and loss of classification, as shown in formula (4).

$$L_{AL} = L_{FE} + L_{CLS} \tag{4}$$

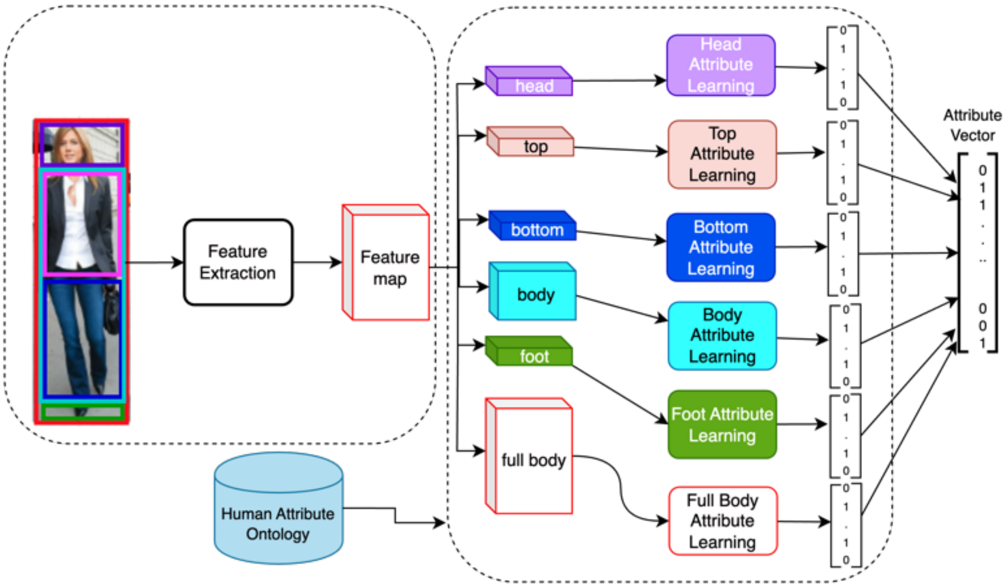



Figure 6: Attribute learning model in Detail

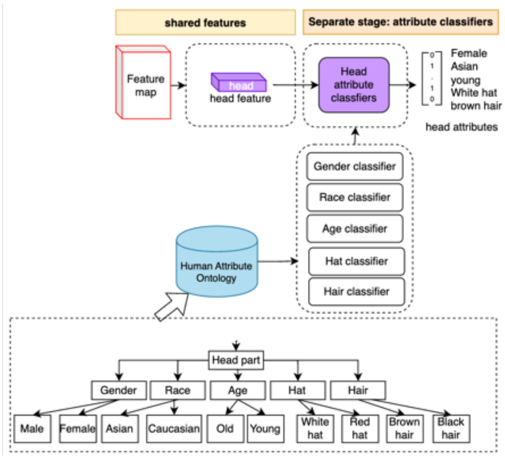



Figure 7: A summary MTL of AL on Head part

With $L_{FE}$ is the loss of feature extraction method, $L_{CLS}$ is the loss of classification. In the next section, we will present the retrieval system that the main task of the topic.

## V. HUMAN IMAGE RETRIEVAL BASED ON ONTOLOGY AND ATTRIBUTE LEARNING

In this section, we present our human image retrieval system based on HAO and AL. The system comprises two stages: the offline stage and the online stage. The offline stage includes steps: the attribute learning and collecting attribute vectors of the dataset and saving them to the attribute vector database. The online stage has two steps: (1) Query set and gallery set, and (2) searching with matching, filtering and Ranking for the results of the retrieval. The model of the retrieval is shown in Figure 8.

- Query set and gallery set

In the offline stage, after the training step, we collected the parameters of the model and computed attribute vectors of images in the test set. The attribute vectors will be stored in the attribute vectors database. We divided the test set into query set $Q$ and gallery set $G$ that satisfies: $Q \cap G = \emptyset$ With an image query input $q \in Q$, we used the attribute learning model to infer the attribute vector of the query: aq. It is also in the attribute vector database if the query is in the test set. After gaining the attribute vector of query, we used it as input for the next step which is the matching and filtering step.

- Searching: Matching and filtering

In the step, from the attribute vector of the query in the previous step, we input it into the matching and filtering module by comparing the attribute vectors in the attribute vector database in the offline stage. We used the $L2$ measurement for comparing the distance of the query and the attribute vectors of the gallery in the database, to get the top 10 smallest distances. It returns the result of the retrieval. The search algorithm presents the method for matching and filtering. The algorithm is presented in Algorithm. 1.

## VI. EXPERIMENTS

### 1. System configuration and datasets

We used Google Colab Pro for experiments with the programming language Python 3.9 with libraries such as Torch,

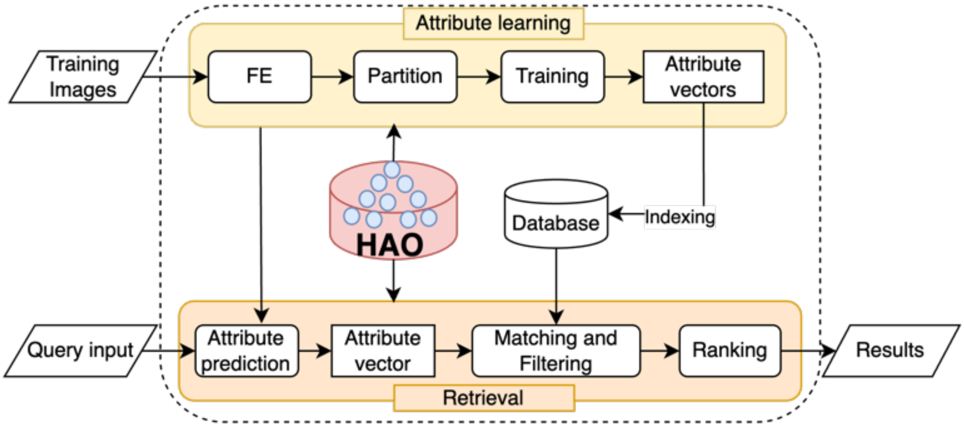



Figure 8: Human retrieval system

**Algorithm 1:** Search Algorithm

```
Inputs: image query q, image galleries G.
Outputs: top-10 image results.
begin
    aq=attributePredicting (q); //get attribute vector of q
    aG=attributePredicting (G);//get attribute vectors of G
    dis=distance(aG,aq); //computing distance attribute
      vectors of aG and aq
    candidates=argSort(dis); //sort the best candidates go
      to the head list
    result=getTopK(candidates); //get top-10 candidates
    Return result;
end
```

TABLE V: DATASETS.

| Datasets | No. Img | No. Atts | Train set | Val set | Test set |
|---|---|---|---|---|---|
| PETA | 19,000 | 35 | 9,500 | 1,900 | 7,600 |
| Pa100k | 100,000 | 26 | 80,000 | 10,000 | 10,000 |

keras, tensorflow, numpy, and OpenCV. We demonstrated our proposed methods on a large Re-ID dataset named PETA [67], and PA100k [36]. The PETA (PEdesTrian Attribute) dataset contains 19,000 indoor and outdoor images with 35 attributes. The PA100k (pedestrian attributes) dataset consists of 100,000 outdoor pedestrian images with 26 attributes. The information on datasets is shown in the Table V.

### 2. Ontology building

In the experiment of building HAO, we focused on the human knowledge domain with three relevant objects: Person (Pedestrian), Fashion, and Face object for HAO. We described Person object as follows the Figure 3. with parts: head, top, bottom, foot, wholebody. Each part has a set of attributes.

### 3. Attribute learning

In attribute learning, we use transfer learning EfficientNetB5 [65] for feature extraction by resizing the input image to 224x224. The result of the step is to return a feature vector of 2048 lengths (the default length of the feature vectors of EfficientNetB5). After that, we used the information about parts of objects in HAO to divide the feature vector into the number of parts of objects. In each part, we used a classifier for them. In this case, we used a Linear classifier combining the sigmoid function of the Torch library for classifiers. In the training step, we used SGD optimization, Binary Entropy Loss by following the experiment of author Yutian Lin [12]. We trained 40 epochs and a learning rate of $10^{-4}$. We evaluate our attribute learning with attributes to follow each dataset. The result of AL will be shown in the next section.

Partitition is depended on the object in experiments. In the case, without loss of generality, we divided an image (as an object) into 4 segments lengthwise with the rate: the head part(1/6), the top part(3/8), the bottom part(2/6), the foot part(2/6), and the wholepart (6/6). Each part will input to the Attribute training step.

### 4. Image retrieval

With the query set and gallery set, we implemented the search algorithm (in Algorithm 1) and gained the results. We separated into three cases for the retrieval stage belonging to the organizing query set and gallery set as follows Table X. In case 1, we filtered images from the test set with each person having greater or equal 10 images, in case 2 with each person having greater or equal 5 images, and in case 3 with 500 random images from the test set for query set, and the remaining images of the test set for gallery set. The results of the retrieval are shown in the next section.

## VII. RESULTS

### 1. Evaluation Metrics

In the evaluation metric, we followed the work [68] with measurements: Mean Accuracy (mA), Accuracy (Acc), Precision (Prec), Recall (Rec), F1-score. In which, mA is as follows the equation (5):

$$mA = \frac{1}{2N}\sum_{i=1}^{M}\left(\frac{TP_i}{P_i} + \frac{TN_i}{N_i}\right), \quad (5)$$

where $N$ is the number of examples and $M$ is the number of attributes; $P_i$ and $TP_i$ are the number of positive examples and correctly predicted positive examples of the $i^{th}$ attribute respectively; $N_i$ and $TN_i$ are defined for negative cases. The measurements: Accuracy (Acc), Precision (Prec), Recall (Rec), F1-score have display in [68] in detail.

In the evaluation of the retrieval system, we used the metric Top-k with k={1, 5, 10}, and Mean Average Precision (mAP)[68]. Top-k is the true positive in fisrt k results, it is also the rank-k accuracy.

### 2. Attribute learning

We compare our AL method with and without using HAO to popular studies about attribute learning on two datasets PETA and Pa100k. The results are shown in Table VI and Table VII.

The results of 35 attributes on PETA and Pa100k datasets in Table VI and Table VII were not on top 1, but they kept sTable With the support of ontology HAO and attribute multitask learning, the results are effective in the retrieval stage. Besides, we evaluated 6 attributes on the PETA dataset and gained high results in the Table VIII and Table IX.

### 3. Image retrieval

we experimented on the PETA dataset with cases in the Table X, our results in the Table XI, and compared to some popular studies and gained high results on the PETA dataset as follows Table XII. An example is in Figure 9 with query is in blue box.

The same experiment on the Pa100k dataset is shown in the Table XIII, our method results in Table XIV, and compared to popular studies in the Table XV.

Our results showed that the person search based on attributes effectively, not only finding out the person in the query image but also finding out people with the same attributes of the query. For example in the Figure 11 and 12. (the query is in the blue box).

TABLE VI: AL RESULTS OF 35 ATTRIBUTES ON THE PETA DATASET.

| Topics | ma | Acc | Prec | Rec | F1-score |
|---|---|---|---|---|---|
| Zichang Tan [33] | 84.88 | 79.46 | 87.42 | 86.33 | 86.87 |
| HydraPlus-Net [36] | 81.77 | 76.13 | 84.92 | 83.24 | 84.07 |
| ALM [37] | 86.3 | 79.52 | 85.65 | 88.09 | 86.85 |
| Strong Baseline [38] | 85.11 | 79.14 | 86.99 | 86.33 | 86.09 |
| Jian Jia 2021 [39] | 86.52 | 78.95 | 86.02 | 87.12 | 86.99 |
| feature pyramid [40] | 87.69 | 81.2 | 87.59 | 89.2 | 88.32 |
| DAFL[41] | 87.07 | | | | 86.4 |
| JLAC [42] | 86.96 | 80.38 | 87.81 | 87.09 | 87.45 |
| Rein-PAR[43] | 85.51 | 78.45 | 84.08 | 88.77 | 85.91 |
| Dai[34] | 88.24 | 79.14 | 88.79 | 84.7 | 86.7 |
| UPAR [35] | 88.40 | | | | 89.9 |
| Ours:AL+w/o HAO | 85.50 | 79.50 | 87.32 | 86.45 | 86.60 |
| Ours:AL+HAO | 85.41 | 80.17 | 88.09 | 86.59 | 87.07 |

TABLE VII: AL RESULTS OF 26 ATTRIBUTES ON THE PA100K DATASET.

| Topics | ma | Acc | Prec | Rec | F1-score |
|---|---|---|---|---|---|
| Zichang Tan [33] | 81.61 | 78.89 | 86.83 | 87.73 | 87.27 |
| HydraPlus-Net [36] | 74.21 | 72.19 | 82.97 | 82.09 | 82.53 |
| ALM [37] | 80.68 | 77.08 | 84.21 | 88.84 | 86.46 |
| Strong Baseline [38] | 79.38 | 78.56 | 89.41 | 84.78 | 86.25 |
| Jian Jia 2021 [39] | 81.87 | 78.86 | 85.98 | 89.1 | 86.87 |
| feature pyramid [40] | 81.45 | 79.13 | 86.24 | 89.46 | 87.94 |
| DAFL[41] | 83.54 | | | | 88.09 |
| JLAC [42] | 82.31 | 79.47 | 87.45 | 87.77 | 87.61 |
| Rein-PAR[43] | 80.55 | 77.2 | 84.76 | 87.67 | 85.7 |
| Dai[34] | 77.89 | 79.71 | 90.26 | 85.37 | 87.75 |
| UPAR [35] | 84.80 | | | | 90.2 |
| Ours:AL+w/o HAO | 82.49 | 79.89 | 87.39 | 88.32 | 87.49 |
| Ours:AL+HAO | 75.46 | 75.34 | 88.86 | 81.34 | 84.42 |

TABLE VIII: AL RESULTS: ACCURACY OF TOP 6 ATTRIBUTES ON THE PETA DATASET.

| Attributes | Acc | Prec | Rec | F1-score |
|---|---|---|---|---|
| lowerBodyCasual | 94.22 | 96.13 | 97.94 | 97.03 |
| upperBodyCasual | 93.72 | 95.87 | 97.66 | 96.76 |
| personalLarger60 | 90.19 | 98.14 | 91.76 | 94.84 |
| accessoryMuffler | 89.97 | 97.33 | 92.25 | 94.72 |
| personalMale | 89.16 | 93.73 | 94.82 | 94.27 |
| accessoryNothing | 88.80 | 93.41 | 94.74 | 94.07 |
| Average | 91.01 | 95.77 | 94.86 | 95.28 |

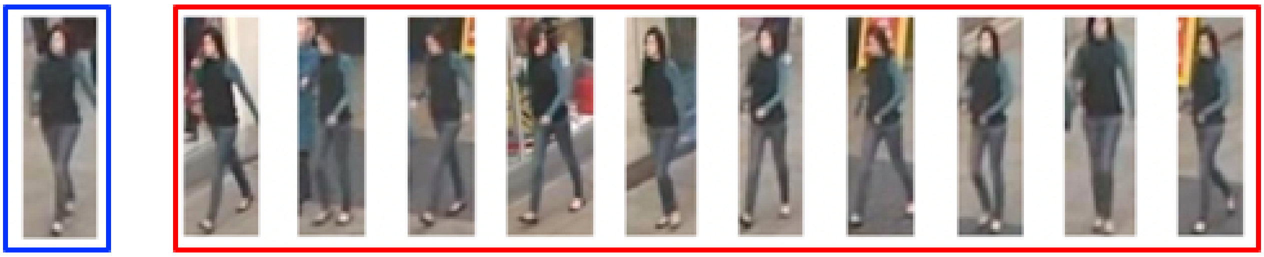


Figure 9: A result top-10 in the best case on the PETA dataset

Figure 10: A result top-10 in the best case on the Pa100k dataset

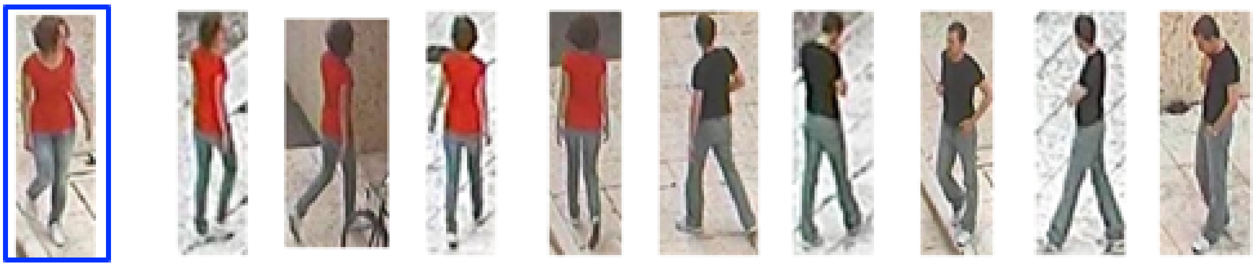


Figure 11: Example search person based on attributes: ✓accessoryNothing, ✓upperBodyCasual, ✓upperBodyShortSleeve, ✓upperBodyTshirt, ✓lowerBodyCasual, ✓lowerBodyJeans, ✓footwearShoes, ✓carryingNothing, ✓personalLess30

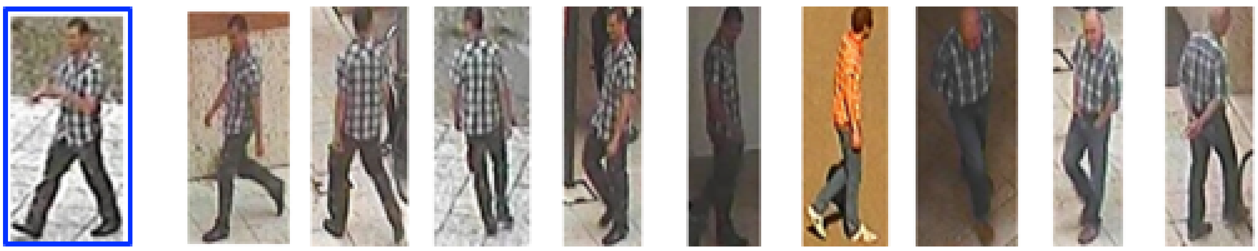


Figure 12: Example search person based on attributes: ✓accessoryNothing, ✓upperBodyPlaid, ✓upperBodyShortSleeve, ✓lowerBodyCasual, ✓lowerBodyJeans, ✓footwearShoes, ✓carryingNothing, ✓personalLess30, ✓personalMale.

- Discussion

In the dataset PETA, there are over 4000 persons who have one image/person, which is why the random case for query (case 3 in Table XI) is the worst.

Our results in case 1 and case 2 also show that: with attribute learning, we could re-identification a person without training person-ID. It improves the performance of the system because it eliminates a training stage for the Person Re-Identification task.

Our result on the Pa100k dataset is not better than on the PETA dataset because the number of images in PA100k is bigger. It is a disadvantage for improving in the future works.

## VIII. CONCLUSION

This topic is to improve the image understanding capacity of image retrieval systems. First, we developed HAO for storing the knowledge of images. Second, we built an MTL

TABLE IX: AL RESULTS: COMPARING THE ACCURACY OF THE TOP 6 ATTRIBUTES ON THE PETA DATASET.

| Attributes | SAR [69] | MAR [69] | Ours | +/- |
|---|---|---|---|---|
| lowerBodyCasual | 81.60 | 84.90 | 94.22 | + |
| upperBodyCasual | 81.10 | 84.40 | 93.72 | + |
| personalLarger60 | 92.00 | 94.80 | 90.19 | - |
| accessoryMuffler | 94.40 | 96.10 | 89.97 | - |
| personalMale | 85.10 | 89.90 | 89.16 | - |
| accessoryNothing | 81.50 | 85.80 | 88.80 | + |
| Average | 85.95 | 89.32 | 91.01 | + |

TABLE X: CASES OF QUERY AND GALLERY SET IN THE PETA DATASET.

| Cases | description | query set | gallery set |
|---|---|---|---|
| 1 | 10 images/person | 144 | 2313 |
| 2 | 5 images/person | 250 | 2953 |
| 3 | 500 random images | 500 | 7100 |

TABLE XI: 3 CASE RESULTS ON THE PETA DATASET

| Cases | Top-1 | Top-5 | Top-10 | mAP |
|---|---|---|---|---|
| 1 | 92.36 | 97.22 | 97.22 | 94.98 |
| 2 | 86.00 | 92.80 | 93.20 | 87.66 |
| 3 | 40.60 | 48.30 | 49.80 | 40.39 |
| Average | 72.99 | 79.44 | 80.07 | 74.34 |

TABLE XII: RESULTS IR ON THE PETA DATASET

| Cases | Top-1 | Top-5 | Top-10 | mAP |
|---|---|---|---|---|
| UPAR [35] | 29.70 | | | 30.20 |
| SAL [53] | 47 | 66.50 | 74 | 41.50 |
| ASMR [58] | 56.50 | 80.00 | 83.50 | 50.20 |
| Ours (avg) | 72.99 | 79.44 | 80.07 | 74.34 |

TABLE XIII: CASES FOR RETRIEVAL IN THE PA100K DATASET.

| Cases | description | query set | gallery set |
|---|---|---|---|
| 1 | 10 images/person | 191 | 8917 |
| 2 | 5 images/person | 248 | 9244 |
| 3 | 500 random images | 500 | 9500 |

TABLE XIV: 3 CASE RESULTS ON THE PA100K DATASET

| Cases | Top-1 | Top-5 | Top-10 | mAP |
|---|---|---|---|---|
| 1 | 65.44 | 67.53 | 69.63 | 25.35 |
| 2 | 62.90 | 64.51 | 66.12 | 15.15 |
| 3 | 19.40 | 33.60 | 43.60 | 7.22 |
| Average | 49.25 | 55.21 | 59.78 | 15.91 |

TABLE XV: RESULTS IR ON THE PA100K DATASET

| Cases | Top-1 | Top-5 | Top-10 | mAP |
|---|---|---|---|---|
| UPAR [35] | 39.50 | | | 30.50 |
| ASMR [58] | 31.9 | 49.1 | 58.2 | 20.6 |
| AIHM [49] | 31.3 | 45.1 | 50.0 | 17.0 |
| Ours (avg) | 49.25 | 55.21 | 59.78 | 15.91 |

attribute learning with the support of HAO and deep neural networks. Finally, we constructed the original HR system based on HAO and AL. The retrieval could understand human images in detail at the attribute level, and be more intelligent. We demonstrated that the performance of our retrieval system was better than some state-of-the-art on datasets PETA and Pa100k.

In the future, we could use the rule component of HAO for attribute reasoning based on rules. This helps AL decrease the costs of training and improve the performance of the system because no need for training the reasonable attributes from the rules. With rule-based reasoning, we can infer new knowledge from prior knowledge without training. The reasoning module also supports detecting the relation between an attribute and relevant attributes in an object, an object, and relevant objects. HAO and reasoning modules based on rules will help to reuse prior knowledge and close the semantic gap between low-level features and high-level semantics of images.

## Acknowledgements

This research is supported by research funding from the Faculty of Information Technology, University of Science, Vietnam National University-Ho Chi Minh City, Saigon University, and by the JSPS Core-to-Core Program (grant number: JPJSCCB20230005).

## References

[1] E. Yaghoubi, F. Khezeli, D. Borza, S. Kumar, J. Neves, and H. Proença, "Human attribute recognition-a comprehensive survey," *Appl. Sci*, 2020.

[2] S. Dubey, "A decade survey of content based image retrieval using deep learning," 2021.

[3] M. Kalayeh, E. Basaran, M. Gokmen, M. Kamasak, and M. Shah, "Human semantic parsing for person re-identification," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit*, p. 1062–1071, 2018.

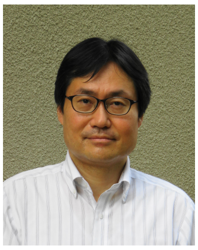
[4] T. He, X. Shen, J. Huang, Z. Chen, and X.-S. Hua, "Partial person re-identification with part-part correspondence learning," *Cvpr*, p. 9105–9115, 2021.

[5] J. Liu, Z.-J. Zha, W. Wu, K. Zheng, and Q. Sun, "Spatial-temporal correlation and topology learning for person re-identification in videos," *CVPR*, p. 4370–4379, 2021.

[6] Z. Zhang, H. Zhang, and S. Liu, "Person re-identification using heterogeneous local graph attention networks," *Cvpr*, p. 12136–12145, 2021.
[7] B. Nguyen, B. Nguyen, T. Do, E. Tjiputra, Q. Tran, and A. Nguyen, "Graph-based person signature for person re-identifications," 2021.
[8] D. Parikh and K. Grauman, "Relative attributes learning," p. 503–510, 2011.
[9] K. Grauman and B. Leibe, *Visual object recognition*, 2011.
[10] X. Wang, S. Zheng, R. Yang, B. Luo, and J. Tang, "Pedestrian attribute recognition: A survey," p. 1–32, 2019.
[11] J. Joo, S. Wang, and S. Zhu, "Human attribute recognition by rich appearance dictionary," *Proc. IEEE Int. Conf. Comput. Vis*, p. 721–728, 2013.
[12] Y. Lin, "Improving person re-identification by attribute and identity learning," *Pattern Recognit*, vol. 95, p. 151–161, 2017.
[13] G. Zhang and J. Xu, "Person re-identification by mid-level attribute and part-based identity learning," *Proc. Mach. Learn. Res*, vol. 95, p. 220–231, 2018.
[14] S. Li, T. Xiao, H. Li, B. Zhou, D. Yue, and X. Wang, "Person search with natural language description," p. 5187–5196, 2017.
[15] H. Galiyawala and M. Raval, "Person retrieval in surveillance using textual query: a review," *Multimed. Tools Appl*, 2021.
[16] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. Hoi, "Deep learning for person re-identification: A survey and outlook," p. 1–20, 2020.
[17] C. Tay, S. Roy, and K. Yap, "Aanet: Attribute attention network for person re-identifications," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit*, vol. 2019-June, p. 7127–7136, 2019.
[18] M. Wieczorek, A. Michalowski, A. Wroblewska, and J. Dabrowski, "A strong baseline for fashion retrieval with person re-identification models," 2020.
[19] T. Gruber, "Toward principles for the design of ontologies used for knowledge sharing," *Int. J. Hum. - Comput. Stud*, vol. 43, no. 5–6, p. 907–928, 1993.
[20] H. Nguyen, N. Ly, and T. Phung, "Large-scale face image retrieval system at attribute level based on facial attribute ontology and deep neuron network," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics*, vol. 10752, p. 539–549, 2018.
[21] M. Xu, "Bio-inspired deep attribute learning towards facial aesthetic prediction," *IEEE Trans. Affect. Comput*, vol. 3045, no. c, p. 1–1, 2018.
[22] R. Feris, C. Lampert, and D. Parikh, *Visual Attributes*. Springer, 2017.
[23] V. Ferrari and A. Zisserman, "Learning visual attributes," *Adv. Neural Inf. Process. Syst*, vol. 20, 2008.
[24] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth, "Describing objects by their attributes," 2009.
[25] C. Lampert, H. Nickisch, and S. Harmeling, "Attribute-based classification for zero-shot visual object categorization," 2013.
[26] O. Russakovsky and L. Fei-Fei, "Attribute learning in large-scale datasets," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics*, no. PART 1, p. 1–14, 2012.
[27] N. Kumar, "Describable visual attributes for face images," 2011.
[28] Y. Fu, "Attribute learning for image / video understanding," *Queen*, 2015-03.
[29] E. Rudd, M. Günther, and T. Boult, "Moon: A mixed objective optimization network for the recognition of facial attributes," *LNCS Bioinforma*, p. 19–35, 2016.
[30] K. He, Z. Wang, Y. Fu, R. Feng, Y. Jiang, and X. Xue, "Adaptively weighted multi-task deep network for person attribute classification," *MM*, p. 1636–1644, 2017.
[31] D. Jayaraman, F. Sha, and K. Grauman, "Decorrelating semantic visual attributes by resisting the urge to share," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit*, p. 1629–1636, 2014.
[32] Z. Huo, Y. Xia, and B. Zhang, "Vehicle type classification and attribute prediction using multi-task rcnn," p. 564–569, 2016.
[33] Z. Tan, Y. Yang, J. Wan, H. Hang, G. Guo, and S. Li, "Attention-based pedestrian attribute analysis," *IEEE Trans. Image Process*, vol. 28, no. 12, p. 6126–6140, 2019.
[34] Y. Hu, X. Bai, P. Zhou, F. Shang, and S. Shen, "Data augmentation imbalance for imbalanced attribute classification," 2020.
[35] A. Specker, M. Cormier, and J. Beyerer, "Upar: Unified pedestrian attribute recognition and person retrieval," *Appl. Comput. Vision, WACV*, p. 981–990, 2023.
[36] X. Liu, "Hydraplus-net: Attentive deep features for pedestrian analysis," *Proc. IEEE Int. Conf. Comput. Vis*, no. c, p. 350–359, 2017.
[37] C. Tang, L. Sheng, Z. Zhang, and X. Hu, "Improving pedestrian attribute recognition with weakly-supervised multi-scale attribute-specific localization," *Proc. IEEE Int. Conf. Comput. Vis*, no. c, p. 4996–5005, 2019.
[38] J. Jia, H. Huang, W. Yang, X. Chen, and K. Huang, "Rethinking of pedestrian attribute recognition: Realistic datasets and a strong baseline," p. 1–12, 2020.
[39] J. Jia, X. Chen, and K. Huang, "Spatial and semantic consistency regularizations for pedestrian attribute recognition," *Proc. IEEE Int. Conf. Comput. Vis*, p. 942–951, 2021.
[40] M. Moghaddam, M. Charmi, and H. Hassanpoor, "Jointly human semantic parsing and attribute recognition with feature pyramid structure in efficientnets," *IET Image Process*, vol. 15, no. 10, p. 2281–2291, 2021.
[41] J. Jia, N. Gao, F. He, X. Chen, and K. Huang, "Learning disentangled attribute representations for robust pedestrian attribute recognition," *Proc. AAAI Conf. Artif. Intell*, vol. 36, no. 1, p. 1069–1077, 2022.
[42] Z. Tan, Y. Yang, J. Wan, G. Guo, and S. Li, "Relation-aware pedestrian attribute recognition with graph convolutional networks," 2020.
[43] Z. Ji, Z. Hu, Y. Wang, Z. Shao, and Y. Pang, "Reinforced pedestrian attribute recognition with group optimization," *SSRN Electron. J*, 2022.
[44] Y. Shi, Z. Wei, H. Ling, Z. Wang, J. Shen, and P. Li, "Person retrieval in surveillance videos via deep attribute mining and reasoning," *IEEE Trans. Multimed*, vol. 14, no. 8, 2020.
[45] J. Wang, X. Zhu, S. Gong, and W. Li, "Transferable joint attribute-identity deep learning for unsupervised person re-identification," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit*, no. 1, p. 2275–2284, 2018.
[46] Y. Men, Y. Mao, Y. Jiang, W. Ma, and Z. Lian, "Controllable person image synthesis with attribute-decomposed gan," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit*, p. 5083–5092, 2020.
[47] Y. Chen, S. Duffner, A. Stoian, J. Dufour, and A. Baskurt, "Pedestrian attribute recognition with part-based cnn and combined feature representations," *Jt. Conf. Comput. Vision, Imaging Comput. Graph. Theory Appl*, vol. 5, p. 114–122, 2018.
[48] X. Song, H. Yang, and C. Zhou, "Pedestrian attribute recognition with graph convolutional network in surveillance scenarios," *Futur. Internet*, vol. 11, no. 11, 2019.
[49] Q. Dong, X. Zhu, and S. Gong, "Person search by text attribute query as zero-shot learning," *Proc. IEEE Int. Conf.*

*Comput. Vis*, vol. 2019-Octob, p. 3651–3660, 2019.
[50] N. Sarafianos, C. Nikou, T. Giannakopoulos, and I. Kakadiaris, "Curriculum learning for multi-task classification of visual attributes," p. 2608–2615, 2017.
[51] S. Pei, "Multitask model for person re-identication by," 2021.
[52] N. Ly, T. Do, and B. Nguyen, "Large-scale coarse-to-fine object retrieval ontology and deep local multitask learning," *Comput. Intell. Neurosci*, vol. 2019, 2019.
[53] Y. Cao, J. Wang, and D. Tao, "Symbiotic adversarial learning for attribute-based person search," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics*, vol. 12359, p. 230–247, 2020.
[54] Z. Liu, H. Mao, C. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit*, vol. 2022-June, p. 11966–11976, 2022.
[55] Z. Wang, Z. Fang, J. Wang, and Y. Yang, "Vitaa: Visual-textual attributes alignment in person search by natural language," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics*, vol. 12357, p. 402–420, 2020.
[56] T. Ma, M. Yang, H. Rong, Y. Qian, Y. Tian, and N. Al-Nabhan, "Dual-path cnn with max gated block for text-based person re-identification," *Image Vis. Comput*, vol. 111, p. 1–14, 2021.
[57] S. Wang, R. Wang, Z. Yao, S. Shan, and X. Chen, "Cross-modal scene graph matching for relationship-aware image-text retrieval," *Appl. Comput. Vision, WACV*, p. 1497–1506, 2020.
[58] B. Jeong, J. Park, and S. Kwak, "Asmr: Learning attribute-based person search with adaptive semantic margin regularizer," *Proc. IEEE Int. Conf. Comput. Vis*, vol. 1, no. 1, p. 11996–12005, 2021.
[59] N. Maillot, M. Thonnat, and C. Hudelot, "Ontology based object learning and recognition: Application to image retrieval," *Proc. - Int. Conf. Tools with Artif. Intell. ICTAI*, no. 1, p. 620–625, 2004.
[60] N. Maillot, "Ontology based object learning and recognition-thesis," 2008.
[61] V. Mezaris, I. Kompatsiaris, and M. Strintzis, "An ontology approach to object-based image retrieval," p. –511–14, 2004.
[62] C. Hudelot, "Towards a cognitive vision platform for semantic image interpretation; application to the recognition of biological organisms," 2008.
[63] R. Contreras, O. Starostenko, V. Alarcon-Aquino, and L. Flores-Pulido, "Facial feature model for emotion recognition using fuzzy reasoning," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics*, vol. 6256, p. 11–21, 2010.
[64] R. Bashar, S. Kang, P. Dawadi, and P. Rhee, "A context-aware statistical ontology approach for adaptive face recognition," p. 704–709, 2007.
[65] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," 2019.
[66] Y. Zhang and Q. Yang, "A survey on multi-task learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 12, pp. 5586–5609, 2022.
[67] Y. Deng, P. Luo, C. Loy, and X. Tang, "Pedestrian attribute recognition at far distance," *Proc. ACM Int. Conf. Multimed*, vol. MM'14, p. 789–792, 2014.
[68] D. Li, Z. Zhang, X. Chen, and K. Huang, "A richly annotated pedestrian dataset for person retrieval in real surveillance scenarios," *IEEE Transactions on Image Processing*, vol. 28, no. 4, pp. 1575–1590, 2019.
[69] D. Li, X. Chen, and K. Huang, "Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios," p. 111–115, 2015.

**Ly Quoc Ngoc** is currently Associate Professor and Head of the Computer Vision and Cognitive Cybernetics Department at the Faculty of Information Technology, University of Science, Ho Chi Minh City. He received his Ph.D in Computer Science from VNUHCM-HCMUS, Vietnam, in 2009. His research interests lie in the fields of Computer Vision, Digital Image and Video Processing, Computer Graphics, Artificial Intelligence, Multimodal Deep Learning, Medical Image Processing.
Email: lqngoc@fit.hcmus.edu.vn



**Masayuki Fukuzawa** is currently an associate professor of Faculty of Information and Human Sciences in Kyoto Institute of Technology, Japan. His specialty is image instrumentation and processing as multidimensional signal. His current research interests include image and video processing for clinical diagnosis, optical instrumentation of semiconductor crystals, and intelligent image sensors.
Email: fukuzawa@kit.ac.jp

**Trang T.T. Phung** is currently a lecturer in the Faculty of Information Technology at Saigon University (SGU), Vietnam. Her research interests lie in the fields of Computer Vision, especially in Image Retrieval and Image Classification.
Email: trangphung@sgu.edu.vn