IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received June 28, 2019, accepted July 6, 2019, date of publication July 9, 2019, date of current version July 26, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2927626

Deep Regression Neural Network for End-to-End Person Re-Identification

YINGCHUN GUO[1,2], (Member, IEEE), KUNPENG ZHAO[1], XIAOKE HAO[1,2], (Member, IEEE), AND MING YU[1,2], (Member, IEEE)

[1]School of Artificial Intelligence, Hebei University of Technology, Tianjin 300401, China
[2]State Key Laboratory of Reliability and Intelligence of Electrical Equipment, Hebei University of Technology, Tianjin 300401, China

Corresponding author: Xiaoke Hao (haoxiaoke@scse.hebut.edu.cn)

This work was supported in part by the Natural Science Foundation of China under Grant 61806071, in part by the Natural Science Foundation of Hebei Province of China under Grant F2019202381, in part by the Sci-tech Development Strategic Research Project of Tianjin under Grant 18ZLZXZF00660, and in part by the Innovation Ability Training Subsidy Project of Hebei Education Department for Postgraduates under Grant CXZZSS2019024.

ABSTRACT Person re-identification can be seen as a process of open set recognition. Usually, the deep learning models consider the person re-identification model as a classification model with a softmax layer. However, the softmax layer cannot be extended to unknown classes because of its closed nature, so the classification model is just regarded as the feature extractor. To overcome the problem mentioned above and make the person re-identification process end-to-end, this paper cast the person re-identification into a regression process and calculates the probability that persons in the images belong to the same identity. First, this paper proposes a deep regression model, named deep regression neural network integrating adaptive multi-attribute fusion method (DRNN-AMAF), which can make the person re-identification as regression analysis. Second, attributes are taken as the basis of this model for calculating the probability of persons belonging to the same identity, and each attribute corresponds to each branch of the deep regression neural network. Finally, hard labels of multiple attributes are adaptively fused into a soft label by the proposed multi-label fusion method based on the idea of Bayesian inference, which makes the attribute labels suitable for regression tasks. The comprehensive experiments on available public databases are conducted, and the experimental results show that our model produces competitive performance compared with the state-of-the-art approaches.

INDEX TERMS Person re-identification, adaptive multi-label fusion, deep regression neural network, probabilistic regression.

I. INTRODUCTION

The task of person re-identification is to discriminate whether the identities of the persons are the same in the images taken by different cameras with non-overlapping fields of view. Affected by factors, such as changes in lighting, pose, viewing distances, or occlusion, images of the same identity captured by different cameras or at different times by the same camera may have significant differences in visual appearance. Persons with different identities may have strong visual similarity due to similar color of clothes, physical features, or pose (see Fig.1).

The persons in the two images in Fig. 1(a) belong to different identities, but have strong visual similarities, while two images belonging to the same identity in Fig. 1(b) have many visual differences.

In a real-world deployment, the test identities of persons are usually not in the training sets, so the identity classes in the deployment are unknown classes for the re-identification model. At the same time, because of the images continuously captured by a surveillance camera, a tremendous amount of data is generated. In order to ensure the real-time accuracy of the system, the industry prefers to simple and effective models rather than models which concatenating lots of local features in the inference stage. Moreover, the simple and effective model of person re-identification system still needs to extract semantic features that are comprehensive and robust for the changes of lighting, pose, view angle, view distance, and occlusion, and efficiently determine whether these features belong to the

The associate editor coordinating the review of this manuscript and approving it for publication was Jingchang Huang.

(a) Persons with different identities, but similar appearances

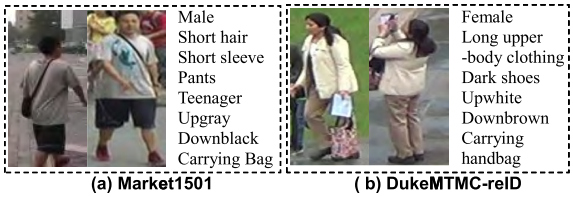
(b) Persons with the same identity, but discrepancies in appearances

FIGURE 1. Parts of samples in the Market1501 dataset.

same identity. All of these make person re-identification a challenging task.

The existing person re-identification methods are divided into two aspects: detecting discriminative identity features [1]–[25], designing a proper similarity measurement by using multi-information fusion method [26]–[30].

With the development of deep learning, more and more person re-identification methods use deep model to detect discriminative features and thus achieve better performance than traditional methods by using hand-crafted feature descriptor. Usually, these methods consider the person re-identification model as (1) verification model structures, which determine whether persons in two input images have the same identity [1]–[7], (2) identification model structures, which recognize the person identity in the input image [8]–[16], or (3) the combination of a verification model and an identification model [17]–[19]. However, there are some problems for the deep classification model. For example, the identification model needs to clarify how many identity classes are there in the real-world deployment in advance of the training process. This kind of model has inadaptability [31]. In the real-world deployment, the system is in an open environment, and it is possible that the test images of the person were not encountered in the training process. Take an example, if there are person A and person B in the training process, and there are person A, B, and C in the actual application process. Since the softmax layer in deep learning models for person re-identification cannot be extended to unknown classes because of its closed nature. In this condition, the softmax layer has two neurons in the training process, and there is no corresponding neuron of class C in practice. In addition, person re-identification models are treated as the verification model, and the outputs of the models are hard labels. If a query image is verified to have the same identity as diverse gallery images belonging to different identities, a distance metric is still needed to recognize the query identity. Thus, the deep classification models are just regarded as the feature extractors and the process of person re-identification then needs to use distance metric for the measure of similarity. These two-step methods can't make person re-identification become a real end-to-end process. And also, the cross-entropy



(a) Market1501 (b) DukeMTMC-reID

FIGURE 2. Person visual attributes representation in Market-1501 and DukeMTMC-reID.

of the softmax function is used as the loss function to train the model during the training process, and the distance of the feature maps is used as the measure during the application process. So the learned knowledge of this kind of model is not fully utilized. For example, a single softmax prediction may correspond to multiple different logit (i.e., feature) inputs. Specifically, even if two logit inputs are different, as long as the corresponding element is relatively larger than others, both of their softmax outputs will be close to the same one-hot vector. In other words, for deep classification models, the underlying feature representations of each class are not unique [31].

In order to overcome the inadaptability of the classification model which makes person re-identification cannot be in a real end-to-end process, and to obtain an identity similarity rank end-to-end which makes the model suitable for real-world deployment, we propose a person re-identification method based on deep regression, which aims to calculate the probability of the same identity of persons in the image pair end-to-end. Unlike the outputs of the classification model are hard labels, the output of this model is a soft label. By combining the query image with the images in the gallery to form the image pairs as the inputs of the model, the rank with the probabilities of the same identity can be obtained to determine the identity of the query image.

Since the information of a single image is limited, multi-information fusion methods [26]–[30] have been proposed and are intended to utilize combined information, such as image information, motion information between video frames, spatio-temporal information, multi-label information, etc., to improve the accuracy of the matching model.

In the process of re-identification by human experts, visual attributes are checked one by one until getting the likelihood of two persons in an image pair belonging to the same identity. For example, human experts first analyze whether persons in the two images are the same gender, and then consider whether their apparels are the same or not, and so on. Through an item-by-item investigation, the likelihood that the persons belonging to the same identity is obtained. The examples of person attribute labels are shown in Fig.2 marked by Lin [32].

In practice, each visual attribute has a different level of importance. For example, due to the immutability of gender, two persons with different genders cannot be the same identity. Therefore, greater confidence is given to a gender-based decision. As for a hat or a backpack, etc., less confidence is

(a)backpack for person-Ⅰ (b) hat for person-Ⅱ

FIGURE 3. The backpack and hat effect on person Re-ID.

given to decisions made by these attributes because of visual changeability. The backpack or hat is likely to have a negative impact on the single-shot person re-identification task due to changes in viewing angle and occlusion, shown in Fig.3.

For person re-identification, it is crucial to apply attributes appropriately. Inspired by the decision process of experts and the idea of Bayesian inference, we propose a Bayesian inference based multi-attribute fusion and optimization method. The main ideas are as follows: (1) hard labels of multiple attributes are fused as a soft label to make supervisory information more suitable for regression tasks; (2) making the representation feature learned by the model powerful for each attribute; (3) Bayesian inference based multi-label fusion and optimization theory is integrated into the end-to-end iterative process of the network which also makes the model learn the adaptive fusion weight of each attribute decision and obtain the accurate regression probability.

To demonstrate that the proposed approach can improve the accuracy of person re-identification (see comparison experiment IV-C baseline2 for details), make the feature more discriminative (see comparison experiment IV-C baseline1 for details) and make the deep regression model easy to converge (see comparison experiment IV-C baseline2 for details), we applied datasets Market1501 and DukeMTMC-reID for evaluation.

Our contributions can be summarized as follows:

(1)Take the person re-identification model as a regression model. In this way, we have solved the problem that person re-identification process cannot be end-to-end due to the inadaptability of the classification model and made the proposed DRNN-AMAF model more suitable for real-world deployment.

(2)In order to solve the problem that the deep regression models are difficult to converge, the Siamese network, multi-branch structure, and multi-branch fusion structure are the first time employed in regression setting for person re-identification.

(3)The multi-attribute labels fusion and optimization theory are proposed by fusing multiple attribute hard labels as a soft label to make supervisory information more suitable for regression tasks and integrated into the end-to-end iterative process of the network to make the deep regression model easy to converge.

II. RELATED WORK

In this section, we introduce the existing person re-identification work from three parts: methods of learning discriminative feature representation, multi-information fusion methods, and several ''look-alike'' works.

A. METHODS OF FEATURE REPRESENTATION

Inspired by the effective performance of deep learning in the field of computer vision, many researchers begin to apply deep features in person re-identification tasks. Most of the early person re-identification models based on deep learning tried to propose new network structures to improve model performance [1]–[3]. As far as we know, the earliest application of deep learning in the field of person re-identification are deep re-id proposed by Li [1], and deep metric learning for person re-identification proposed by Yi [2]. Li *et al.* [1] treated the person re-identification task as a verification problem and proposed a filter pairing neural network (FPNN), in which patch matching layer and convolution and max-pooling layer are added to make the feature robust to pose changes, and maxout-grouping layer is added to make the features robust to lighting changes. Yi *et al.* [2] divided the person image into three overlapped parts: top, middle, and bottom. The higher-order semantic features of the three parts are detected by two pooling layers and two convolution layers. The features of these three parts are fused by a fully connected layer. Literature [3] further improved the network structure and proposed cross-input neighborhood differences layer, patch summary layer, across-Patch layer, higher-order relationships layer to improve the performance of the model. Recent person re-identification researches are usually based on is still not converging network, such as VGG [33], GoogleNet [34], [35], ResNet [36], DenseNet [37], etc., and improves the recognition rates by providing a new theory. Chen *et al.* [11] comprehensively considered the global context information of small-scale images and the local information of large-scale images and proposed Deep Pyramidal Feature Learning (DPFL) CNN, which includes multi-scale inputs and the corresponding network branch to each scale. It makes the feature more discriminative. In literature [15], Multi-Scale Context-Aware Network (MSCAN) is proposed by using multi-scale information, which can capture the local context information and integrate global features over the whole body and different body parts by stacking multi-scale filters on each layer. As for using a multi-model structure to obtain the robust features, literature [18] combined the loss functions of the verification model and the identification model to learn more discriminative features. In order to solve the problem of the insufficient data sample and domain adaptation in a single dataset, Xiao [10] proposed a Domain Guided Dropout (DGD) algorithm to learn generalized features on multiple datasets.

Metric learning is another commonly used method in person re-identification tasks. A good metric function from feature space to distance space is generally obtained by organizing pairs of constraints—positive constraints and negative constraints—which make the distance of positive image pairs closer and the distance of negative image pairs further apart

in the feature space [22]–[25], [38]. Deep metric learning methods aim to learn the similarity of image pairs through the network, which makes features with the same identity more similar than features belong to different identities by using triplet loss [20] and quadruplet loss [21]. These methods undoubtedly make the features of person re-identification more discriminative.

B. MULTI-INFORMATION FUSION METHODS

Due to the limited information contained in a single image, many researchers have proposed a comprehensive application of multiple information to improve the performance of the model. Using the combination of image feature information and spatio-temporal information, Lv [29] proposed an unsupervised transformative learning theory. The fusion model is composed of spatio-temporal model constructed by spatio-temporal information of data label based on Bayesian inference and deep image model. The labels of each image pair are generated by the fusion model and the learning to rank method, and the deep image model is further trained with the labels and data, thereby improving the whole performance of the model.

Person re-identification is a cross-view image matching task, and the view-specific bias has a great influence on the matching accuracy. The use of sequence-level image feature information and motion information between frames of a video sequence can suppress the negative effects of view-specific bias. Zhang [30] applied reinforcement learning theory to person re-identification for the first time and considered the decisions based on the feature sequence composed of each frame as Markov Decision Processes. In Zhang's model, an agent is trained to aggregate sequence-level image features and furthermore, persons in two images are determined whether they belong to the same identity or not. Literature [27] integrates the CNN model, the RNN model and the temporal pooing layer in Siamese network architecture. The features of each frame in a video sequence are extracted by the CNN model, then RNN and temporal pooling are used to aggregate these features. This model effectively utilizes the image information of each frame and the motion information between frames to improve the performance of the person re-identification model.

C. SEVERAL "LOOK-ALIKE" WORKS

We propose a multi-attribute adaptive fusion and optimization algorithm to make the attribute labels more adaptive to the regression model and to make the deep regression model learn powerful features of attributes. Then, the probability of the same personal identity is regressed with the adaptive fusion of the multi-attribute feature. Feature extraction and probability regression are embedded in the end-to-end process. Some existing methods, such as [31] and [39], seem similar to the proposed method but are fundamentally different.

Wang *et al.* [31] proposed the identity regression space, in which all of the persons in the images belonging to the same identity are represented by one point in the space, and the ridge regression is used to find the optimal solution for the embedding function of the person images. The difference with Wang's method is that the proposed method solves the probability that the persons in the input image pair belong to the same person identity by the deep regression model in an end-to-end process.

Su *et al.* [39] proposed a weakly-supervised multi-type attribute learning framework, which is a multi-branch classification based model taking attributes as supervisory information. This model not only predicts the identity of the person but also predicts the person attributes, which makes the model learn more discriminative and generalized features. However, this method still aims to find better feature representation, and it is still a two-step process while not dealing with the person re-identification in an end-to-end process. Our framework is based on the regression model and considers the relationship among multiple attributes and get the re-identification results in an end-to-end process.

III. OUR METHOD

In this section, we will start with a formal formulation of the person re-identification problem. Then, in part B, we present each component of DRNN-AMAF, the motivation of each component, and the Bayesian inference based multi-attribute fusion and optimization theory embedded in the end-to-end process of deep learning. Bayesian inference based multi-attribute fusion and optimization theory is presented in part C. Part D introduces the training method proposed in this paper to solve the problem that the regression model is difficult to converge. The framework of the proposed adaptive multi-attribute fusion optimization for DRNN-AMAF is shown in Fig.4. In Fig.4 (a), the DRNN-AMAF network consists of two major components: a shared weight Siamese network using as the feature extraction, and a multi-branch network structure to regress the similarity of various attributes individually. The scores of various attributes are fused as the probability of persons in the input pair belonging to the same identity. In Fig.4 (b), Bayesian inference based multi-attribute fusion and optimization theory is used to fuse multiple attribute hard labels as a soft label to make supervisory information more suitable for regression tasks. " $\otimes$ " denotes element-wise multiplication, " $\oplus$ " denotes element-wise addition, and " $\ominus$ " denotes element-wise subtraction in our framework setting.

A. PROBLEM DEFINITION

Suppose that the types of attributes in the dataset are denoted as $\{\boldsymbol{a_1}, \ldots, \boldsymbol{a_j}, \ldots, \boldsymbol{a_m}\}$, where m represents number of attribute types including the identity attribute. Let $\boldsymbol{a_j}$ represent a type of attribute, for example, $\boldsymbol{a_j}$ represents gender and $\boldsymbol{a_j} = \{0$: *male*,1: *female*$\}$. $\boldsymbol{a_j} = 0$ represents male, and $\boldsymbol{a_j} = 1$ represents female. We set $\mathbf{y}_i(\delta_{i1}, \ldots, \delta_{ij}, \ldots, \delta_{im})^{\mathrm{T}}$ s.t.: $\delta_{ij} \in \boldsymbol{a_j}$ as the attribute labels of the i-th training samples x_i, $i = 1, \ldots, n$. The attribute labels of training samples can be expressed as $\mathbf{Y}=\{\mathbf{y}_1, \ldots, \mathbf{y}_i, \ldots, y_n\} \in \mathbb{R}^{m \times n}$, where n is

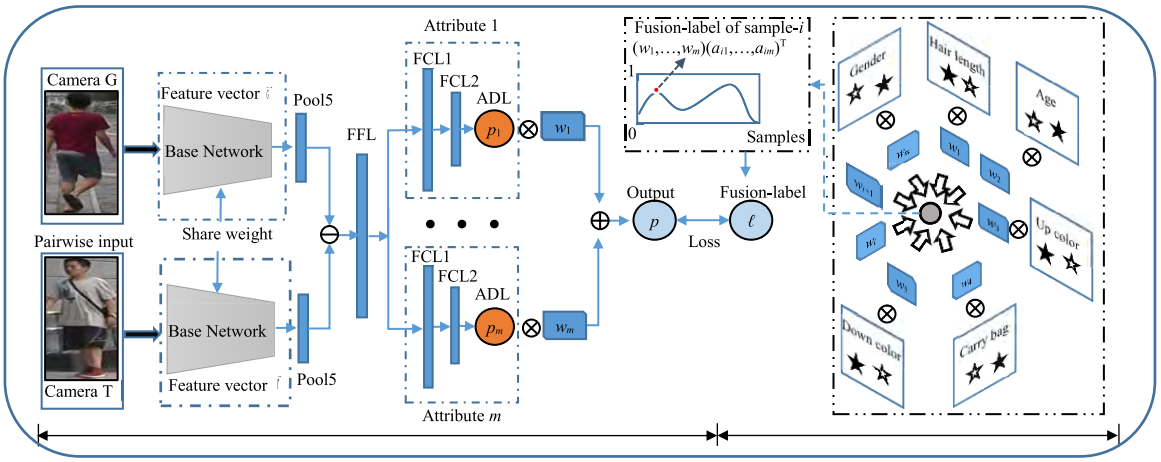


FIGURE 4. Overview of the proposed adaptive multi-attribute fusion optimization framework for DRNN-AMAF.

number of samples in training dataset. So the training set can be represented as $\mathbf{T}=\{\mathbf{X},\mathbf{Y}\}$, and $\mathbf{X}=\{x_i, i = 1, \ldots, n\}$.

For image pairs from two cameras $\theta = \{(x_h, x_q)\}$, $h, q = 1, \ldots, n$, $h \neq q$ in training set, we wish to learn a mapping function from image pairs to the probability distribution F which persons in the image pair belong to the same identity. The mapping function $\mathbf{O}$ can be resolved by the proposed DRNN-AMAF model, as shown in Equ. (1).

$$\mathbb{F} \sim \begin{bmatrix} \theta_1 \cdots \theta_i \cdots \\ p_1 \cdots p_i \cdots \end{bmatrix}, p_i = \mathbf{O}(\theta_i) \tag{1}$$

where p_i is the probability of the persons, in the i-th image pair θ_i, belong to the same identity. The attribute label vector of image pairs (x_h, x_q) are shown in Equ. (2).

$$\boldsymbol{\ell}_i = \left(\ell_1, \ldots, \ell_j \ldots, \ell_m\right)^T s.t. : \ell_j = \begin{cases} 1, & if\ \mathbf{y}_{hj} = \mathbf{y}_{qj} \\ 0, & otherwise \end{cases} \tag{2}$$

where y_{hj} and y_{qj} are the j-th element in attribute label vectors of the images x_h and x_q.

B. FRAMEWORK OF MULTI-ATTRIBUTE FUSION PERSON IDENTITY SIMILARITY REGRESSION NETWORK

As shown in Fig. 4(a), we aim to get the probability that persons belonging to the same identity. So we prefer that the feature extraction part can get the difference between the image pair. Siamese network proposed in [40] is used to measure the similarity between two inputs. Thus a shared weight Siamese network is used to extract features. The feature vectors extracted from the input image pair are combined by element-wise subtraction to calculate the visual difference between the input image pair.

It is difficult for the model to converge by regressing the probability directly. However, this problem can be resolved by refining the probability of each attribute, which is the same individually. Therefore, a multi-branch structure is applied to regress the similarity of various attributes individually. Finally, the scores generated by each attribute branch are weighted fused, which are shared with the multi-attribute label fusion task (i.e., Equ.(5)).

The attribute fusion weight vector $\mathbf{w}$ is shown in Equ.(3).

$$\mathbf{w} = \left(w_1, \ldots, w_j \ldots, w_m\right)^T, s.t. : \sum_{j=1}^{m} w_j = 1, w_j \in [0, 1] \tag{3}$$

The mapping function $\mathbf{O}$ can be decomposed into the inner product of two vectors as Equ.(4).

$$\mathbf{O} = \mathbf{w}^T \cdot \left(o_1, \ldots, o_j, \ldots, o_m\right)^T \tag{4}$$

where o_j represents the jth attribute branch in the mapping function and "$\cdot$" means inner product.

For the ith image pair θ_i, the probabilities of having the same attributes can further be represented as a vector $(p_{i1}, \ldots, p_{ij}, \ldots, p_{im})^{\mathrm{T}}$, $p_{ij} \in [0,1]$, where $p_{ij} = o_j(\theta_i)$. Thus the probability of the ith image pair belonging to the same identity can be represented as Equ.(5).

$$\begin{aligned} p_i &= \mathbf{O}(\theta_i) = \mathbf{w}^T \cdot \left(o_1(\theta_i), \ldots, o_j(\theta_i), \ldots, o_m(\theta_i)\right)^T \\ &= \mathbf{w}^T \cdot \left(p_{i1}, \ldots, p_{ij}, \ldots, p_{im}\right)^T \end{aligned} \tag{5}$$

Therefore, the person re-identification task is embodied in such a way that the model learns both the probabilities that each attribute of the image pair is the same and the fusion weight of the probabilities of each attribute.

The regression network structure of the proposed model is given in Table 1.

In Table 1, m is the number of attributes.

As shown in Fig. 4(b), the Bayesian inference based multi-attribute fusion and optimization theory, proposed in the next section, is embedded into the end-to-end training process of DRNN-AMAF to generate soft labels which can adapt to regression tasks at the beginning of backpropagation.

TABLE 1. Regression network structure.

Name	Output size
Base Network	2048*2
Feature fusion Layer(FFL)	2048*1
Full Connected Layer(FCL1)	1024*m
Full Connected Layer (FCL2)	512*m
Attribute Decision Fusion (ADL)	1*m
Output	1

In Table 1, m is the number of attributes.

C. BAYESIAN INFERENCE BASED MULTI-ATTRIBUTE FUSION AND OPTIMIZATION THEORY

Existing person re-identification models generally use the hard label given by the dataset as the supervision information. A hard label is a label assigned to a member of a class where membership is binary: either the element is a member of the class or not. A soft label is one which has a score (probability or likelihood) attached to it. The soft label is more suitable for regression analysis than the hard label. As represented in the last section, the person re-identification task can be regarded as multi-attribute similarity probability analysis. This motivates us to propose the multi-attribution fusion method, which obtains a soft label by fusing multi-hard-labels with different weights, to transform existing hard attribute labels into soft labels which adapt to regression tasks.

Here reference to Bayesian inference method, the optimal fusion weights can be obtained by integrated Bayesian reference in the end-to-end process of deep learning to generate the soft label. During the process of the end-to-end deep neural network, the fusion weights of the previous iteration are used as the monitoring data to keep iterating until the fusion weight can be obtained with a small enough loss function.

The multi-attribute fusion and the iterative optimization process are described as follows:

***Step* (1):** Before the iteration, the attribute fusion weight is randomly initialized as $\mathbf{w}_0$, and $\Pr(\mathbf{w}_0)$ is the prior probability distribution of $\mathbf{w}_0$. The prior probability distribution of the initialized mapping function $\mathbf{O}_0$ is $\Pr(\mathbf{O}_0)$, and $[\theta_b]_i$ represents the i-th mini-batch with b samples, and $[\ \ell_b]_i$ represents the corresponding attribute labels.

***Step*(2):** According to the Bayesian rule, the posterior probability of the attribute fusion weight vector is obtained as in Equ.(6).

$$\begin{aligned}\Pr(\mathbf{w}_1) &= \Pr\left(\mathbf{w}_0 \left| [\theta_b]_1, \mathbf{w}_0^T \cdot [\boldsymbol{\ell}_b]_1 \right.\right) \\ &= \frac{\Pr\left([\theta_b]_1, \mathbf{w}_0^T \cdot [\boldsymbol{\ell}_b]_1 | \mathbf{w}_0\right) \Pr(\mathbf{w}_0)}{\Pr\left([\theta_b]_1, \mathbf{w}_0^T \cdot [\boldsymbol{\ell}_b]_1\right)} \end{aligned} \quad (6)$$

***Step* (3):** The posterior probability of the mapping function is obtained as in Equ.(7).

$$\begin{aligned}\Pr(O_1) &= \Pr\left(O_0 \left| [\theta_b]_1, \mathbf{w}_0^T \cdot [\boldsymbol{\ell}_b]_1 \right.\right) \\ &= \frac{\Pr\left([\theta_b]_1, \mathbf{w}_0^T \cdot [\boldsymbol{\ell}_b]_1 | O_0\right) \Pr(O_0)}{\Pr\left([\theta_b]_1, \mathbf{w}_0^T \cdot [\boldsymbol{\ell}_b]_1\right)} \end{aligned} \quad (7)$$

After n times iterations by repeating step(2)∼(3), the learned mapping function becomes $\mathbf{O}_n$ and the probability distribution of all image pairs produced by the mapping function $\mathbf{O}_n$ is $\mathbb{F}_n$. The attributes fusion weight vector is $\mathbf{w}_n$.

We use $\mathbf{O}_n$ and $\mathbf{w}_n$ as the prior knowledge for the $(n+1)$th iteration. The output mini-batch $[\theta_b]_{n+1}$ of the model is $\{p_1, \ldots, p_b\} \sim \mathbb{F}_n$. By using the fusion label $\mathbf{w}_n^T \cdot [\boldsymbol{\ell}_b]_{n+1} = \{k_1, \ldots, k_b\}$ as the supervision information for training and applying Equ. (6) and Equ. (7) to the training process, we can get Equ. (8) and Equ. (9):

$$\begin{aligned}\Pr(\mathbf{w}_{n+1}) &= \Pr\left(\mathbf{w}_n \left| [\theta_b]_{n+1}, \mathbf{w}_n^T \cdot [\boldsymbol{\ell}_b]_{n+1} \right.\right) \\ &= \frac{\Pr\left([\theta_b]_{n+1}, \mathbf{w}_n^T \cdot [\boldsymbol{\ell}_b]_{n+1} | \mathbf{w}_n\right) \Pr(\mathbf{w}_n)}{\Pr\left([\theta_b]_{n+1}, \mathbf{w}_n^T \cdot [\boldsymbol{\ell}_b]_{n+1}\right)} \end{aligned} \quad (8)$$

$$\begin{aligned}\Pr(\mathbf{O}_{n+1}) &= \Pr\left(\mathbf{O}_n \left| [\theta_b]_{n+1}, \mathbf{w}_n^T \cdot [\boldsymbol{\ell}_b]_{n+1} \right.\right) \\ &= \frac{\Pr\left([\theta_b]_{n+1}, \mathbf{w}_n^T \cdot [\boldsymbol{\ell}_b]_{n+1} | \mathbf{O}_n\right) \Pr(\mathbf{O}_n)}{\Pr\left([\theta_b]_{n+1}, \mathbf{w}_n^T \cdot [\boldsymbol{\ell}_b]_{n+1}\right)} \end{aligned} \quad (9)$$

The optimization iteration process of the model is embodied in the above Equations (1)∼(9).

As for the loss function, the cross-entropy is used to calculate the difference between the output probability distribution and the ground-truth probability distribution. It is described as Equ. (10).

$$L = -\frac{1}{b}\sum_{i=1}^{b}\left[k_i \cdot \log p_i + (1-k_i) \cdot \log(1-p_i)\right] \quad (10)$$

In order to avoid the multi-attribute fusion weight vector to be the one-hot tendency, the L2 regularization term is added to the loss function. The final loss function is shown as follows:

$$L = -\frac{1}{b}\sum_{i=1}^{b}\left[k_i \cdot \log p_i + (1-k_i) \cdot \log(1-p_i)\right] + \lambda\sqrt{\sum_{j=1}^{m} w_j^2} \quad (11)$$

where λ is the regularization coefficient.

Usually, the loss function of the regression model is a squared loss function, as follows:

$$L = \frac{1}{b}\sum_{i=1}^{b}(p_i - k_i)^2 \quad (12)$$

In Section IV-D, we compare the performances of these two loss functions.

D. TRAINING METHOD

In order to make the deep regression model more easy to converge, we proposed a training method, named adaptive multi-attribute fusion optimization method as follows and summarized in **Algorithm 1**.

In **Algorithm 1**, ResNet-50 [36] pre-trained on the image classification dataset ImageNet [41] is used as the base network. First, ResNet-50 is fine-tuned on the training set of the person re-identification dataset, to obtain the classification model for pedestrian re-identification (CR-ID, for short). Then the softmax layer of the fine-tuned network is removed, and the parameters of the remaining layers are retained as a base network to construct the multi-attribute fusion regression network(DRNN-AMAF). Finally, the base network is kept in a low learning rate($lr \searrow$), and the regression layers are kept in a relatively higher learning rate($\boldsymbol{lr} \nearrow$) training on the training set for several epochs.

Algorithm 1 Adaptive Multi-Attribute Fusion Optimization

Input: Pre-trained model $\phi(O_o)$, Re-ID training data $\mathbf{X}$, Attribute labels $\mathbf{Y}$, Maximum Iterations T_c, T_r for CR-ID, and DRNN-AMAF, respectively.
Output: Learned DRNN-AMAF model $\phi(\hat{O}_b, \hat{O}_w, \hat{O}_r)$, where $\hat{O}_b$, $\hat{O}_w$ and $\hat{O}_r$ denote parameters of BN, fusion layers and regression layers, respectively.
Initialization: $O_b \leftarrow O_o$, random initialization for $\hat{O}_w$ and $\hat{O}_r$
Fine-tuning on person re-id benchmark datasets
for $t = 1 : T_c$ **do**
  Keep O_b $\boldsymbol{lr} \searrow$ and softmax layer in CR-ID $\boldsymbol{lr} \nearrow$
  Update $CR - ID^t$ using cross-entropy loss function
end for
$O_b \leftarrow O_b^{T_c}$
DRNN-AMAF learning
for $t = 1 : Tr/2$ **do**
  Keep O_b fixed
  Fuse hard labels $\mathbf{Y}$ into a soft label using inner product of Equ.(2) and Equ.(3)
  Update O_r^t and O_w^t using Equ.(11) by the theory Equ.(6) to Equ.(9)
end for
$O_r \leftarrow O_r^{T_r/2}$,$O_w \leftarrow O_w^{T_r/2}$
for $t = Tr/2 : Tr$ **do**
  Keep O_b $\boldsymbol{lr} \searrow$, O_w and O_r $\boldsymbol{lr} \nearrow$
  Fuse hard labels $\mathbf{Y}$ into a soft label using inner product of Equ.(2) and Equ.(3)
  Update O_b^t, O_r^t and O_w^t using Equ.(11) by the theory Equ.(6) to Equ.(9)
end for
$\hat{O}_b \leftarrow O_b^{T_r}$,$\hat{O}_r \leftarrow O_r^{T_r}$,$\hat{O}_w \leftarrow O_w^{T_r}$
Return: $\phi(\hat{O}_b, \hat{O}_w, \hat{O}_r)$

The former two steps of the training method train the person re-identification model as a classification model, which greatly speeds up the convergence of the feature extraction layers. The third step training is based on the knowledge of feature extraction layers learned by the classification model, and the soft label is used as supervised information to train the final regression model. Note that, the soft label, which is adaptively fusing of multiple attributes hard labels by DRNN-AMAF model, is the supervisory information that more suitable for regression tasks. These all play a direct role in the convergence of the regression model. By comparing Baseline2 with our proposed method, we demonstrate the positive impact of our training method on the accuracy of the regression model and making the deep regression model easy to converge.

IV. EXPERIMENTS

The steps of person re-identification are more similar to image retrieval tasks than to classification tasks. Images with the same identity taken as the query image are retrieved in the gallery images. We tested the proposed model at Market-1501 and DukeMTMC-reID datasets and achieved high accuracy.

A. DATASETS AND SETTINGS

1) DATASETS

Based on the assumption in [42] that pedestrian images are captured in a short period, so clothes and shape of the body do not change much, and can be used as cues to recognize the identity. If the persons in the image pair have the same identity, all of the attribute labels must be the same. If the person identities are different, parts of the attribute labels maybe the same. Here for each type of attributes in image pair, the value 0 means different, and the value 1 means the same, shown in Equ. (2).

2) MARKET-1501

[43] is a multi-shot pedestrian re-identification dataset which includes more than 32,000 person images captured by 6 cameras on the university campus. These images belong to 1501 identities, and each of which contains multiple images of different view angles or poses. This dataset is divided into a training set and a test set. The test set is composed of a query set and a gallery set. The training set includes 12,936 images of 751 identities. The test set has 750 identities, including 3,368 images for the query and 19,732 images for the gallery. Lin *et al.* [32] labeled 27 types of attribute labels for this dataset. Moreover, for each identity, the upper-body clothing and the lower-body clothing only have one color individually. In order to reduce the number of network branches and the overall parameter amount, the labeled 8 colors of upper-body clothing and the labeled 9 colors of lower-body clothing are taken as one attribute individually, and the original 27 attributes and the ID label are summarized into 13 attributes, see Table 2.

3) DUKEMTMC-reID

[45] dataset includes 36,361 images of 1,404 identities captured by 8 cameras. The gallery dataset contains 17,661 images of 1,110 identities. Query dataset contains

TABLE 2. Reorganized attributes of market1501.

Attribute	Label
Gender	Male/female
Hair length	Short/long
Sleeve length	Short/long
Length of lower-body clothing	Short/long
Type of lower-body clothing	Dress/pants
Wearing hat	No/yes
Carrying backpack	No/yes
Carrying bag	No/yes
Carrying handbag	No/yes
Age	Young/teenager/adult/old
Color of lower-body clothing	Downblack/downwhite/downpink/ downpurple/downyellow/downgray/ downblue/downgreen/downbrown
Color of upper-body clothing	Upblack/upwhite/upred/uppurple/ upyellow/upgray/upblue/upgreen
ID	Identities of person

2,228 images of 702 identities. The training set has 702 identities of 16,522 images. Lin *et al.* [32] labeled 23 kinds of attribute labels for this dataset. Referring to the attribute organization method of the Market-1501 data set, we summarized these 23 types of attributes and the ID label into 11 types of attributes. Our intuition is that ID, gender and age are global information, and local information includes hair length, sleeve length, length of lower-body clothing, type of lower-body clothing, wearing a hat, carrying a backpack, carrying a bag, or carrying a handbag, colors of upper-body clothing, and colors of lower-body clothing, and so on.

4) IMPLEMENTATION SETTING DETAILS

We use the adaptive multi-attribute fusion optimization method to train the DRNN-AMAF. The size of the mini-batch is set to 32, and the initial learning rate of the fusion weight is 0.01, and the initial learning rate of the Base Network parameter is set to 0.001. The initial learning rate of the regression subnetwork parameters is 0.1, the weight decay is 5E-4, and the momentum is set to 0.9. There are 13,000 pairs of images in every epoch. Data preprocessing is performed on the input images, and the images are randomly flipped horizontally and erased [46] during the training process to improve data diversity. Therefore, the robustness of the final trained model to the occlusion and pose is improved. For the training data, we control the ratio of the positive and negative pairs to 1:1.

B. EVALUATION

In this section, we evaluate the performance of our model by applying three evaluation protocols for comparison including CMC curve, Top-1 prediction accuracy and mean average precision (mAP).

1) CMC CURVE

In order to fit the non-overlapping camera views of the person re-identification task, we randomly take one image from each gallery identity as a sample, which is captured by cameras different from the query image, so that we get the single-shot gallery set. First, for each image in a single-shot gallery set, the 2048-dims feature vector is extracted from the feature vector layer, and the Euclidean distance between the query image feature vector and the gallery image feature vector is calculated. Then the distance list is re-ranked according to the Euclidean distance from near to far. Finally, we use the Rank-1 CMC to evaluate the pros and cons of the model for feature extraction.

2) TOP-1 PREDICTION ACCURACY

The single-shot gallery is also used to evaluate the Top-1 prediction accuracy. The given query is combined with images in the gallery set to form image pairs, and the trained model is used to calculate the probability that the image pair has the same identity. The rank is discharged according to the magnitude of the probability. We use the Top-1 accuracy to measure the discriminative power of the model.

3) MEAN AVERAGE PRECISION (mAP)

mAP is a common method to evaluate the model performance on multi-shot datasets.

In the experiments, we listed the differences in experimental results with different attributes on Marker1501 and. DukeMTMC-reID. In the proposed network structure, each attribute corresponds to a network branch. In order to taking into account the accuracy and the number of parameters, not all of attributes are used in the experiments and we only adopt the attributes that have a great impact on the accuracy of the model(shown in Table 3 and Table 4). Identity information is strong supervisory information for pedestrian re-identification, so the weakly supervised learning setting in this paper is the training process without applying identity information. And we compare the probability regression method by using Equ.(11) as the loss function with the numerical regression method using Equ.(12) as the loss function. The differences between the two methods are shown in Table 3.

Model performances using different attribute combinations on DukeMTMC-reID are shown in Table 4.

Note that in TABLE 3 and TABLE 4, compared with the feature representation, the prediction accuracy is mainly affected in the case of weak supervised learning. It is mainly due to the fact that in the first step of the training strategy, we transfer representations learned from large image classification dataset ImageNet.

We test the single-shot Rank-1 CMC and mAP on the Market-1501. The experimental results of our method and the results of state-of-the-art methods are shown in Table 5.

The experimental results of our method and the results of state-of-the-art methods on DukeMTMC-reID are shown in Table 6.

TABLE 3. Probabilistic regression and numerical regression on marke1501(%).

	Attributes	Probability Regression		Numerical Regression	
		Top-1	Rank-1	Top-1	Rank-1
Fully-supervised learning	Baseline1 (single shot)	-	76.93	-	-
	Baseline2 (single shot)	13.20	-	-	-
	①+②	69.47	70.50	53.93	60.27
	①+②+③+④	56.67	62.13	45.33	57.60
	①+②+③+④+⑤+⑥+⑦	85.73	88.80	68.00	74.13
Weakly-supervised learning	②+③+④	45.60	60.27	-	-
	②+③+④+⑤+⑥+⑦	66.80	80.27	-	-

①ID ②gender ③downcolor ④upcolor ⑤age ⑥sleeve length ⑦length of lower-body clothing

TABLE 4. Model performance by using different attributes on DukeMTMC-reID(%).

	Attributes	Top-1	Rank-1
Fully-supervised learning	Baseline1(single shot)	-	71.65
	Baseline2(single shot)	15.38	-
	①+②	46.15	53.85
	①+②+③+④	66.38	70.23
	①+②+③+④+⑤	76.78	79.06
	①+②+③+④+⑤+⑥+⑦	68.95	70.52
Weakly-supervised learning	②+③+④+⑤	42.74	61.97
	②+③+④+⑤+⑥+⑦	59.97	70.37

①ID ②gender ③downcolor ④upcolor ⑤length of upper-body clothing ⑥shoes ⑦hat

Notice that in Table 5 and Table 6, we do not compare with part-based methods [14], [55] which exploit part or patch matching-based architecture to learn discriminative feature representation in local regions of persons. Although the performance of the part-based methods is generally better than the global-based methods, in order to straightforward express the positive impact of our method, we have chosen a simple global-based model as the base network.

Through the performance shown in the tables above, we can find that our method has obvious advantages in the accuracy of prediction (Top-1 and mAP). Most of the methods, using the deep classification model as the feature extractor, usually use SVM, KNN or the deep classification model fine-tuned on the test set as the classifier, but our method does not need these steps, and it can obtain more accurate identity prediction end-to-end.

C. ABLATION STUDY

We conduct an ablation study of the multi-branch structure and Bayesian inference based multi-attribute fusion theory to illustrate their positive impact on both feature representation and efficient training.

TABLE 5. Performance on the market1501(%).

	Methods	Rank-1	mAP	Methods	Rank-1	mAP
Fully-supervised learning	BoW+KISSME[43]	44.42	20.76	Baseline1 (single shot)	18.27	-
	MultiRegion[47]	66.36	41.17	Baseline2 (single shot)	76.93	-
	MSCAN[15]	86.79	66.70	APR[32]	84.29	64.67
	S-CNN[4]	65.88	39.55	MTA[48]	84.14	64.07
	GAN[9]	84.29	64.67	IRS[31]	72.70	48.10
	PAN[16]	82.81	66.07	IC-TL[49]	86.60	70.10
	SVDNet[13]	82.30	62.10	Ours	**88.80**	**79.35**
Weakly-supervised or unsupervised learning	TJ-AIDL[51]	58.20	26.50	WSMTAL[39]	56.60	31.20
	UTAL[50]	69.20	46.20	MAR[53]	67.70	40.00
	TFusion[29]	73.13	-	SSFR[54]	72.15	49.61
	DPR[52]	**84.82**	-	Ours	80.27	**66.93**

TABLE 6. Model performance on DukeMTMC-reID(%).

	Methods	Rank-1	mAP
Fully-supervised learning	BoW+KISSME[43]	25.16	12.17
	LOMO+XQDA[49]	30.75	17.04
	APR[32]	70.69	51.88
	MTA[48]	73.56	52.87
	Ours	**79.06**	**72.10**
Weakly-supervised or unsupervised learning	TJ-AIDL[51]	44.30	23.00
	SSFR[54]	52.90	33.60
	UTAL[50]	62.30	44.60
	MAR[53]	67.10	48.00
	Ours	**70.37**	**59.13**

1) EFFECTIVENESS ON FEATURE REPRESENTATION

The network structure of Baseline 1 is shown in Fig.5. This Base Network is Resnet-50 pre-trained in ImageNet dataset, which is fine-tuned on the training set of Market-1501. The Rank-1 CMC of baseline1 on the Market-1501 test set is 76.93%, and the Rank-1 CMC of DRNN-AMAF on the Market1501 test dataset is 88.80%, and the similar performance also appears on the DukeMTMC-reID dataset (shown in TABLE 4).

By comparing Rank-1 CMC between Baseline1 and DRNN-AMAF, we can see the positive impact of our approach on feature representation. The improvement of CMC means that the feature descriptor becomes more discriminative. Multi-branch structure in DRNN-AMAF has a similar positive impact as multi-task transfer learning to the model. Due to the task of multi-attribute regression and fusion, the network learns the feature descriptor which extracts not only discriminative patterns for identity but

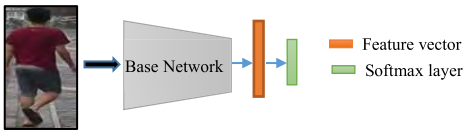


FIGURE 5. Baseline1 classification network structure.

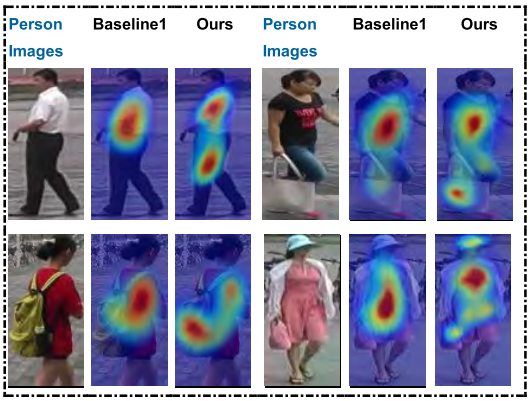


FIGURE 6. Comparisons of feature visualization with Baseline1 and the proposed DRNN-AMAF method in Person Re-ID.

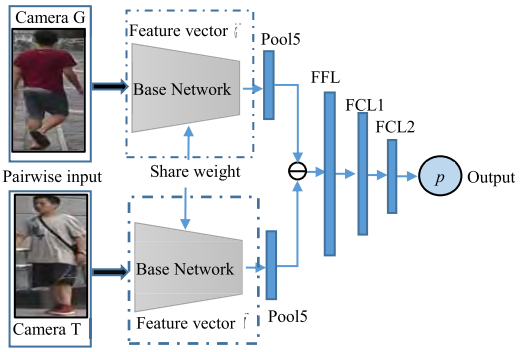
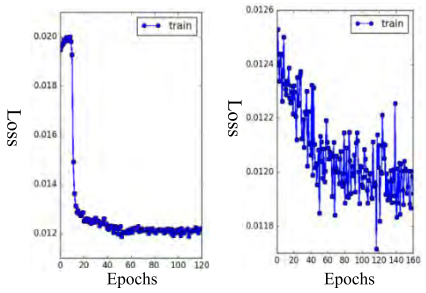


FIGURE 7. The regression network structure of Baseline2.



(a)Training loss of DRNN-AMAF (b) Training loss of Baseline2

FIGURE 8. Training loss visualization.

also extracts discriminative patterns for multiple attributes, which undoubtedly improves the generalization of the feature descriptor.

The visualization of feature representations which are extracted by Baseline1 and DRNN-AMAF is shown in Fig.6. From Fig.6, we can see Baseline1 can detect global information such as apparel and body shape, which are useful for person re-identification. The proposed model DRNN-AMAF can detect local details such as backpack, handbag and clothing-style, which is even more helpful for re-identification. This confirms our intuition about identity feature detected by DRNN-AMAF model has better generalization ability and higher discriminant power than baseline1.

2) EFFECTIVENESS ON EFFICIENT TRAINING

The network structure of Baseline2 is shown in Fig.7. The base network is also the ResNet-50 [36] pre-trained on the image classification dataset ImageNet [41]. Note that it is similar to DRNN-AMAF when DRNN-AMAF only has one branch. We aim to control the changes in the hyperparameter of the network, and only make Baseline2 different from our method in training strategy and supervised information, so as to reflect the advantages of our method by experiments.

In Baseline2, only the ID label is used as supervisory information. The value of the label is obtained by Equ. (2). Note that the ID label is a hard label. And the Baseline2 model is trained with the same iteration times as DRNN-AMAF training setting. The accuracy of the Baseline2 model prediction on Market1501 is low, with a Top-1 prediction accuracy of 13.2%. And the Top-1 prediction accuracy of DRNN-AMAF can reach to 85.73%.

TABLE 7. The training time efficiency.

Method	Train Time(s)/ epoch	Convergence Time(s)	Eval Time(s)/ image pair
Baseline2	165	More than 165*160	0.0012
DRNN-AMAF	183	183*53	0.0013

The training loss of DRNN-AMAF, changing along with the increase in the number of training epochs, is shown in Fig. 8(a), and the training loss of Baseline2 is shown in Fig. 8(b). From Fig. 8(a) and Fig. 8(b), it is obvious that DRNN-AMAF converges after 53 epochs training, while Baseline2 still doesn't converge after 160 epochs training. The specific time is shown in Table 7. Evaluation time is computed for a single image pair consist of probe and gallery image.

On analyzing the principle of the proposed method, one of the factors that the deep regression model is difficult to converge is that the deep regression model is difficult to be trained to obtain good feature extraction layers (e.g., convolutional layer and pooling layer in CNN). The proposed method resolves this problem by applying the idea of transfer learning. We first train the classification model to get the base network as the discriminative feature extraction, then transfer the knowledge to the training process of the final deep regression model. The difficulty of convergence is reduced. In addition, it is clear that the regression models have continuous outputs and usually require a continuous soft label. But for the classification problems, these models

TABLE 8. Performance with or without fusion weights L2 regular terms on the model.

Dataset	Market1501		DukeMTMC-reID	
Attribute	ID/gender/downcolor/ upcolor/age/up/down		ID/gender/downcolor/ upcolor/up	
With L2	Yes	No	Yes	No
Top-1(%)	85.73	62.93	76.78	23.65
Rank-1(%)	88.80	81.60	79.06	45.87

have discrete outputs and require one-hot hard label as the supervisory information. As for attributes of persons in image pairs, each attribute label can be only a hard label as either the same or different. Human experts cannot give quantitative similarity degrees objectively when labeling. These labels are not suitable for regression tasks. We fuse the multi-attribute labels to make the one-hot hard labels into a continuous soft label, which is more conducive to the regression task.

D. SOME PHENOMENA DURING THE EXPERIMENT

During the experiments, we found some interesting phenomena. After training, the multi-attribute fusion weight vector will tend to be one-hot. However, this is not the result we want to see. To this end, we add a regularization item (i.e. L2 norm of fusion weight) to the loss function:

$$\left(\sum_{i=1}^{m} w_i\right)^{1/2} \tag{13}$$

The L2 norm of the one-hot vector greater than the L2 norm of the soft vector when $w_i \in [0, 1]$. Therefore, the loss function with the regularization item can effectively limit the one-hot trend during the minimization process. For the loss function, with or without the regularization item, the difference in model performance is shown in Table 8. From Table 8 we can see our model can be well trained by the loss function with the L2 norm regularization item than the loss function without the L2 norm regularization item.

Another phenomenon, the training of regression models is relatively difficult, and the effect of training depends largely on the initialization of multi-attribute types, the model is easy to converge. From Table 3 we can see that when there are only ID and gender attributes, the model is easily optimized and has a relatively good performance. According to experiments, in general, for global attributes such as ID, age, gender, etc., given a larger initialization weight, the model training process is easier to converge. However, when randomly initializing the fusion weight, by using the adaptive multi-attribute fusion optimization method we propose, as well as the hyperparameter settings, we still get a good performance model after training.

The last phenomenon, for our experiments, using probability regression converge more quickly than using numerical regression (see Table 3). In the case of training the same epoch, probability regression has a better performance.

V. CONCLUSION

Taking full account of the real-world deployment of the person re-identification system and the process of human expert decision-making on person re-identification, this paper proposes a novel modeling method for the probability of identity-like regression. And we propose a training method named adaptive multi-attribute fusion optimization method to make the training of the deep regression model easy to converge. By applying the multi-attribute fusion and optimization method in the end-to-end person re-identification process, the neural network fuses the one-hot hard labels of multiple attributes into one soft label to make the label more suitable for the regression task. Moreover, the proposed method also suitable for weakly-supervised setting without identity supervised information. The feasibility of the proposed method is verified on current mainstream datasets and achieves good performance. In future work, two directions can be considered to improve our model. First, the proposed method needs a large number of labeled matching pairs for training due to the increasing in the amount of parameters brought by the multi-branch structure. It can be resolved by researching a novel cross-dataset transfer learning approach to train on multiple small datasets with different visual attributes and fuse the "knowledge" learned by the model together. Second, inspired by the literature [56], we are going to further improve our model by adding new network structure to explicitly resist the adverse effect such as viewpoint or lighting variations and further improve the training strategy of the regression model.

REFERENCES

[1] 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.*, Columbus, OH, USA, Jun. 2014, pp. 152–159.

[2] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Deep metric learning for person re-identification," in *Proc. 22nd Int. Conf. Pattern Recognit.*, Stockholm, Sweden, Aug. 2014, pp. 34–39.

[3] E. Ahmed, M. Jones, and K. T. Marks, "An improved deep learning architecture for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, USA, Jun. 2015, pp. 3908–3916.

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

[5] L. Wu, C. Shen, and A. Hengel, "PersonNet: Person re-identification with deep convolutional neural networks," Jun. 2016, *arXiv:1601.07255*. [Online]. Available: https://arxiv.org/abs/1601.07255

[6] X. Qian, Y. Fu, Y.-G. Jiang, T. Xiang, and X. Xue, "Multi-scale deep learning architectures for person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 5399–5408.

[7] H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan, "End-to-end comparative attention networks for person re-identification," *IEEE Trans. Image Process*, vol. 26, no. 7, pp. 3492–3506, Jul. 2017.

[8] S. Wu, Y. Chen, X. Li, A. Wu, J. You, and W. Zheng, "An enhanced deep feature representation for person re-identification," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Lake Placid, NY, USA, Mar. 2016, pp. 1–8.

[9] 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. (ICCV)*, Venice, Italy, Oct. 2017, pp. 3754–3762.

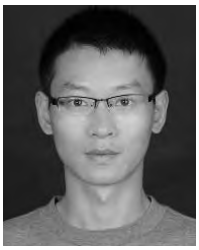
[10] T. Xiao, H. Li, W. Ouyang, and X. Wang, "Learning deep feature representations with domain guided dropout for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 1249–1258.

[11] Y. Chen, X. Zhu, and S. Gong, "Person re-identification by deep learning multi-scale representations," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Istanbul, Turkey, Oct. 2017, pp. 2590–2600.
[12] T. Matsukawa and E. Suzuki, "Person re-identification using CNN features learned from combination of attributes," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Cancun, Mexico, Dec. 2016, pp. 2428–2433.
[13] Y. Sun, L. Zheng, W. Deng, and S. Wang, "SVDNet for pedestrian retrieval," in *Proc. IEEE Int. Conf. Comput. Vis.*, Venice, Italy, Oct. 2017, pp. 3800–3808.
[14] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Salt Lake City, UT, USA, Sep. 2018, pp. 480–496.
[15] 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. CVPR*, Honolulu, HI, USA, Jul. 2017, pp. 384–393, 2017.
[16] Z. Zheng, L. Zheng, and Y. Yang, "Person alignment network for large-scale person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, to be published. doi: 10.1109/TCSVT.2018.2873599.
[17] M. Geng, Y. Wang, T. Xiang, and Y. Tian, "Deep transfer learning for person re-identification," Nov. 2016, *arXiv:1611.05244*. [Online]. Available: https://arxiv.org/abs/1611.05244
[18] Z. Zheng, L. Zheng, and Y. Yang, "A discriminatively learned cnn embedding for person reidentification," *ACM Trans. Multimed. Comput. Commun. Appl.*, vol. 14, no. 1, p. 13, Jan. 2017.
[19] Z. Zhang and T. Si, "Learning deep features from body and parts for person re-identification in camera networks," *EURASIP J. Wireless Commun. Netw.*, vol. 2018, p. 52, Dec. 2018.
[20] 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. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 1335–1344.
[21] 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. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 403–412.
[22] Y. Yuan, J. Zhang, and Q. Wang, "Modeling unknown class centers for metric learning on person re-identification," *IEEE Access*, vol. 6, pp. 40602–40610, 2018.
[23] M. Köstinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, Jun. 2012, pp. 2288–2295.
[24] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *J. Mach. Learn. Res.*, vol. 10, pp. 207–244, Feb. 2009.
[25] Z. Li, S. Chang, F. Liang, T. S. Huang, L. Cao, and J. Smith, "Learning locally-adaptive decision functions for person verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Portland, OR, USA, Jun. 2013, pp. 3610–3617.
[26] D. Chung, K. Tahboub, and E. J. Delp, "A two stream siamese convolutional neural network for person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 1983–1991.
[27] N. McLaughlin, J. M. del Rincon, and P. Miller, "Recurrent convolutional network for video-based person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1325–1334.
[28] Z. Zhou, Y. Huang, W. Wang, L. Wang, and T. Tan, "See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 6776–6785.
[29] J. Lv, W. Chen, Q. Li, and C. Yang, "Unsupervised cross-dataset person re-identification by transfer learning of spatial-temporal patterns," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, Jun. 2018, pp. 7948–7956.
[30] J. Zhang, N. Wang, and L. Zhang, "Multi-shot pedestrian re-identification via sequential decision making," May 2018, *arXiv:1712.07257*. [Online]. Available: https://arxiv.org/abs/1712.07257
[31] H. Wang, X. Zhu, S. Gong, and T. Xiang, "Person re-identification in identity regression space," *Int. J. Comput. Vis.*, vol. 126, no. 12, pp. 1288–1310, Dec. 2018.
[32] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, Z. Hu, C. Yan, and Y. Yang, "Improving person re-identification by attribute and identity learning," Apr. 2017, *arXiv:1703.07220*. [Online]. Available: https://arxiv.org/abs/1703.07220
[33] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," Apr. 2015, *arXiv:1409.1556*. [Online]. Available: https://arxiv.org/abs/1409.1556
[34] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 1–9.
[35] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, Jun. 2016, pp. 2818–2826.
[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778.
[37] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 4700–4708.
[38] H.-X. Yu, A. Wu, and W.-S. Zheng, "Cross-view asymmetric metric learning for unsupervised person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, Venice, Italy, Oct. 2017, pp. 994–1002.
[39] C. Su, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Multi-type attributes driven multi-camera person re-identification," *Pattern Recognit.*, vol. 75, pp. 77–89, Mar. 2017.
[40] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, San Diego, CA, USA, Jun. 2005, pp. 539–546.
[41] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Miami Beach, FL, USA, Jun. 2009, pp. 248–255.
[42] X. Wang and R. Zhao, "Person re-identification: System design and evaluation overview," in *Person Re-Identification*. London, U.K.: Springer, 2014, pp. 351–370.
[43] 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.*, Boston, MA, USA, Dec. 2015, pp. 1116–1124.
[44] J. Yosinski, J. Clune, and Y. Bengio, "How transferable are features in deep neural networks?" in *Proc. Adv. Neural Inf. Process. Syst.*, Montréal, QC, Canada, 2014, pp. 3320–3328.
[45] 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.
[46] Z. Zhong, L. Zheng, and G. Kang, "Random erasing data augmentation," Nov. 2017, *arXiv:1708.04896*. [Online]. Available: https://arxiv.org/abs/1708.04896
[47] E. Ustinova, Y. Ganin, and V. Lempitsky, "Multi-region bilinear convolutional neural networks for person re-identification," in *Proc. 14th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Lecce, Italy, Aug./Sep. 2017, pp. 2993–3003.
[48] J. Wang and M. Lyu, "Multi-task network learning representation features of attributes and identity for person re-identification," in *Proc. Chin. Conf. Biometric Recognit.* Cham, Switzerland: Springer, 2018, pp. 689–699.
[49] D. Wu, S.-J. Zheng, W. Z. Bao, X.-P. Zhang, C.-A. Yuan, and D.-S. Huang, "A novel deep model with multi-loss and efficient training for person re-identification," *Neurocomputing*, vol. 324, pp. 69–75, Jan. 2019.
[50] M. Li, X. Zhu, and S. Gong, "Unsupervised tracklet person re-identification," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published.
[51] J. Wang, X. Zhu, S. Gong, and W. Li, "Transferable joint attribute-identity deep learning for unsupervised person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, Jun. 2018, pp. 2275–2284.
[52] R. Panda, "Visual learning with weak supervision: Applications in video summarization and person re-identification," Ph.D. dissertation, Univ. California, Riverside, Riverside, CA, USA, 2018. [Online]. Available: https://escholarship.org/content/qt7w08k4sd/qt7w08k4sd.pdf
[53] H.-X. Yu, W.-S. Zheng, A. Wu, X. Guo, S. Gong, and J.-H. Lai, "Unsupervised person re-identification by soft multilabel learning," Mar. 2019, *arXiv:1903.06325*. [Online]. Available: https://arxiv.org/abs/1903.06325
[54] X. Xin, J. Wang, R. Xie, S. Zhou, W. Huang, and N. Zheng, "Semi-supervised person re-identification using multi-view clustering," *Pattern Recognit.*, vol. 88, pp. 285–297, Apr. 2019.

[55] H. Yao, S. Zhang, R. Hong, Y. Zhang, C. Xu, and Q. Tian, "Deep representation learning with part loss for person re-identification," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 2860–2871, Jun. 2019.
[56] J. Zhu, H. Zeng, J. Huang, S. Liao, Z. Lei, C. Cai, and L. Zheng, "Vehicle re-identification using quadruple directional deep learning features," *IEEE Trans. Intell. Transp. Syst.*, to be published.

YINGCHUN GUO received the Ph.D. degree from the School of Information, Tianjin University, Tianjin, China, in 2006. She is currently an Associate Professor with the School of Artificial Intelligence, Hebei University of Science and Technology, Tianjin. Her research interests include image saliency and its application, image processing, artificial intelligence, and image compression.

KUNPENG ZHAO received the B.S. degree in software engineering from the City Institute, Dalian University of Technology, China, in 2017. He is currently pursuing the M.S. degree with the Hebei University of Technology. His research interests include pedestrian re-identification, pattern recognition, and deep learning.



XIAOKE HAO received the B.S. and M.S. degrees from the Nanjing University of Information Science and Technology, Nanjing, China, in 2009 and 2012, respectively, and the Ph.D. degree in computer science and technology from the Nanjing University of Aeronautics and Astronautics, Nanjing, in 2017. He is currently an Assistant Professor with the School of Artificial Intelligence, Hebei University of Technology. His research interests include machine learning, pattern recognition, and medical image analysis.

MING YU (M'15) received the Ph.D. degree in communication and information systems from the Beijing Institute of Technology. Since 1989, he has been a Teacher with the Hebei University of Technology and a Full Professor with the School of Computer Science and Engineering, since 2000. His research interests include biometrics of voice and image vision information fusion, image mathematical transformation, efficient algorithm of image and video coding, computer networks, and high compression image video transmission under wireless mobile networks.