Learning Sparse and Identity-Preserved Hidden Attributes for Person Re-Identification

Zheng Wang, *Member, IEEE*, Junjun Jiang, *Member, IEEE*, Yang Wu, *Member, IEEE*, Mang Ye, Xiang Bai, *Senior Member, IEEE*, and Shin'ichi Satoh, *Member, IEEE*

***Abstract*— Person re-identification (Re-ID) aims at matching person images captured in non-overlapping camera views. To represent person appearance, low-level visual features are sensitive to environmental changes, while high-level semantic attributes, such as "short-hair" or "long-hair", are relatively stable. Hence, researches have started to design semantic attributes to reduce the visual ambiguity. However, to train a prediction model for semantic attributes, it requires plenty of annotations, which are hard to obtain in practical large-scale applications. To alleviate the reliance on annotation efforts, we propose to incrementally generate Deep Hidden Attribute (DHA) based on baseline deep network for newly uncovered annotations. In particular, we propose an auto-encoder model that can be plugged into any deep network to mine latent information in an unsupervised manner. To optimize the effectiveness of DHA, we reform the auto-encoder model with additional *orthogonal* generation module, along with *identity-preserving* and *sparsity* constraints. 1) *Orthogonally generating*: In order to make DHAs different from each other, Singular Vector Decomposition (SVD) is introduced to generate DHAs orthogonally. 2) *Identity-preserving constraint*: The generated DHAs should be distinct for telling different persons, so we associate DHAs with person identities. 3) *Sparsity constraint*: To enhance the discriminability of DHAs, we also introduce the sparsity constraint to restrict the number of effective DHAs for each person. Experiments conducted on public datasets have validated the effectiveness of the proposed network. On two large-scale datasets, *i.e.*, Market-1501 and DukeMTMC-reID, the proposed method outperforms the state-of-the-art methods.**

***Index Terms*— Person re-identification, attribute learning, generation, discrimination.**

Manuscript received January 30, 2019; revised May 21, 2019 and August 21, 2019; accepted October 2, 2019. Date of publication October 17, 2019; date of current version December 30, 2019. This work was supported in part by JST CREST under Grant JPMJCR1686, in part by the Grant-in-Aid for JSPS Fellows under Grant 18F18378, and in part by the Microsoft Research Asia Collaborative Research Grant. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Giulia Boato. *(Corresponding author: Yang Wu.)*

Z. Wang and S. Satoh are with the Digital Content and Media Sciences Research Division, National Institute of Informatics, Tokyo 101-8430, Japan (e-mail: wangz@nii.ac.jp; satoh@nii.ac.jp).

J. Jiang is with the School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China, and also with the Peng Cheng Laboratory, Shenzhen 518066, China (e-mail: junjun0595@163.com).

Y. Wu is with the International Collaborative Laboratory for Robotics Vision, Institute for Research Initiatives, Nara Institute of Science and Technology, Nara 630-0192, Japan (e-mail: wuyang0321@gmail.com).

M. Ye is with the Inception Institute of Artificial Intelligence, Abu Dhabi, United Arab Emirates (e-mail: mangye16@gmail.com).

X. Bai is with the School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan 430074, China (e-mail: xbai@hust.edu.cn).

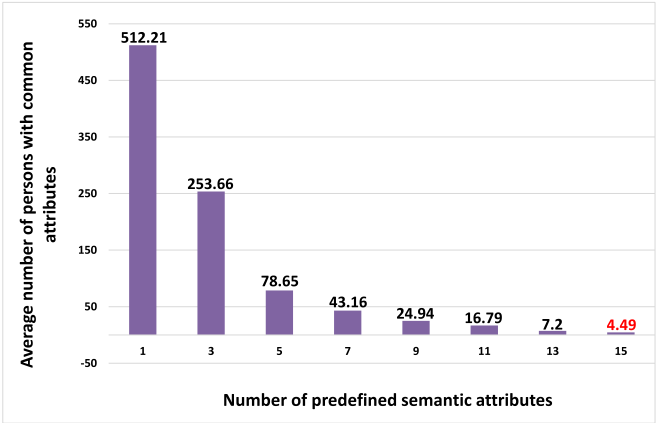
Digital Object Identifier 10.1109/TIP.2019.2946975

I. Introduction

PERSON re-identification (Re-ID) is attracting increasing attentions in computer vision and artificial intelligence community [1]–[5], due to its important application to video surveillance and criminal investigation systems [6], [7]. Re-ID seeks to match persons across non-overlapping surveillance camera views through their visual appearance. Hence, designing a good appearance descriptor is necessary and essential to the Re-ID task.

To represent person image, a lot of approaches try to design low-level visual features [8], [10], [11], which are discriminative but sensitive to appearance disturbances, such as illumination variation, scale change and viewpoint change. In comparison, high-level attributes, *e.g.*, age, gender and dressing, are relatively robust to different imaging conditions. Hence, person representation with semantic attributes start to be investigated [9], [12]–[16]. Most of these methods require predefined semantic attributes with thousands of annotated samples for each of them, and train an attribute prediction model based on such annotated data. As we know, the more types of semantic attributes we predefine and annotate, the more benefits the person representation will gain. In other words, the performance is likely to be further improved by just adding more attributes.

However, the quantity of types of semantic attributes is always limited, compared with the number of persons to be distinguished. By measuring the attribute statistics on the VIPeR dataset [17], we find that even exploiting 15 predefined semantic attributes [14], 4.49 out of totally 632 persons in average still hold common semantic attributes (as Fig. 1(a) shows). Analogously, we have 750 and 1110 different persons for the Market-1501 [8] and the DukeMTMC-reID [18] datasets, respectively, while only 27 and 23 types of semantic attributes are correspondingly designed [9], which are still not enough to separate all different persons. Fig. 1(b) gives an example that eight persons hold common semantic attribute annotations. As we know, enlarging the types of semantic attributes is a direct way to improve Re-ID performance. However, when we design a new semantic attribute, it requires massive annotated samples. For instance, if we attempt to raise only one new semantic attribute for the DukeMTMC-reID dataset, we have to re-annotate 16522 training samples. It is unrealistic in the practical application.



(a)

(b)

Fig. 1. (a) **The tendency of average number of persons with common semantic attributes V.S. the number of predefined semantic attributes.** The statistic data is counted on the VIPeR dataset (632 persons in total). As the quantity of types of attributes increases, the average number of persons with common semantic attributes will decrease. Even exploiting 15 predefined semantic attributes for 632 persons, 4.49 persons in average still hold common semantic attributes. (b) **An example to illustrate the idea of this paper.** These eight persons are selected from Market-1501 dataset [8]. All of them are with the same semantic attribute annotations. The semantic attributes are partly selected from the predefined attributes in [9]. In this paper, we propose to mine incremental hidden attributes for newly uncovered annotations, and to make attribute representation more discriminative.

To address this challenge, as Fig. 1(b) shows, we propose to create some hidden attributes without any new annotation. Generally, semantic attributes are mined from low-level features, which contains detailed descriptions of person appearance. Specifically, taking the low-level features as the input space and high-level attributes as the latent space, an encoding step is taken to simulate the process of inferring semantic and hidden attributes from visual feature *respectively*. Then, we introduce a decoding step to reconstruct the visual feature with semantic and hidden attributes *simultaneously*. To make inferred attributes sensible, the latent space is learned by minimizing the reconstruction loss of input space. In this way, we employ an auto-encoder model to the baseline deep network and mine the Deep Hidden Attribute (DHA).

As we know, deep learning based person representation has been widely exploited, and showed their strengths in Re-ID task recently. It is reasonable to add the hidden attribute prediction progress into a deep learning network. In particular, we plug-in the auto-encoder model into general deep learning network, and construct the Deep Hidden Attribute Network (DHA-Net). In the proposed DHA-Net, a new reconstruction loss is introduced to learn the hidden attributes, along with the person identities and predefined semantic attributes. To optimize the effectiveness of DHA, we reform the auto-encoder model with additional *orthogonal* generation module, along with *identity-preserving* and *sparsity* constraints.

A. Orthogonally Generating

It is reasonable to make the DHA orthogonal to each other. Inspired from [19], we add an eigenlayer into the deep framework to find a set of orthogonal projection directions by exploiting Singular Vector Decomposition (SVD). Then, the DHA is generated orthogonally in the latent part of the auto-encoder model.

B. Identity-Preserving Constraint

The DHA should be discriminative for telling different persons. Thus, we associate the generated DHA with person identities. In particular, DHA inferred from different images should be similar if the corresponding images are from the same identity, and vice versa. In this paper, we introduce the cross entropy to exploit the hidden attribute discrimination. If the generated attributes are related to the corresponding identity, they are judged as true ones. In contrast, the attributes would be judge as false, if they are not related to corresponding person identity.

C. Sparsity Constraint

Besides identity-preserving constraint, to enhance the discriminability of DHAs, we also introduce the sparsity constraint to restrict the number of effective DHAs for each person. The philosophy is that the representation of certain individual will be discriminative, if only a small number of unique attributes are activated.

The baseline deep learning network for person Re-ID with semantic attributes includes an identification loss and several semantic attribute losses [9]. In the proposed DHA-Net, a new reconstruction loss is introduced. To this end, besides person identities and predefined semantic attributes, DHAs are deeply learned simultaneously. We further reform the network with the new module and constraints described above. After that, our new network can mine more informative DHAs. The main contributions of this paper are summarized as follows:

- *The plugged-in auto-encoder model:* We plug-in the auto-encoder model to the baseline deep network, which extends novel DHAs by mining latent information from visual features for newly uncovered attribute annotations. Along with original identification loss and semantic attribute losses, a reconstruction loss is designed for DHA generation. Without more annotations, the hidden attribute generation process acts in an unsupervised manner. After optimizing, we make each layer of the network physical meaningful.
- *The eigenlayer and two constraints:* We embed the eigenlayer in the auto-encoder model. With SVD, hidden

attributes are projected orthogonally. For the DHA discrimination, a cross entropy loss associated with identities and a sparsity loss are further designed. They are optimized together. Meanwhile, the weights of eigenlayer are updated with SVD during the network training process.

- *Simple and effective:* The proposed method is very simple to construct and easy to implement. Experiments conducted on two large-scale Re-ID datasets, as indicated in [20], have validated the effectiveness of the proposed model, which outperforms most of the state-of-the-art methods.

A preliminary conference version of this paper was published in [21]. In this journal version, we propose to reform the DHA-Net with additional Identity-preserving and Sparsity constraints and the Orthogonal generation module (ISO), which make DHAs more informative and effective. We also include in this paper more related works on person identification and more experiments to show the influence of different configurations and parameters.

II. Related Work

The approaches [20], [22] give detailed reviews of person Re-ID. Here, we mainly investigate two aspects, 1) Deep learning for person Re-ID, and 2) Person Re-ID with semantic attributes.

A. Deep Learning for Person Re-ID

Deep learning methods in Re-ID task can be grouped into two categories. The first type is deep metric learning, in which image pairs, triplets, quadruplets or batch of images are fed into the network [23]–[30]. Generally speaking, this kind of methods are effective in learning image similarities in an adaptive manner, but may have efficiency problems under large-scale galleries [19]. The second one focuses on feature learning, which categorizes the training samples into predefined classes and the *FC* descriptor is used for retrieval [11], [31]–[35]. Reference [11] proposed to learn a generic feature embedding by training a classification model from multiple domains with a domain guided dropout. Reference [33] jointly learned multiple classification losses of local and global discriminative feature optimization subject to the same Re-ID labeled information. Reference [32] designed a classification model by learning powerful features over full body and body parts. Reference [31] combined the verification and classification losses together, and learned an effective presentation. Reference [34] exploited the collaboration between handcrafted and deep learning features. Reference [35] utilized the view information in the feature extraction stage. With only person identities, we take this kind of networks as the base part of Re-ID template deep learning network.

B. Person Re-ID With Semantic Attributes

High-level semantic attributes, used as auxiliary information, have been investigated in some works. Reference [12] proposed to learn a selection and weighting of predefined semantic attributes to describe persons. Reference [13] applied semantic color names to represent person. Reference [36] tuned deep Re-ID network with additional semantic attributes. Reference [37] exploited the inter-attribute correlations to improve the representation. Reference [15] combined the semantic attribute learning process into the CNN framework. Reference [9] followed the idea of [15] that learns a Re-ID embedding and predicts the pedestrian semantic attributes simultaneously. Reference [16] attempted to directly use semantic attributes to retrieval person images in a cross-modality way. As we know, human annotations are the key information for Re-ID with semantic attributes. Hence, [14] released attribute-augmented versions for different Re-ID datasets, such as VIPeR. Reference [9] released similar ones for the two largest datasets, *i.e.*, Market-1501 and DukeMTMC-reID. Fig. 2 illustrates the general framework of this kind of methods.

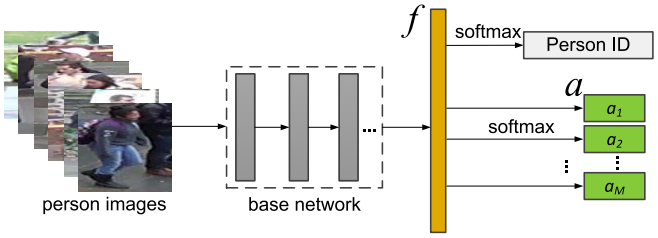


Fig. 2. **The baseline network**. The general framework of approaches with person identities and predefined semantic attributes.

Although [9] achieved the top performance among Re-ID methods with semantic attributes. Compared with the quantity of person identities, the number of the types of semantic attributes is still compromised. Designing more attributes is an effective way for better Re-ID performances. However, making more annotations is not in accord with practical situation. To this end, we propose to enlarge types of attributes with a hidden part, without additional annotations.

III. Proposed Network

As Fig. 3 shows, the proposed DHA-Net consists of a baseline network, a plugged-in auto-encoder model. In [21], there are three kinds of losses, *i.e.*, an identification loss for visual feature, several semantic attribute losses for their prediction, a loss for visual feature reconstruction. The identification loss and the semantic attribute losses are used for the feature layer, as the baseline network does. The reconstruction loss is used for the discrepancy of two feature layers. In this paper, we further reform the network with additional ISO module, where we introduce an embedded eigenlayer. In addition, a cross entropy loss and a sparsity loss for hidden attribute discrimination are taken into consideration. The cross entropy loss and the sparsity loss are used for the orthogonal hidden attribute layer. In the following subsections, we will respectively demonstrate the baseline network, the architecture of DHA-Net with ISO module, in particular the auto-encoder model, the orthogonal generation module, the identity-preserving constraint and the sparsity constraint.

A. Baseline Network

Following [9], we construct the baseline network for Re-ID with person identities and semantic attributes. The baseline

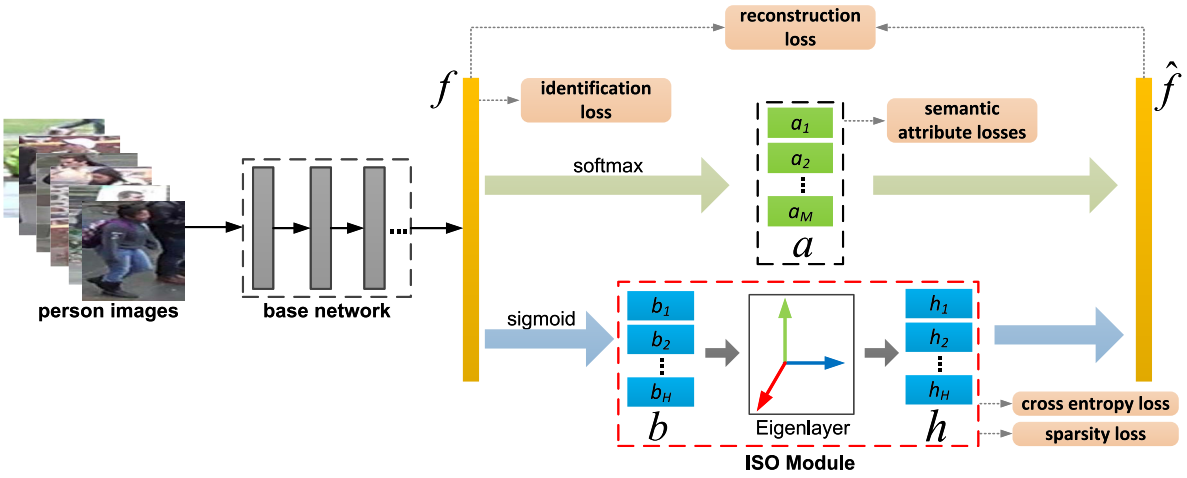


Fig. 3. **The architecture of the proposed DHA-Net with ISO module**. The DHA-Net consists of a baseline network, a plugged-in auto-encoder model and several losses, including an identification loss, several semantic attribute losses, and a reconstruction loss. The ISO module consists of an embedded eigenlayer and two kinds of losses, including a cross entropy loss and a sparsity loss. f, $\hat{f}$, a and b (h) respectively stand for feature, reconstructed feature, inferred semantic attributes and mined hidden attributes. The identification loss and semantic attribute losses are original losses designed in the baseline network. The auto-encoder model and reconstruction loss are used to generate DHAs. In addition, we substitute the DHA b, with orthogonal generation, identity-preserving and sparsity constraints. The eigenlayer is designed for hidden attribute generation, and used to project DHA to orthogonal DHA. The cross entropy loss and the sparsity loss are designed for hidden attribute discrimination. In this paper, besides f, $[a, h]$ are also treated as representations, but from the semantic attributes perspective.

network consists of a base network, and fully-connected layer for person identification and semantic attribute prediction. The survey [22] has proved that among all the base networks, such as ResNet [38] and CaffeNet [39], ResNet-50 shows to yield competitive Re-ID performance. We select ResNet-50 as the base network, and pre-trained it on ImageNet [40]. After that, we use $M + 1$ fully-connected (FC) layers followed by the softmax layers for person identification and semantic attribute prediction, where M denotes the number of types of predefined semantic attributes. We set the number of neurons in FC layer for identification to K, where K denotes the number of training identities. For a certain semantic attribute with m classes, the FC layer is m-dim.

Suppose we have N images of K identities in training stage. Each identity has M types of semantic attributes. Let $D_i = \{X_i, ID_i, Attr_i\}$ be the training set, where X_i denotes the i-th image, ID_i denotes the identity of image X_i, and $Attr_i = \{Attr_i^1, Attr_i^2, \ldots, Attr_i^M\}$ is a set of M predefined semantic attribute labels of the image X_i. Given a training sample $(X_i, ID_i, Attr_i)$, a feature description $f_i \in \mathbb{R}^d$ is first extracted from the feature layer ($pool5$), where d is 2,048-dimension. Exploiting the feature f_i, the model predicts semantic attributes a_i. For the baseline network, an identification loss is for the person ID, semantic attribute losses are for the person's semantic attributes.

1) Identification Loss: Given a training sample X, the output of person ID layer is $z = [z_1, z_2, \ldots, z_K] \in \mathbb{R}^K$. Thus the predicted probability of each ID label k is calculated as:

$$p(k|X) = \frac{exp(z_k)}{\sum_{i=1}^{K} exp(z_i)}. \tag{1}$$

The cross entropy of identification loss is formulated as below:

$$L_{ID} = -\sum_{k=1}^{K} log(p(k|X))q(k|X), \tag{2}$$

where $q(\cdot)$ stands for the ground-truth identity class distribution. Let y be the ground-truth ID label of X, so that $q(y) = 1$, and $q(k) = 0$ for all $k \neq y$. In this case, minimizing the identification loss is equivalent to maximizing the possibility of being assigned to the ground-truth class.

2) Semantic Attribute Losses: As [9], [15] did, we also use M softmax losses for semantic attributes prediction. Assume that m classes for a certain semantic attribute $s = [s_1, s_2, \ldots, s_m] \in \mathbb{R}^m$, and the probability of assigning sample X to the semantic attribute class $j \in \{1, \ldots, m\}$ will be written as

$$p_s(j|X) = \frac{exp(s_j)}{\sum_{i=1}^{m} exp(s_i)}. \tag{3}$$

Similarly, the certain semantic attribute loss of classifying sample X will be computed as $-\sum_{j=1}^{m} log(p_s(j))q_s(j)$. Let y_m denote the ground-truth attribute label, so that $q_s(y_m) = 1$, and $q_s(j) = 0$ for all $j \neq y_m$. The total is combining multiple semantic attribute losses together as:

$$L_S = -\frac{1}{M}\sum_{i=1}^{M}\sum_{j=1}^{m} log(p_s(j|X))q_s(j|X). \tag{4}$$

B. Auto-Encoder Model

We define the semantic attributes for the training sample X as $a \in \mathbb{R}^M$. Our goal is to augment a hidden part DHA $b \in \mathbb{R}^H$, and form a hybrid attribute vector $[a, b]$ for each person image, where H is the number of hidden attributes. The auto-encoder model is adopted, which has the following two characteristics: (1) information in the input feature is preserved in the reconstructed future as much as possible; (2) the hidden part is learned automatically instead of learned by classifiers. It is achieved by a two-step construction: encoding step and

decoding step. Consequently, we design a simple way to reveal the lacking expressive information in semantic attributes.

1) Encoding Step: The network encodes the original feature into hybrid attribute vector, which is composed of two parts: a predefined semantic part a obtained by general softmax predictors and a hidden part DHA b learned from visual features by an encoding function. Each component of semantic attributes a is obtained from prediction of learned attribute classifiers. The encoding function E encodes feature vector f only for the hidden attribute part as

$$b = E(f) = \phi(W_b f), \tag{5}$$

where $W_b \in \mathbb{R}^{H*d}$ is the augmentation matrix containing all the attribute augmentation parameters. $\phi(z) = 1/(1+\exp(-z))$ is a sigmoid function which ensures values of b are in a range comparable to the a. This encoding process is shown as

$$[a, b] = [a, E(f)] = [a, \phi(W_b f)]. \tag{6}$$

Decoding step. The decoding function D aims at reconstructing the input feature space from hybrid attribute vector $[a, b]$. This process is shown as Eq. 7, where $R \in \mathbb{R}^{d*(M+H)}$ is the reconstruction matrix.

$$\hat{f} = D([a, b]) = R[a, b]. \tag{7}$$

2) Reconstruction Loss: The reconstruction loss measures the loss incurred in the reconstruction of input feature vectors of all samples, which is used to guide the learning of W_b. We use a squared error loss [41] as Eq. 8.

$$L_R = \|\hat{f} - f\|^2 = \|R[a, \phi(W_b f)] - f\|^2. \tag{8}$$

With help of the plugged-in auto-encoder model with reconstruction loss, the Re-ID network gains the ability of DHA generation. By using the identity loss, the semantic attribute loss and the feature reconstruction loss, the DHA-Net is trained to predict person identity and semantic attribute labels, and to reconstruct feature. The total loss function is defined as:

$$L = \alpha_1 L_{ID} + \alpha_2 L_S + \alpha_3 L_R, \tag{9}$$

where L_{ID}, L_S and L_R denote the cross entropy loss of identity classification and semantic attribute prediction, and the squared error loss of feature reconstruction respectively. Parameters α_1, α_2 and α_3 ($\alpha_1, \alpha_2, \alpha_3 > 0$) balance the contributions of the three losses and is determined on a validation set.

In the following, as Fig. 3 shows, we substitute the DHA b, with orthogonal generation module, identity-preserving and sparsity constraints. The eigenlayer is designed for hidden attribute generation, and used to project DHA to orthogonal DHA. The cross entropy loss and the sparsity loss are designed for hidden attribute discrimination.

C. Hidden Attribute Orthogonal Generation

SVD for Eigenlayer. After encoding process, we add an eigenlayer to project hidden attribute vector to be an orthogonal one, as shown in Fig. 3. The Eigenlayer contains an orthogonal weight matrix and is a linear layer without bias. The reason for not using bias is that the bias will disrupt the learned orthogonality [19]. During training, the input DHA b is passed through the Eigenlayer. Its inner products with the weight vectors of the Eigenlayer form the output orthogonal DHA $h \in \mathbb{R}^H$. The optimization method of the Eigenlayer is learned from [19].

- Step 1-Decorrelation. We perform SVD on the weight matrix as follows:
$$W_e = USV^\top, \tag{10}$$
where W_e is the weight matrix of the Eigenlayer, U is the left unitary matrix, S is the singular value matrix, and V is the right-unitary matrix. After the decomposition, we replace W_e with US. Then the linear layer uses all the eigenvectors of $W_e W_e^\top$ as weight vectors and is named as Eigenlayer.
- Step 2-Restraint. The backbone model is fine-tuned till convergence, but the Eigenlayer is fixed.
- Step 3-Relaxation. The fine-tuning goes on for some more epochs with Eigenlayer unfixed.

After Step 1 and Step 2, the weight vectors are orthogonal, *i.e.*, in an eigen state. But after Step 3, *i.e.*, relaxation training, W_e shifts away from the eigen state. The training procedure works in restraint and relaxation iteration.

Then, the total attributes become to $[a, h]$. The reconstruction is based on the hybrid attribute vector $[a, h]$, and the reconstruction loss becomes to Eq. 11.

$$L_R = \|\hat{f} - f\|^2 = \|R[a, h] - f\|^2. \tag{11}$$

With help of the embedded eigenlayer with SVD, the ability of hidden attribute orthogonal generation is acquired for the Re-ID network.

D. Identity-Preserving and Sparsity Constraints

Considering that the generated DHA should be suitable for Re-ID independently, we also have to design a strategy for its discrimination. An easy way is to re-exploited the person ID label to train the DHA representative for identification. Similar to the identification Loss, we propose a cross entropy loss for hidden attribute. As Fig. 3 shows, we connect a person ID layer after orthogonal DHA layer, and use the softmax activation to construct the loss.

1) Cross Entropy Loss: Following the identification loss described above, we suppose that the output of person ID layer is $\bar{z} = [\bar{z}_1, \bar{z}_2, \ldots, \bar{z}_K] \in \mathbb{R}^K$. So the predicted probability of each ID label k is also calculated as Eq. 1. To simplify the equation, the cross entropy loss is formulated as below:

$$L_C = -\sum_{k=1}^{K} log(p_c(k)) q_c(k). \tag{12}$$

Let y be the ground-truth ID label, so that $q_c(y) = 1$, and $q_c(k) = 0$ for all $k \neq y$. In this case, minimizing the cross entropy loss is equivalent to maximizing the possibility of being assigned to the ground-truth class.

2) Sparsity Loss: In addition, to enhance the discriminability of DHAs, we also introduce the sparsity constraint to restrict the number of effective DHAs for each person. It is defined as follows:

$$L_{Sparsity} = |h|. \tag{13}$$

The total loss is a combination of the cross entropy loss and the sparsity loss for the DHA h.

$$L_D = L_C + L_{Sparsity}. \tag{14}$$

With help of the cross entropy loss and the sparsity loss, the ability of hidden attribute discrimination is acquired for the Re-ID network.

E. Implementation Procedure

By using the identity classification loss function, the multiple semantic attribute prediction loss function, the feature reconstruction loss function, the cross entropy loss function and the sparsity loss function, the DHA-Net with ISO module is trained to predict person identity and semantic attribute labels, to mine DHA, and to reconstruct feature. The final loss function is defined as:

$$L = \alpha_1 L_{ID} + \alpha_2 L_S + \alpha_3 L_R + \alpha_4 L_D, \tag{15}$$

where L_{ID}, L_S, L_R and L_D denote the loss of identity classification and semantic attribute prediction, the squared error loss of feature reconstruction, and the cross entropy loss of generated hidden attribute discrimination respectively. Parameters α_1, α_2, α_3, and α_4 ($\alpha_1, \alpha_2, \alpha_3, \alpha_4 > 0$) balance the contributions of the four kinds of losses and can be determined on a validation set. In essence, L_{ID} leads the feature layer representation discriminative. L_S guides the prediction of predefined attributes robust. L_R gifts the network the ability of generating hidden attributes. L_D provides the network the ability of discriminating hidden attributes.

To this end, for the DHA-Net with ISO module, different kinds of losses are optimized simultaneously. We utilize Tensorflow [42] and Keras [43] to implement the codes of baseline network, the DHA-Net and the ISO module. The training process includes four steps. Note that Step ③ and Step ④ process in iteration.

- Step ①: We use ResNet-50 as the base network. The base network is pre-trained on ImageNet [40].
- Step ②: The baseline network is constructed taking advantage of the trained ResNet-50, which is essentially the same as [9]. Network parameters in trained ResNet-50 are shared to the baseline network. The baseline network is fine-tuned using the currently available semantic attribute annotations and identity labels. To avoid over-fitting, a dropout layer is inserted before the FC layers, and the dropout rate is 0.9. After training the baseline network, its network parameters are shared to the DHA-Net.
- Step ③: We re-use semantic attribute annotations and identity labels to fine-tune the DHA-Net with ISO module. The batch size is set to 64. Learning rate is initialized to 0.001. The stochastic gradient descent (SGD) is implemented in each mini-batch to update the parameters. We set the number of epochs to 120.
- Step ④: For each 10 epochs, we decompose W_e with SVD decomposition, and then the linear layer uses all the eigenvectors of $W_e W_e^\top$ as weight vectors. Then, we continue fine-tuning the network.

During testing, the outputs will be extracted respectively from the activations of the feature layer (f) and the attribute layer ($[a, h]$). The L2 normalized outputs are taken as the person representations, and Euclidean metric is utilized to measure the distances between representations. Note that although it takes more time to train the DHA-Net and the ISO module, compared with the baseline network, the testing time does not change too much, due to the tiny change for extracting person representation.

IV. Experiments

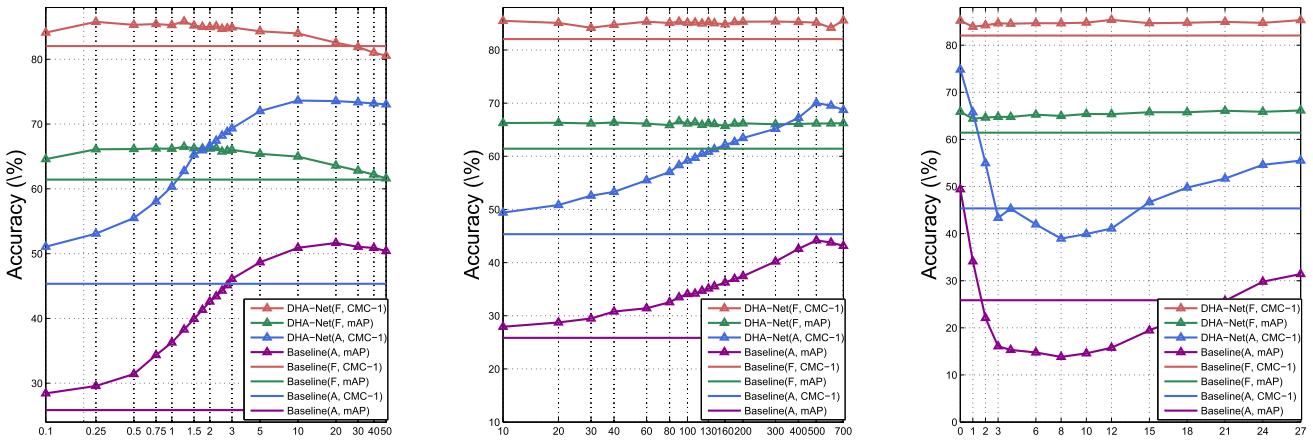
A. Datasets and Evaluation Protocol

1) Image Datasets: The Market-1501 dataset [8] contains 32,668 annotated bounding boxes of 1,501 identities. Images are captured from six cameras. As far as we know, it is one of the most popular and largest person Re-ID dataset. The Market-1501 dataset is split into 751 identities for training and 750 identities for testing. The DukeMTMC-reID dataset [18] is a subset of the DukeMTMC dataset [44]. It contains 1,812 identities captured from eight cameras. A number of 1404 identities appear in more than two cameras. The rest 408 are distractor images. Following the evaluation protocol in [18], the training and testing subsets both have 702 identities. There are 16,522 training images, in which 2,228 are query images, and 17,661 are gallery images.

2) Attribute Annotations: We used the predefined semantic attributes designed in [9]. The authors manually annotated the Market-1501 and DukeMTMC-reID datasets with attribute labels. For Market-1501 dataset, they labeled 27 types of semantic attributes, such as gender (male, female), length of lower-body clothing (long, short), carrying backpack (yes, no) and so on. For DukeMTMC-reID dataset, they labeled 23 types of semantic attributes.

3) Evaluation Metrics: In common, the Cumulative Matching Characteristic (CMC) [45] value and the mean average precision (mAP) [22] are used to evaluate the results for the Re-ID task. Given a query, its average precision (AP) is computed from its ranking result or precision-recall curve. Then, the mAP calculates the mean value of precisions of all queries. The presumption is that CMC reflects retrieval precision (CMC value at rank 1 is mainly evaluated, *i.e.*, CMC-1), and mAP reflects the recall.

4) Parameters Setting: As Eq. 15 shows, α_1, α_2, α_3 and α_4 are key parameters balancing the contributions of different losses. When $\alpha_3 = 0$, L_R is removed from the loss function, and the DHA-Net degenerates to the baseline network. When $\alpha_2 = 0$ in further, L_S is removed from the loss function, and the baseline network reduces to Resnet-50. The approach [9] had an observation on the setting of α_1 and α_2. When $\alpha_1 = 8 * M * \alpha_2$, a relatively higher Re-ID performance can be obtained for the baseline. Hence, in this paper, we set $\alpha_2 = 0.1$, $\alpha_1 = 21.6$ and 18.4, respectively for Market-1501 ($M = 27$) and DukeMTMC-reID ($M = 23$). In this paper, to simply the method, we set α_3 and α_4 as fixed values. If not specified, we set $\alpha_3 = \alpha_4 = \alpha_1/2$, *i.e.*, $\alpha_3 = \alpha_4 = 10.8$ and $\alpha_3 = \alpha_4 = 9.2$ respectively.



(a) Strength of reconstruction loss α_3/α_1 (b) Number of hidden attributes H (c) Number of predefined attributes M

Fig. 4. **Evaluation on different parameters, with experiments on the Market-1501 dataset**. In the figures, 'DHA-Net' and 'Baseline' respectively denote the accuracy curves generated by DHA-Net and the baseline network. The suffixes 'F' and 'A' respectively stand for evaluations with the output of feature layer and attribute layer as person representation. CMC-1 and mAP results are respectively demonstrated in all the evaluations. In order to facilitate comparison, the results of baseline are drawn as well. (1) Evaluation as the value of α_3 changes. (2) Evaluation on different number of hidden attributes H. (3) Evaluation on different number of predefined attributes M. The number of predefined attributes decreases from 27 to 0. Note that this figure was used in the conference version [21]. We make the analysis more clear in this version.

B. Investigation on the Parameters of DHA-Net

The proposed framework consists of different components, which give different contributions to the Re-ID model. These components are balanced by the corresponding parameters. In this subsection, we evaluate each component by adjusting the corresponding parameter. Here, we take the Market-1501 dataset as an example. We fix two of the following parameters, change the value of the other observed parameter, and drew its Re-ID results. Note that, in previous researches, the output of the feature layer is taken as the sole person representation. Although the baseline [9] has introduced a lot of predefined semantic attributes, it still only took the output of the feature layer as the person representation. The output of the attribute layer has not been taken into evaluation. Indeed, the semantic attributes are not discriminative enough. Since multiple hidden attributes are generated in DHA-Net, we evaluated the Re-ID performances of both views of the representations, by the output of the feature layer and the output of attribute layer.

- α_3: the strength of reconstruction loss
- H: the number of hidden attributes
- M: the number of predefined attributes

1) The Influence of the DHA Component: We first fixed M and H, then changed the value of parameter α_3. As Eq. 9 described, the parameter α_3 determines the strength of reconstruction loss L_R. Larger the value is, more contribution will be made by the auto-encoder part of DHA-Net, consequently DHAs will be more influential to the Re-ID performance. We evaluated the Re-ID performances by the output of feature layer and the output of attribute layer. We set α_1 and α_2 as recommended values. As the number of predefined semantic attributes $M = 27$, we fixed the number of hidden attributes H as 60, and changed the value of α_3 based on α_1 step by step, *i.e.*, $0.1 * \alpha_1$, $0.5 * \alpha_1$, $1 * \alpha_1$, $10 * \alpha_1$, $50 * \alpha_1$, and so on. Fig. 4(a) shows the Re-ID accuracy curves. We can find that the accuracy of the output of attribute layer 'DHA-Net (A)' increases as the value of α_3 becomes larger when $\alpha_3/\alpha_1 < 10$. It proves that more weights the reconstruction loss has, more benefits the output of attribute layer will gain. We can also find that although the accuracy of the output of feature layer 'DHA-Net (F)' goes down a little, DHA-Net obtains a considerable improvement, compared with the baseline method (nearly 3% CMC-1 and 5% mAP promotions in average when $\alpha_3/\alpha_1 < 10$). Note that we selected $H = 60$ because our focus is on demonstrating the function of DHA component and the influences to DHA-Net as parameters change, rather than finding out the most effective parameters. We believe that when we chose different H values, the tendencies of the curves would be similar.

2) The Number of the DHA Component: We fixed M and α_3, then changed the value of parameter H. The number of hidden attributes H is essential to the discrimination of DHAs, which is very important to be investigated. We also evaluated the Re-ID performances of both two views of the representations, by the output of feature layer and the output of attribute layer. We set α_1 and α_2 as recommended values, and fixed $\alpha_3/\alpha_1 = 0.5$. Then, we changed the number of hidden attributes H. Fig. 4(b) show the Re-ID accuracy curves. We can find that the Re-ID accuracy of the attribute layer 'DHA-Net (A)' goes better as H becomes larger when $H < 500$. That is to say, the hidden attributes are useful to promote the discrimination ability of the output of attribute layer. We can also find that, with fixed α_3, the number of hidden attributes does not intensively make the result of the output of feature layer 'DHA-Net (F)' fluctuated. Nevertheless, similar to the evaluation above, the Re-ID performance of the output of feature layer has a considerable improvement with DHA-Net, compared with the baseline network.

3) The Number of the Semantic Attribute Component: We fixed H and α_3, then changed the value of parameter M. As we know, the number of predefined semantic attributes M stands for the number of attribute annotations. In a practical

situation, we cannot always obtain so many types of semantic attributes and their annotations. In this situation, we investigate whether the DHA-Net is still effective with less or no predefined semantic attributes. We set the number of hidden attributes as $H = 60$. We also set α_1 and α_2 as recommended values, and fixed $\alpha_3/\alpha_1 = 0.5$. Then, we changed the number of predefined semantic attributes M. Fig. 4(c) shows the Re-ID accuracy curves. We can find that the change of M does not intensively influence the results of the output of the feature layer 'DHA-Net (F)'. Whereas, the results of the output of attribute layer 'DHA-Net (A)' go down and up, as the number of semantic attributes M changes from 27 to 0. As one of the important parts of the attribute layer, the predefined semantic attributes will reduce their contribution as the number goes small. Hence, we can find that the curves 'DHA-Net (A)' firstly go down. However, the curves again go up when $M < 4$. We consider the reason may be that the hidden attribute part starts to take the responsibility of the semantic part, when the number of semantic attributes is too small. The accuracy of the attribute layer even performs very well, when $M = 0$. $M = 0$ means that there are no predefined semantic attributes. It is interesting that even without predefined semantic attributes, the accuracy of the feature layer 'DHA-Net (F)' obtains 85.2% CMC-1 and 65.85% mAP value, which outperforms that of the baseline. It tells us that the plugged-in auto-encoder model will refine the feature layer, during the process of mining latent information to form hidden attributes. We consider that DHA-Net may learn something similar to the predefined semantic attributes. It should be mentioned that DHA-Net with $H = 60$ hidden attributes outperforms that with the combination of $M = 27$ predefined attributes and $H = 60$ hidden attributes (as the blue and purple curves show). The reason is that when $M = 0$, $\alpha_1 = 8 * M * \alpha_2 = 0$, and we set $\alpha_3 = 0.5 * \alpha_2$, and then relatively more weights are assigned to the reconstruction loss. In this situation, DHA-Net learns more discriminative and robust hidden attributes.

C. Comparison With the State-of-the-Art Methods

In this subsection, we make comparisons with the state-of-the-art methods. We exploit the Market-1501 and the DukeMTMC-reID datasets. For the baseline network [9], we set the training batch size as 64 and chose Tensorflow as our deep learning platform. Note that the reported batch size of the baseline network [9] is 128 and the deep learning platform is different. In this situation, the baseline network results are a bit smaller than the reported ones in [9]. Whatever, the results are still competitive. Besides the results of the state-of-the-art methods, we list the results of the baseline, DHA-Net, and DHA-Net+ISO, to prove the effectiveness of the DHA-Net and the ISO module.

1) The Market-1501 Dataset: In Table. I, we compare with existing state-of-the-art methods. Among these methods, (1) SSDAL [36], ACRN [51], and APR [9] are typical methods utilizing semantic attribute annotations. APR [9] is our referred baseline method with an attribute re-weighting module. 'w/o ARM' denotes APR without the attribute re-weighting module.

TABLE I
COMPARISON OF OUR METHOD WITH STATE-OF-THE-ART ON THE MARKET-1501 DATASET. 'A' AND 'F' DENOTE RESULTS RESPECTIVELY BY THE OUTPUT OF ATTRIBUTE LAYER AND FEATURE LAYER. 'AGGR' MEANS AGGREGATION RESULTS OF ATTRIBUTE AND FEATURE VIEWS. 'DHA-NET' STANDS FOR THE BASELINE NETWORK WITH AUTO-ENCODER, AND 'ISO' STANDS FOR IDENTITY-PRESERVING AND SPARSITY CONSTRAINTS, AND ORTHOGONAL GENERATION MODULE

Methods	Ref	CMC-1	mAP
SSDAL [36]	ECCV'16	39.40	19.60
LOMO+XQDA [10]	CVPR'15	43.79	22.22
SCSP [46]	CVPR'16	51.90	26.35
Null [47]	CVPR'16	61.02	35.68
LSTM [48]	ECCV'16	61.60	35.30
Re-ranking [49]	CVPR'17	77.11	63.63
GAN [18]	ICCV'17	78.06	56.23
SSM [50]	CVPR'17	82.21	68.80
ACRN [51]	CVPRW'17	83.61	62.60
MGCAM [52]	CVPR'18	83.55	74.25
JLML [33]	IJCAI'17	83.90	64.40
One-Example$^{\#}$ [53]	TIP'19	55.80	26.20
Full-Supervised$^{\#}$ [53]	TIP'19	83.10	63.70
AACN [54]	CVPR'18	85.90	66.87
APR (w/o ARM) [9]	PR'19	85.71	66.59
APR [9]	PR'19	87.04	66.89
ResNet-50	Ours	73.90	47.78
Baseline (A)	Ours	45.34	25.85
DHA-Net (A)	Ours	79.82	55.52
DHA-Net+ISO (A)	Ours	81.90	57.11
Baseline (F)	Ours	82.04	61.43
DHA-Net (F)	Ours	85.84	66.50
DHA-Net+ISO (F)	Ours	85.88	66.49
DHA-Net (Aggr)	Ours	86.15	67.22
DHA-Net+ISO (Aggr)	Ours	88.19	70.08
HACNN* [55]	CVPR'18	90.86	75.57
HACNN*+DHA-Net	Ours	**91.27**	**75.95**

(2) XQDA [10], SCSP [46], and Null [47] are typical methods focusing on learning distance metrics, while our method only uses the standard Euclidean metric. (3) Some methods design an additional re-ranking step for Re-ID tasks, such as Re-ranking [49] and SSM [50]. (4) The rest methods try to design special and promotional deep learning networks for Re-ID task, such as LSTM [48], GAN [18], JLML [33], MGCAM [52], AACN [54], and HACNN [55], while our baseline just takes the simple ResNet-50 as the base network. The results of 'Baseline (A)' and 'Baseline (F)' are obtained by setting $\alpha_3 = 0, \alpha_4 = 0$, where the network degenerates to the baseline network. The results of Base ResNet-50 are obtained by setting $\alpha_2 = 0$ in further, where L_S is removed from the loss function, and the network reduces to ResNet-50. Note that One-Example [53] proposed a network to address the challenge that each identity has only one labeled example along with many unlabeled examples. In this paper, its upper

bound base network achieves 83.1% in CMC-1, which is much higher than our base network does. The reason is that we set the training batch size as 64 and chose Tensorflow as our deep learning platform, which is different from [53].

For our methods, we evaluated the baseline network, DHA-Net, *i.e.*, the baseline network with auto-encoder. We also evaluated DHA-Net with ISO module, *i.e.*, Identity-preserving and Sparsity constraints, and Orthogonal generation module. For the symbols in the table, the suffixes 'A' and 'F' denote results respectively by the output of attribute layer and feature layer. 'Aggr' means aggregation results of attribute and feature views. To achieve state-of-the-art results, we reformed HACNN [55] with our DHA-Net, and fine-tuned the network with semantic attribute annotations. The method is denoted as HACNN+DHA-Net. Note that HACNN is marked with a asterisk, which means that the method is implement with Torchreid.[1]

The table shows that our method improves the results of the baseline with a big margin. Taking the output of attribute layer as the representation, the accuracy promotions are over 36% for CMC-1 and 31% for mAP. Taking the output of feature layer as the representation, the accuracy promotions are over 3% for CMC-1 and 5% for mAP. The table also shows that DHA-Net improves the results of the baseline, and DHA-Net+ISO improves the results of DHA-Net. 'DHA-Net+ISO (A)' and 'DHA-Net+ISO (F)' respectively stand for attribute and feature views. The CMC-1 value of 'DHA-Net+ISO (F)' performances better than all the other methods. Motivated by the philosophy of multi-view verification, DSRA [56] aggregates methods with different views. We utilized this aggregation method to fuse the ranking results of these two views together. We can see that DSRA works effectively, and it proves that 'DHA-Net+ISO (A)' and 'DHA-Net (A)'learn discriminative representations different from 'DHA-Net+ISO (F)' and 'DHA-Net (F)'. We can also see that HACNN combined with our method, HACNN+DHA-Net, obtained the state-of-the-art performance.

2) The DukeMTMC-reID Dataset: The evaluation follows the process on the Market-1501 dataset. In Table. II, we compare with other state-of-the-art methods, such as LOMO+XQDA [10], GAN [18], EquiDML [57], ACRN [51], AACN [54], and HACNN [55]. Among these methods, ACRN [51] and APR [9] are typical methods utilizing semantic attribute annotations. APR [9] is our referred baseline method with an attribute re-weighting module. 'w/o ARM' denotes APR without the attribute re-weighting module. When we set $\alpha_3 = 0$ and $\alpha_4 = 0$, the network degenerates to the baseline network. Then we obtain the results of 'Baseline (A)' and 'Baseline (F)'. When we further set $\alpha_2 = 0$, the baseline network reduces to Resnet-50. For the symbols in the table, the suffixes 'A' and 'F' denote results respectively by the output of attribute layer and feature layer. 'Aggr' means aggregation results of attribute and feature views. To achieve state-of-the-art results, we also reformed HACNN [55] with our DHA-Net, and fine-tuned the network with semantic attribute annotations.

[1]We refer to Torchreid, which is a library built on PyTorch for deep-learning person re-identification. The code link: https://github.com/KaiyangZhou/deep-person-reid

TABLE II
COMPARISON OF OUR METHOD WITH STATE-OF-THE-ART ON THE DUKEMTMC-REID DATASET. 'A', 'F', 'AGGR', 'DHA-NET' AND 'ISO' ARE THE SAME DENOTATIONS AS IN TABLE. I

Methods	Ref	CMC-1	mAP
LOMO+XQDA [10]	CVPR'15	30.75	17.04
GAN [18]	ICCV'17	67.68	47.13
EquiDML [57]	PR'18	70.24	52.42
ACRN [51]	CVPRW'17	72.58	51.96
AACN [54]	CVPR'18	76.84	59.25
APR (w/o ARM) [9]	PR'19	73.56	54.79
APR [9]	PR'19	73.92	55.56
Baseline (A) [9]	Ours	30.77	16.21
DHA-Net (A)	Ours	66.94	47.11
DHA-Net+ISO (A)	Ours	68.57	49.08
Baseline (F) [9]	Ours	69.46	50.33
DHA-Net (F)	Ours	72.96	53.08
DHA-Net+ISO (F)	Ours	73.01	53.08
DHA-Net (Aggr)	Ours	73.17	53.34
DHA-Net+ISO (Aggr)	Ours	74.18	54.52
HACNN* [55]	CVPR'18	80.10	63.17
HACNN*+DHA-Net	Ours	**81.32**	**64.11**

The table shows that our method improves the results of the baseline with a big margin. Taking the output of attribute layer as the representation, the accuracy promotions are over 37% for CMC-1 and 32% for mAP. Taking the output of feature layer as the representation, the accuracy promotions are over 3% for CMC-1 and 2% for mAP. The table also shows that DHA-Net improves the results of the baseline, and DHA-Net+ISO improves the results of DHA-Net. Again, we utilized the DSRA [56] method to aggregate the ranking results of feature view and attribute view together. As the proved on the Market-1501 dataset, 'DHA-Net+ISO (A)' and 'DHA-Net (A)'learn discriminative representations different from 'DHA-Net+ISO (F)' and 'DHA-Net (F)'. We can also see that HACNN combined with our method, HACNN+DHA-Net, got the state-of-the-art performance.

D. Analysis and Visualization of DHA

To investigate the genuine meaning of DHA, we visualized the class activation mapping of some attributes by the Global Average Pooling (GAP) method [58]. In Fig. 5, we show the activation maps of three types of predefined attributes, *i.e.*, "wearing backpack", "black lower-body clothing" and "white upper-body clothing", and we also show four randomly selected hidden attributes, named as "DHA1", "DHA2", "DHA3" and "DHA4". From the figure, we can find that "DHA1" pays attention to the central part of the upper-body and some background area. "DHA2" activates some local area, such as legs or feet. "DHA3" focuses on the strap of the backpacks in the front/side view. It should be noted that the

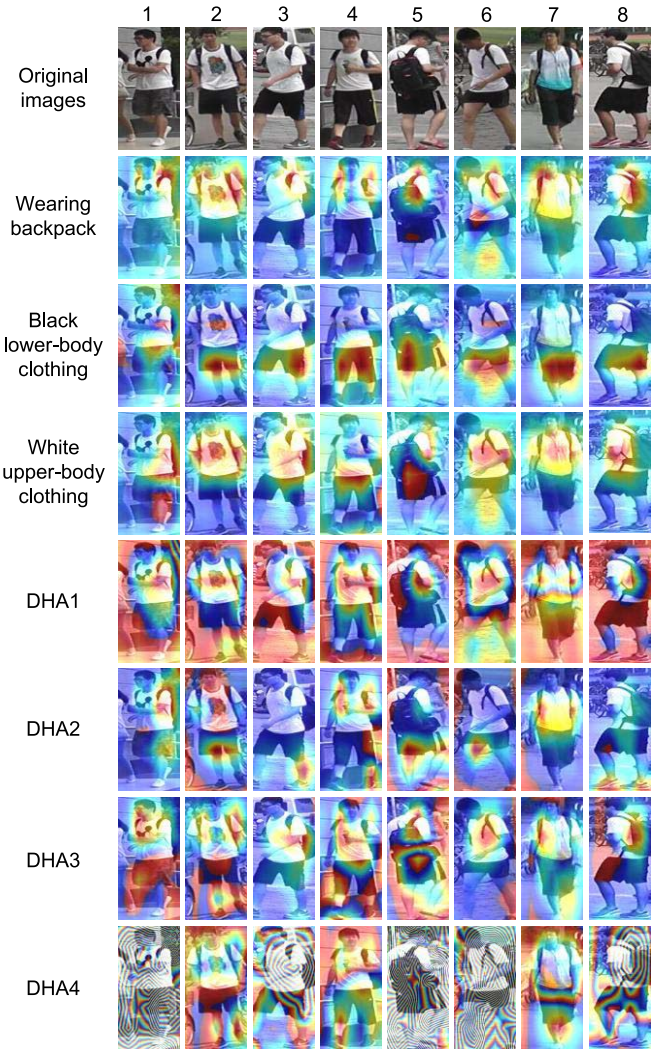


Fig. 5. **Class activation mapping images of different attributes**. We list the original image and the generated class activation mapping images of eight persons. Each column stands for one person. The top row is the original images. The following three rows are generated by predefined semantic attributes, *i.e.*, "wearing backpack", "black lower-body clothing" and "white upper-body clothing". The bottom four rows are generated by hidden attributes. We can find that "DHA1" focuses on the central area of the upper-body and some background areas, "DHA2" focuses on some local areas, "DHA3" focuses on the strap of the backpacks in the front/side view, and "DHA4" focuses on the body with the front viewpoint.

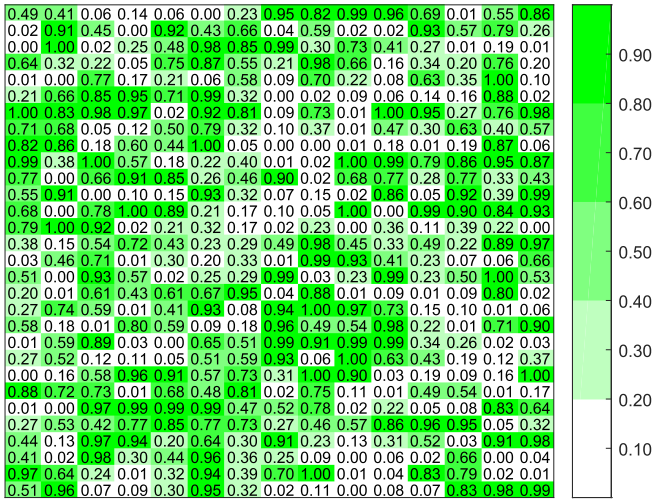


Fig. 6. **The sparsity of DHA**. We randomly selected 30 samples and their 15 DHAs from the generated results. Then, we drew the heat map with the value of DHAs. Each row stands for the DHAs of one person, and each column denotes one type of DHA.

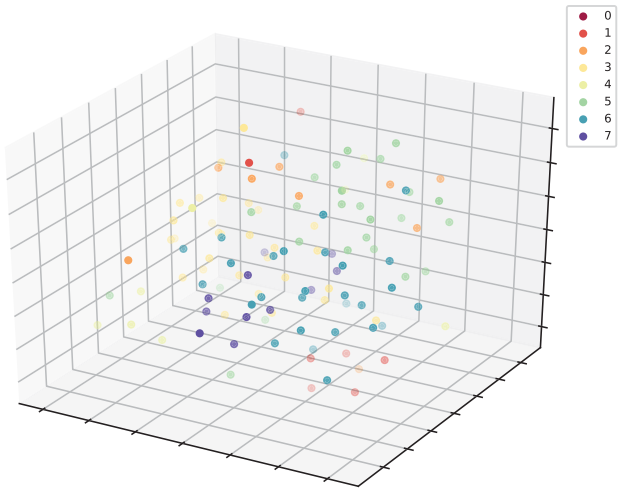
fifth person with the back viewpoint has a confused activation map. For the "DHA4", the activation maps of the second, fourth and seventh persons highlight the whole body. Note that these three persons are with the front viewpoint. The activation maps of the other persons with the side/back viewpoint are in chaos. These incremental DHAs are activated from particular areas of the body, different from that of semantic attributes. These areas indicate some latent information for corresponding DHA, just as the activation areas for semantic attributes. Basically, different areas of the body contain different kinds of details of the person's appearance. Hence, we consider that the DHA-Net is able to mine useful information to produce DHAs. In addition, the activated areas of DHAs are different from each other. It shows that the incremental DHAs are orthogonal.

To show the effectiveness of the sparsity of DHA, we randomly selected 30 samples and their 15 DHAs from the generated results (in total 50 types of DHAs). Then, we drew the heat map with the value of DHAs as Fig. 6 shows. For the Fig. 6, each row stands for the DHAs of one person, and each column denotes one type of DHA. We can see that the results are sparse.

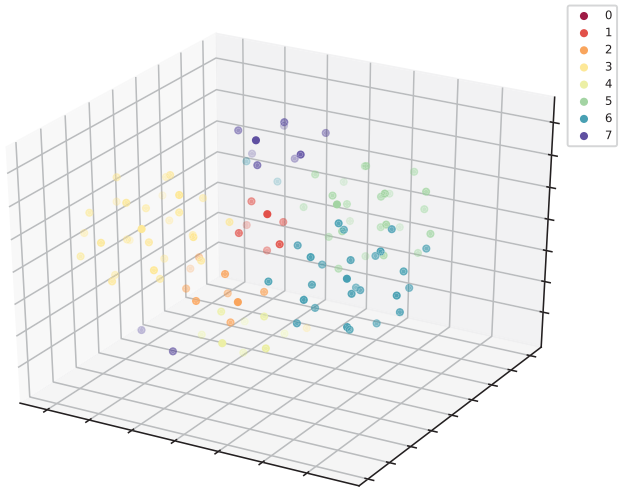
To investigate the discrimination effectiveness of DHA, we also selected images of the eight persons in Fig 1(b), and perceptually visualized their distribution in the PCA 3D attribute space through exploiting t-SNE [59]. As Fig. 7 shows, selected samples are drawn in the PCA 3D attribute space without DHA and with DHA. It shows that samples of different persons are mixed with other and indistinguishable in Fig. 7(a). Whereas, in Fig. 7(b), samples of the same person get together and samples of different persons are relatively separated apart. It shows that the mined hidden attributes do improve the discrimination ability of the model.

E. Evaluation on the Running Cost

We evaluated the running cost of our method in this subsection. We used the NVIDIA Tesla K80 12G (GPU) to extract image features and attributes, and took Market-1501 as an example. In the testing stage, $3,368$ query images and $19,732$ gallery images were used. It costs $128,849$ ms to extract all the features and attributes. Hence, the method takes 5.578 ms in average to extract features and attributes for each image. As we know, a real time monitor application runs 40 ms per frame. If we have less than 8 persons in average for each frame, our method can fulfill the high speed requirement. In addition, it costs $558,969$ ms to conduct all queries. In total, $3,368$ queries were evaluated. So the proposed method takes 165.96 ms in average to obtain a



(a) Without DHA



(b) With DHA

Fig. 7. **3D attribute space visualization of some representative samples**. These samples are selected from 8 persons, as shown in Fig. 1(b). The top figure is the attribute space (PCA to 3D) with only defined semantic attributes, while the bottom figure is the attribute space with defined semantic attributes and DHAs together. In these two attribute space figures, points with different colors indicate different persons.

ranking list. We can see that the proposed method is very fast.

V. Conclusion

In the preliminary conference version, to incrementally generate DHA, we contribute a simple deep learning network, called DHA-Net. In this paper, we reform the DHA-Net. We embed an eigenlayer with SVD, so that the network obtains the ability of hidden attribute *orthogonal* generation. In addition, we supplement a cross entropy loss and a sparsity loss to make the network acquire the ability of hidden attribute *identity-preserving* and *sparsity* discrimination. After optimizing together, we easily boost the Re-ID performance by incremental deep hidden attributes without additional attribute annotations. Comprehensive experiments demonstrate that DHA-Net with ISO module not only achieves better attribute representation but also improves the feature representation.

References

[1] Y.-C. Chen, W.-S. Zheng, and J. Lai, "Mirror representation for modeling view-specific transform in person re-identification," in *Proc. Int. Joint Conf. Artif. Intell.*, Jun. 2015, pp. 3402–3408.

[2] Z. Wang, R. Hu, Y. Yu, J. Jiang, C. Liang, and J. Wang, "Scale-adaptive low-resolution person re-identification via learning a discriminating surface," in *Proc. Int. Joint Conf. Artif. Intell.*, Jul. 2016, pp. 2669–2675.

[3] Z. Wang *et al.*, "Zero-shot person re-identification via cross-view consistency," *IEEE Trans. Multimedia*, vol. 18, no. 2, pp. 260–272, Feb. 2016.

[4] J. Jiao, W.-S. Zheng, A. Wu, X. Zhu, and S. Gong, "Deep low-resolution person re-identification," in *Proc. Assoc. Adv. Artif. Intell.*, Apr. 2018, pp. 6967–6974.

[5] Z. Wang *et al.*, "Person reidentification via discrepancy matrix and matrix metric," *IEEE Trans. Cybern.*, vol. 48, no. 10, pp. 3006–3020, Oct. 2018.

[6] X. Zhu, X.-Y. Jing, X. You, W. Zuo, S. Shan, and W.-S. Zheng, "Image to video person re-identification by learning heterogeneous dictionary pair with feature projection matrix," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 3, pp. 717–732, Mar. 2018.

[7] Z. Wang, J. Jiang, Y. Yu, and S. Satoh, "Incremental re-identification by cross-direction and cross-ranking adaption," *IEEE Trans. Multimedia*, vol. 21, no. 9, pp. 2376–2386, Sep. 2019.

[8] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1116–1124.

[9] Y. Lin *et al.*, "Improving person re-identification by attribute and identity learning," *Pattern Recognit.*, vol. 95, pp. 151–161, Nov. 2019.

[10] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 2197–2206.

[11] T. Xiao, H. Li, W. Ouyang, and X. Wang, "Learning deep feature representations with domain guided dropout for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1249–1258.

[12] R. Layne, T. M. Hospedales, S. Gong, and Q. Mary, "Person re-identification by attributes," in *Proc. BMVC*, 2012, vol. 2, no. 3, p. 8.

[13] C.-H. Kuo, S. Khamis, and V. Shet, "Person re-identification using semantic color names and RankBoost," in *Proc. IEEE Workshop Appl. Comput. Vis.*, Jan. 2013, pp. 281–287.

[14] Y. Deng, P. Luo, C. C. Loy, and X. Tang, "Pedestrian attribute recognition at far distance," in *Proc. ACM Conf. Multimedia*, Nov. 2014, pp. 789–792.

[15] T. Matsukawa and E. Suzuki, "Person re-identification using CNN features learned from combination of attributes," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Dec. 2016, pp. 2428–2433.

[16] Z. Yin *et al.*, "Adversarial attribute-image person re-identification," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, 2018, pp. 1100–1106.

[17] D. Gray, S. Brennan, and H. Tao, "Evaluating appearance models for recognition, reacquisition, and tracking," in *Proc. VS-PETS Workshop*, vol. 3, no. 5, 2007, pp. 1–7.

[18] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by GAN improve the person re-identification baseline *in Vitro*," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 3774–3782.

[19] Y. Sun, L. Zheng, W. Deng, and S. Wang, "SVDNet for pedestrian retrieval," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 3820–3828.

[20] S. Karanam *et al.*, "A systematic evaluation and benchmark for person re-identification: Features, metrics, and datasets," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 3, pp. 523–536, Mar. 2018.

[21] Z. Wang, X. Bai, M. Ye, and S. Satoh, "Incremental deep hidden attribute learning," in *Proc. ACM Conf. Multimedia*, Oct. 2018, pp. 72–80.

[22] L. Zheng, Y. Yang, and A. G. Hauptmann, "Person re-identification: Past, present and future," 2016, *arXiv:1610.02984*. [Online]. Available: https://arxiv.org/abs/1610.02984

[23] W. Li, R. Zhao, T. Xiao, and X. Wang, "DeepReID: Deep filter pairing neural network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 152–159.

[24] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng, "Person re-identification by multi-channel parts-based CNN with improved triplet loss function," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1335–1344.

[25] R. R. Varior, M. Haloi, and G. Wang, "Gated siamese convolutional neural network architecture for human re-identification," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 791–808.

[26] W. Chen, X. Chen, J. Zhang, and K. Huang, "Beyond triplet loss: A deep quadruplet network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 403–412.

[27] J. Wang, Z. Wang, C. Gao, N. Sang, and R. Huang, "DeepList: Learning deep features with adaptive listwise constraint for person reidentification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 3, pp. 513–524, Mar. 2016.

[28] J. Zhu, H. Zeng, S. Liao, Z. Lei, C. Cai, and L. Zheng, "Deep hybrid similarity learning for person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 11, pp. 3183–3193, Nov. 2018.

[29] X. Yang, M. Wang, and D. Tao, "Person re-identification with metric learning using privileged information," *IEEE Trans. Image Process.*, vol. 27, no. 2, pp. 791–805, Feb. 2018.

[30] D. Cheng, Y. Gong, X. Chang, W. Shi, A. G. Hauptmann, and N. Zheng, "Deep feature learning via structured graph laplacian embedding for person re-identification," *Pattern Recognit.*, vol. 82, pp. 94–104, Oct. 2018.

[31] Z. Zheng, L. Zheng, and Y. Yang, "A discriminatively learned CNN embedding for person reidentification," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 14, no. 1, p. 13, 2017.

[32] D. Li, X. Chen, Z. Zhang, and K. Huang, "Learning deep context-aware features over body and latent parts for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 384–393.

[33] W. Li, X. Zhu, and S. Gong, "Person re-identification by deep joint learning of multi-loss classification," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, Aug. 2017, pp. 2194–2200.

[34] D. Tao, Y. Guo, B. Yu, J. Pang, and Z. Yu, "Deep multi-view feature learning for person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 10, pp. 2657–2666, Oct. 2018.

[35] Z. Feng, J. Lai, and X. Xie, "Learning view-specific deep networks for person re-identification," *IEEE Trans. Image Process.*, vol. 27, no. 7, pp. 3472–3483, Jul. 2018.

[36] C. Su, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Deep attributes driven multi-camera person re-identification," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 475–491.

[37] C. Su *et al.*, "Attributes driven tracklet-to-tracklet person re-identification using latent prototypes space mapping," *Pattern Recognit.*, vol. 66, pp. 4–15, Jun. 2017.

[38] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.

[39] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[40] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[41] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *J. Mach. Learn. Res.*, vol. 11, no. 12, pp. 3371–3408, Dec. 2010.

[42] M. Abadi *et al.*, "TensorFlow: A system for large-scale machine learning," in *Proc. OSDI*, vol. 16. 2016, pp. 265–283.

[43] F. Chollet. (2017) *Keras*. [Online]. Available: http://keras.io

[44] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 17–35.

[45] X. Wang, G. Doretto, T. Sebastian, J. Rittscher, and P. Tu, "Shape and appearance context modeling," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2007, pp. 1–8.

[46] D. Chen, Z. Yuan, B. Chen, and N. Zheng, "Similarity learning with spatial constraints for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1268–1277.

[47] L. Zhang, T. Xiang, and S. Gong, "Learning a discriminative null space for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1239–1248.

[48] R. R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang, "A siamese long short-term memory architecture for human re-identification," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 135–153.

[49] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking person re-identification with k-reciprocal encoding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 3652–3661.

[50] S. Bai, X. Bai, and Q. Tian, "Scalable person re-identification on supervised smoothed manifold," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 3356–3365.

[51] A. Schumann and R. Stiefelhagen, "Person re-identification by deep learning attribute-complementary information," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jul. 2017, pp. 20–28.

[52] C. Song, Y. Huang, W. Ouyang, and L. Wang, "Mask-guided contrastive attention model for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1179–1188.

[53] Y. Wu, Y. Lin, X. Dong, Y. Yan, W. Bian, and Y. Yang, "Progressive learning for person re-identification with one example," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 2872–2881, Jun. 2019.

[54] J. Xu, R. Zhao, F. Zhu, H. Wang, and W. Ouyang, "Attention-aware compositional network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2119–2128.

[55] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2285–2294.

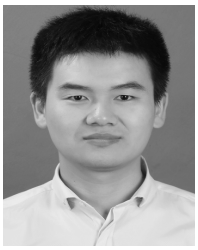
[56] M. Ye *et al.*, "Person reidentification via ranking aggregation of similarity pulling and dissimilarity pushing," *IEEE Trans. Multimedia*, vol. 18, no. 12, pp. 2553–2566, Dec. 2016.

[57] J. Wang, Z. Wang, C. Laing, C. Gao, and N. Sang, "Equidistance constrained metric learning for person re-identification," *Pattern Recognit.*, vol. 74, pp. 38–51, Feb. 2018.

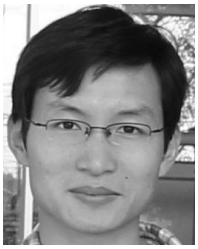
[58] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2921–2929.

[59] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.

Zheng Wang (M'19) received the B.S. and M.S. degrees from Wuhan University, Wuhan, China, in 2006 and 2008, respectively, and the Ph.D. degree from National Engineering Research Center for Multimedia Software, School of Computer Science, Wuhan University, in 2017. He is currently a JSPS Fellowship Researcher with the Shin'ichi Satoh's Lab, National Institute of Informatics, Japan. His research interests focus on person re-identification and instance search.



Junjun Jiang (M'15) received the B.S. degree from the Department of Mathematics, Huaqiao University, Quanzhou, China, in 2009, and the Ph.D. degree from the School of Computers, Wuhan University, Wuhan, China, in 2014. From 2015 to 2018, he was an Associate Professor with the School of Computer Science, China University of Geosciences, Wuhan. Since 2016, he has been a Project Researcher with the National Institute of Informatics, Tokyo, Japan. He is currently a Professor with the School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China. His research interests include image processing and computer vision.



Yang Wu (M'19) received a B.S. and the Ph.D. degrees from Xi'an Jiaotong University in 2004 and 2010, respectively. From 2011 to 2014, he was a Program Specific Researcher with the Academic Center for Computing and Media Studies, Kyoto University. He is currently a Program-Specific Senior Lecturer with the Department of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University. He is also a Guest Associate Professor with the Nara Institute of Science and Technology (NAIST), where he was an Assistant Professor with the NAIST International Collaborative Laboratory for Robotics Vision, from December 2014 to June 2019. His research interests are include in the fields of computer vision, pattern recognition, and image/video search and retrieval.

Mang Ye received the B.S. and M.S. degrees in electronic information from Wuhan University, Wuhan, China, in 2013 and 2016, and the Ph.D. degree in computer science from the Department of Computer Science, Hong Kong Baptist University, Hong Kong, in 2019. He is currently a Research Scientist with the Inception Institute of Artificial Intelligence, Abu Dhabi, United Arab Emirates. His research interests focusing on multimedia content analysis and retrieval, computer vision, and pattern recognition.

Xiang Bai (SM'16) received the B.S., M.S., and Ph.D. degrees in electronics and information engineering from the Huazhong University of Science and Technology (HUST), Wuhan, China, in 2003, 2005, and 2009, respectively. He is currently a Professor with the School of Electronic Information and Communications, HUST. He is also the Vice-Director of the National Center of Anti-Counterfeiting Technology, HUST. His current research interests include object recognition, shape analysis, scene text recognition, and intelligent systems.

Shin'ichi Satoh (M'04) received the B.E. degree in electronics engineering and the M.E. and Ph.D. degrees in information engineering from the University of Tokyo, Tokyo, Japan, in 1987, 1989, and 1992, respectively. He was a Visiting Scientist with the Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, USA, from 1995 to 1997. He has been a Full Professor with the National Institute of Informatics, Tokyo, Japan, since 2004. His current research interests include image processing, video content analysis, and multimedia databases.