Fusion of Multiple Person Re-id Methods With Model and Data-Aware Abilities

De Cheng, Zhihui Li, Yihong Gong, *Fellow, IEEE*, and Dingwen Zhang

***Abstract*—Person re-identification (person re-id) has attracted rapidly increasing attention in computer vision and pattern recognition research community in recent years. With the goal of providing match ranking results between each query person image and the gallery ones, the person re-id technique has been widely explored and a large number of person re-id methods have been developed. As these algorithms leverage different kinds of prior assumptions, image features, distance matching functions, *et al.*, each of them has its own strengths and weaknesses. Inspired by these facts, this paper proposes a novel person re-id method based on the idea of inferring superior fusion results from a variety of previous base person re-id algorithms using different methodologies or features. To achieve this goal, we propose a novel framework which mainly consists of two steps: 1) a number of existing person re-id methods are implemented, and the ranking results are obtained in the test datasets. and 2) the robust fusion strategy is applied to obtain better re-ranked matching results by simultaneously considering the recognition abilities of various base re-id methods and the difficulties of different gallery person images to be correctly recognized under the generative model of labels, abilities, and difficulties framework. Comprehensive experiments show the effectiveness of our proposed method, and we have received state-of-the-art results on recent popular person re-id datasets.**

***Index Terms*—Fusion, generative model of labels, abilities, and difficulties (GLAD), person reidentification (person re-id).**

I. Introduction

PERSON RE-IDENTIFICATION (person re-id) is the problem of matching images of the same individuals across multiple cameras, or across time within a single camera. It has attracted more attention in the computer vision and pattern recognition communities due to its importance for a wide range of applications, such as video surveillance, human–computer interaction, robotics, etc. The task is extremely challenging due to the following reasons: 1) dramatic variations in visual appearance and ambient environment caused by different viewpoints from different cameras; 2) significant changes in human pose across time and space; 3) background clutter and occlusions; and 4) different individuals that share similar appearances. Moreover, with little or no visible faces, in many cases, the use of biometric and soft-biometric approaches is not applicable. Fig. 1 illustrates some examples of the matched pairs in four challenging person re-id datasets: *i*-LIDS, PRID2011, VIPeR [16], and CUHK01 [24]. Images in a red bounding box contain the same person.

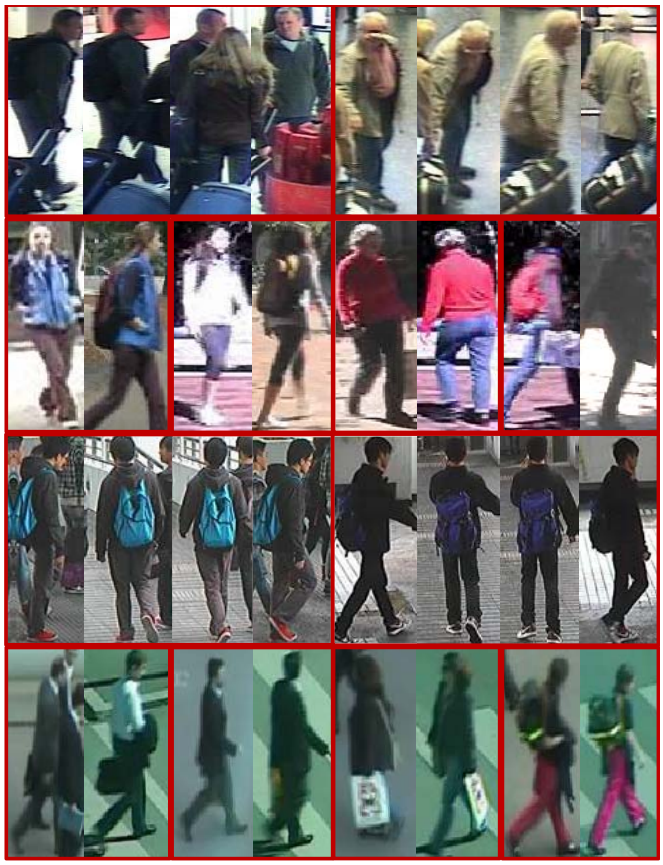


Fig. 1. Matched examples in datasets *i*-LIDS, VIPeR, CUHK01, and PRID2011. Each row shows matched examples from the same dataset. Images in a red bounding box contain the same person.

In recent years, a large number of person re-id methods have been proposed in the literature. These methods range from proposing different visual appearance feature representations to developing advanced similarity metric functions, to exploring deep convolutional neural network (DCNN) models for feature extraction and similarity metric learning [2], [7], [10]–[12], [14], [19], [31], [37]. However, due to the extremely

Manuscript received October 8, 2017; revised July 3, 2018; accepted August 29, 2018. Date of publication October 9, 2018; date of current version December 16, 2019. This paper was recommended by Associate Editor Q. Ji. *(De Cheng and Zhihui Li are co-first authors.) (Corresponding author: Zhihui Li.)*

D. Cheng and Y. Gong are with the School of Electronics and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China.

Z. Li is with the School of Computer Science and Engineering, University of New South Wales, Sydney, NSW 2052, Australia (e-mail: zhihuics@gmail.com).

D. Zhang is with the School of Automation, Northwestern Polytechnical University, Xi'an 710100, China.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCYB.2018.2869739

challenging nature of the task, even the best performances achieved by the latest state-of-the-art methods are still far from satisfactory on the commonly used benchmark datasets shown in Fig. 1. There is still large room for performance improvement to meet the needs for real applications.

Since different person re-id methods employ different kinds of feature representations, similarity metrics, prior assumptions, etc., each of them has its own strengths and weaknesses. It is extremely difficult, or even impossible, for a single method to work well under all scenarios. This observation motivates us to improve person re-id performances by fusing the strengths of the existing methods. More specifically, we aim to develop a fusion method that is able to automatically infer and fuse the strengths of the base person re-id methods, and only depending on the outputs of all the based re-id methods.

In this paper, we propose a novel person re-id fusion method based on the generative model of labels, abilities, and difficulties (GLAD) model [39]. The GLAD approach is based on standard probabilistic inference on a model of the labeling process. It can be used to simultaneously infer the expertise of each labeler, the difficulty of each image, and the most probable label for each image. In this paper, we have extended the GLAD approach as a data fusion method to infer the final ranking results for person re-id. The proposed fusion method has several desired properties. First, it is method-aware. It can automatically infer the person re-id ability of each base method, based on which assigns a different weight to each method. Second, it is data-aware. Rather than treating all gallery images equally, as in most existing fusion approaches, the proposed method additionally infers the labeling difficulty of each gallery image and builds corresponding relationships between gallery images and base person re-id methods, for example, method A is more specialized in ranking images with a high resolution while method B is better at ranking persons wearing texture clothes. Third, the proposed method is of great robustness. It not only considers more factors comprehensively in the fusion process, such as each base method's person re-id ability, the difficulty of each gallery image, etc., but also has a theoretically sound mechanism to infer these key factors. Thus, such an inference scheme can yield a better fusion result yet without additional supervision and labeled training data.

To summarize, the contributions of the proposed work are three-fold.

1) We propose improving person re-id performances by fusing the strengths of the existing base re-id methods. This approach enables us to not only build upon but also make full use of the base person re-id methods, and obtain complementary properties of these methods.
2) We propose a novel person re-id fusion method based on the GLAD model, which considers both the person re-id ability of each base method and the labeling difficulty of each gallery image. It can automatically infer a strong label for each gallery image from multiple base methods in a fusion fashion, which only depends on the ranking outputs of all the based re-id methods.
3) Comprehensive experimental evaluations on four commonly used benchmark datasets demonstrate the remarkable performance gains achieved by the proposed fusion method (about 5% improvement in terms of Top1 matching result).

The rest of this paper is organized as follows. In Section II, we briefly review the related works. Section III introduces the proposed fusion method for person re-id. The experimental results, comparisons, and analysis are presented in Section IV. The conclusion comes in Section V.

II. Related Work

In recent years, many person re-id methods have been proposed in the literature. These methods can be divided into two main categories: 1) methods based on traditional pattern recognition and machine learning approaches and 2) methods using DCNN. Almost all methods attempt to address the following two problems: 1) developing robust and discriminative feature representations for both the query and gallery images and 2) finding suitable similarity metrics or ranking functions to determine whether a query person image appears in the gallery images. Systems in the literature use various combinations of features and similarity/rank learning methods. Below, we review representative works of the above two categories.

A. Traditional Methods

Research works in this category mainly focus on seeking better feature representations and similarity metrics. A great amount of research effort has been made to develop features which aims to be invariant to illumination, pose, and viewpoint changes. The traditional handcraft features that are used for the person re-id task include color histograms and their variants [20], [21], [23], [28]; local binary patterns [20], [21], [23], [28]; Gabor features [23]; color names [43]; and other visual appearance and contextual cues and combinations of them [38]. There are also features especially designed to achieve certain invariance properties. For example, Zhao *et al.* [46] learned a mid-level filter from the patch cluster to achieve the cross-view invariance. Wu *et al.* [40] introduced a person viewpoint-invariant descriptor by integrating the pose prior information of the training data. Köestinger *et al.* [21] proposed one intradistribution structure of the color-based features to make it robust to certain illumination changes. Li and Wang [23] carried out the person re-id task by matching images observed in different camera views with complex cross-view transformations. Liao *et al.* [26] developed a feature descriptor which analyzes the horizontal occurrence of local features, and maximizes the occurrence to make a stable representation against viewpoint changes. Ma *et al.* [28] represented person images via the covariance descriptor which is robust to illumination and background variations. Zhao *et al.* [45] proposed using the distinct salience region to distinguish the correctly matched person from others.

The basic idea behind distance metric learning is to find a mapping function from the feature space to a distance space with certain merits, such as feature vectors from the same person to be closer than those from different ones [47]. Representative works in this category include the KISSME

method [21] that introduces a simple yet effective strategy to learn a distance metric from equivalence constraints based on a statistical inference perspective. The LFDA method [32] that utilizes local Fisher discriminant analysis to map high dimensional features into a more discriminative low dimensional space, and MFA [42] that uses marginal Fisher analysis to tackle the person re-id problem. After that, [41] proposes the kernel-based metric learning methods, denoted as kLFDA and kMFA, which is the extension of the traditional LFDA, MFA, and several other metric learning methods. Also, there are some other metric learning-based methods that are introduced. For instance, [48] proposes a relative distance learning method based on the probabilistic prospective; [29] learns a distance metric by integrating the sparse pairwise similarity constraints into the objectives; [27] casts the person re-id problem as the retrieval related task by considering the list-wise similarity; [5] proposes a kernel-based metric learning method to explore the nonlinearity relationship of samples in the feature space; and [18] learns a discriminative metric by using relaxed pairwise constraints, etc.

B. DCNN-Based Methods

Since deep learning networks have achieved great success in various computer vision tasks [13], [15], [22], [35], [36], it becomes increasingly popular to apply DCNN models to the person re-id task. Recently, the state-of-the-art performances on almost all widely used person re-id benchmark datasets, such as *i*-LIDS, VIPeR, CUHK01, etc., are all obtained by DCNN-based methods. In the following text, we briefly introduce those deep learning-based approaches, and some of them may be selected as the base methods for our proposed algorithm. DeepReID [25] proposed the filter pairing neural network which jointly handles the problems of misalignment, occlusion, black cluster, etc. mFilter [46] learns the mid-level filters to get the local discriminative features for person re-id. Ahmed *et al.* [1] took pair-wise images as its inputs, and outputs a similarity value indicating whether the two input images depict the same person or not. Yi *et al.* [44] constructed a siamese neural network (denoted as SiameseNet in this paper) to learn pairwise similarity, and also used body parts to train their CNN models. In their work, person images are cropped into three overlapped parts which are used to train three independent networks. Ding *et al.* [9] proposed to use the triplet network architecture for person re-id (denoted as TripletNet), and Cheng *et al.* [6] extended the triplet network using multichannel and multiparts way to further improve the re-id performance (denoted as TripletPatrsNet in this paper).

However, our proposed method for person re-id is different from above all. We would like to improve the person re-id performance by an fusion strategy based on some base re-id methods. The base re-id methods that we selected in our proposed fusion algorithm include: KISSME, LFDA, MFA, SiameseNet, TripletNet, and TriepletPartsNet. We choose these methods because their source codes are publicly available, and their methodologies belong to the above two different categories. Yet, few fusion methods for the person re-id task can be found in the literature. The most relevant work to ours is from Paisitkriangkrai *et al.* [30]. Their method employed a combination of multiple hand-crafted low and high level visual features, and realized the person re-id task using a reranking framework. It achieved the state-of-the-art results on most of the benchmark datasets. In this paper, we also use the fusion approach to tackle the person re-id problem. However, we solve the problem in a very different way. In the following section, we will present our proposed method in details.

III. Proposed Fusion Method

Fig. 2 presents the overview of the proposed person re-id fusion method. Our method consists of three main modules: 1) for a given query image, generate the initial rank lists of gallery images by multiple base person re-id methods; 2) initialize the fusion model parameters; and 3) infer the final rank list using the GLAD fusion model. In this section, we provide detailed descriptions of the GLAD fusion model used in module 3.

Consider a person re-id dataset that contains N, L individuals in the gallery and query subsets, respectively.[1] Given a query image, we can get a rank list of individuals in the gallery subset according to the distances between the query image and all the gallery images produced by a person re-id method. Assume that there are M base person re-id methods $\{\Phi_1, \Phi_2, \ldots, \Phi_M\}$. For each query image, M base ranking lists $RL = \{rl_1, rl_2, \ldots, rl_M\}$ can be obtained using the M base person re-id methods. Our goal is to fuse these weak predictions into a strong and reliable one in an unsupervised fusion manner.

The proposed fusion method needs to first binarize the rank list produced by each person re-id method Φ_i. We accomplish this by simply assigning label "1" to the top K gallery images in the rank list, and label "0" to the remaining ones, where labels 1 and 0 denote match and no match to the query image, respectively. We use the line search method to find the optimal K in our implementation. In the following part of this section, we describe the proposed fusion method from the following three aspects: 1) definitions of the important variables; 2) the algorithm for inferring the strong posterior probability prediction; and 3) the optimization algorithm for computing the optimal posterior probability estimation.

A. Important Variables Definition

Our proposed person re-id fusion method employs the following important variables.

1) Recognition Ability of Each Base Person Re-id Method: Different person re-id methods have different performance accuracies for matching query images to the corresponding gallery images. We define the variable $\alpha_i \in [0, +\infty)$ to represent the recognition ability of person re-id method Φ_i.

2) Difficulty of Each Gallery Image: Different gallery images also have different difficulty for being correctly recognized. For example, some gallery images contain a clear shot

[1]The *single shot* evaluation, which is commonly adopted person re-id performance evaluation protocol, assumes that each individual has only one image in both the gallery and the query subsets.

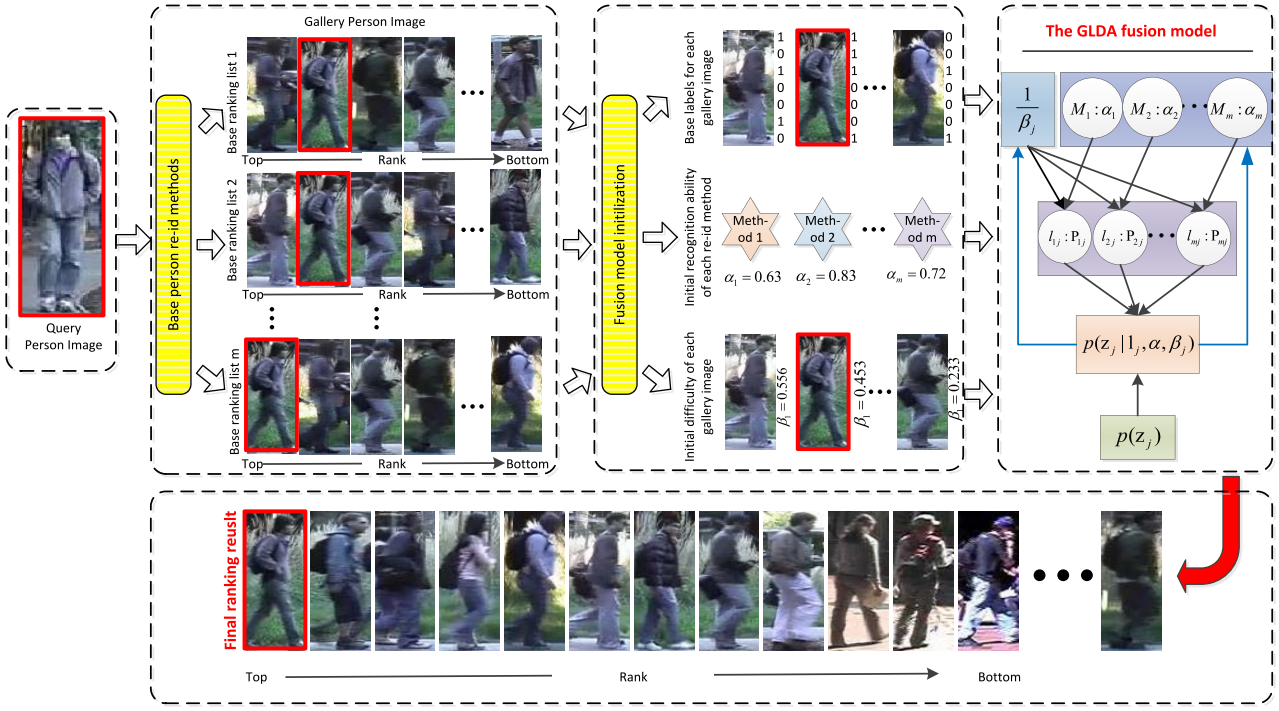


Fig. 2. Overview of the proposed fusion method. It consists of three main modules: 1) for a given query image, generation of the initial rank lists of gallery images by multiple base person re-id methods, 2) initialization of the fusion model parameters, and 3) inference of the final rank list using the GLAD fusion model.

of a single person wearing a cloth with distinct colors, while others contain a person partially occluded by certain object. Clearly, the latter is more difficult to be correctly recognized than the former. We introduce the variable $(1/\beta_j) \in [0, +\infty)$ to indicate the difficulty of correctly matching gallery image j to its query image, where the larger of $(1/\beta_j)$, and the more difficulty of the corresponding gallery image to be correctly recognized.

3) Strong and Weak Labels of Each Gallery Image: We use z_j to denote the strong label of gallery image j, and l_{ij} to denote a weak label assigned to gallery image j by person re-id method Φ_i. Both z_j and l_{ij} take binary values $\{0, 1\}$, where labels 1 and 0 mean match and no match to the query image, respectively. Our goal is to infer the most likely value of z_j from the observed weak labels l_{ij}, where $i = \{1, 2, \ldots, M\}$. We accomplish this goal by estimating the maximum a posterior (MAP) probability $p(z_j|\boldsymbol{l}_j, \boldsymbol{\alpha}, \beta_j)$.

B. Strong Label Posterior Probability Derivation

Let $\boldsymbol{l}_j$ denote the set of weak labels assigned to gallery image j by each of the M person re-id methods, $\boldsymbol{l}_j = \{l_{ij}\}_{i=1}^{M}$, $\boldsymbol{\alpha}$ the set of recognition abilities of the M person re-id methods, $\boldsymbol{\alpha} = \{\alpha_i\}_{i=1}^{M}$, and $(1/\boldsymbol{\beta})$ the set of difficulties of the N gallery images, $\boldsymbol{\beta} = \{\beta_i\}_{i=1}^{N}$. With all the important variables and symbols defined above, the posterior probability of strong label z_j can be computed as follows [39]:

$$p(z_j|\boldsymbol{l}_j, \boldsymbol{\alpha}, \boldsymbol{\beta}) = p(z_j|\boldsymbol{l}_j, \boldsymbol{\alpha}, \beta_j) \propto p(z_j|\boldsymbol{\alpha}, \beta_j) p(\boldsymbol{l}_j|z_j, \boldsymbol{\alpha}, \beta_j) \propto p(z_j) \prod_i p(l_{ij}|z_j, \alpha_i, \beta_j). \quad (1)$$

In the above derivations, line two is obtained by applying Bayes' rule to $p(z_j|\boldsymbol{l}_j, \boldsymbol{\alpha}, \beta_j)$, and $p(z_j|\boldsymbol{\alpha}, \beta_j) = p(z_j)$ is based on the conditional independence assumption that the true label z_j has no direct connections with all the person re-id methods' recognition abilities $\boldsymbol{\alpha}$ and the gallery image's difficulty $(1/\beta_j)$.

In (1), $p(l_{ij}|z_j, \alpha_i, \beta_j)$ can be interpreted as the probability of the base label l_{ij} being correct, which is equivalent to the probability that the base label l_{ij} equals to the strong label z_j. We define this probability as follows [39]:

$$p(l_{ij}|z_j, \alpha_i, \beta_j) = p(l_{ij} = z_j|\alpha_i, \beta_j) = \frac{1}{1 + e^{-\alpha_i \beta_j}}. \quad (2)$$

Equation (2) stipulates that a stronger person re-id method with a higher α_i should have a higher probability of correctly labeling an easy gallery image with a lower $(1/\beta_j)$. In addition, when a gallery image's difficulty $(1/\beta_j)$ approximates $+\infty$, and/or a person re-id method's recognition accuracy α_i approach zero, $p(l_{ij} = z_j|\alpha_i, \beta_j)$ reaches 0.5, which means that the base label l_{ij} is a random guess and makes no contribution to estimating the final strong label.

The casual relationships of the recognition ability of each base person re-id method, the difficulty of each gallery image, initial base labels of the gallery images, and the desired strong fusion labels are shown in Fig. 3.

C. MAP Probability Estimation With EM Algorithm

In (1), the base labels l_{ij} for a given gallery image j are the only observed variables. The unobserved variables include the strong label z_j, the difficulty parameter β_j of each gallery image j, and the recognition ability α_i of each person re-id

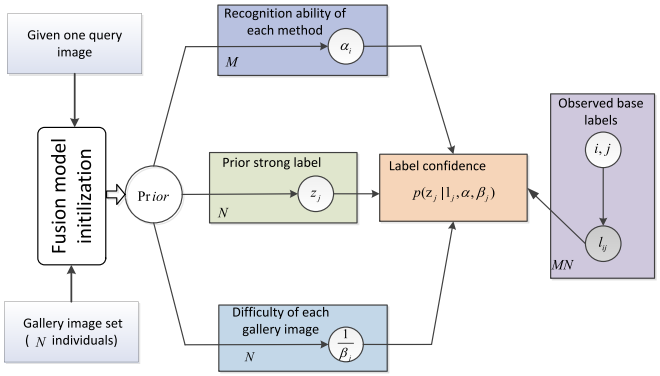


Fig. 3. Relationships between the inferred strong label confidence and the recognition ability of each base method, observed base labels, and the difficulty of each gallery image. The numbers M and N in the left corner of each bounding box represent the total number of these parameters for each query image.

method Φ_i. Our goal is to efficiently search for the most probable values of the unobserved variables z_j, α_i, and β_j given the observed variables. Here, we use expectation-maximization (EM) algorithm to compute the maximum likelihood estimates of the unobserved variables and infer the MAP probability of z_j simultaneously. The main steps of the EM algorithm are described as follows.

E-Step: Use (1) to compute the posterior probability of all z_j's based on the observed base labels l_{ij} and the estimated values of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ from the last M-step.

M-Step: Compute the optimal values of the unobserved variables $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ by maximizing the standard auxiliary function $Q(\boldsymbol{\alpha}, \boldsymbol{\beta})$ [33], [39]

$$\left(\boldsymbol{\alpha}', \boldsymbol{\beta}'\right) = \arg\max_{(\boldsymbol{\alpha}, \boldsymbol{\beta})} Q(\boldsymbol{\alpha}, \boldsymbol{\beta}) \tag{3}$$

where $Q(\boldsymbol{\alpha}, \boldsymbol{\beta})$ is can be defined as the expectation of joint log-likelihood of the observed and unobserved variables $(\boldsymbol{l}, \boldsymbol{z})$, given the variables $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ [33], [39]

$$\begin{aligned} Q(\boldsymbol{\alpha}, \boldsymbol{\beta}) &= \mathrm{E}\left[\ln p(\boldsymbol{l}, \boldsymbol{z}|\boldsymbol{\alpha}, \boldsymbol{\beta})\right] \\ &= \mathrm{E}\left[\ln(p(\boldsymbol{z}|\boldsymbol{\alpha}, \boldsymbol{\beta}) \cdot p(\boldsymbol{l}|\boldsymbol{z}, \boldsymbol{\alpha}, \boldsymbol{\beta}))\right] \\ &= \mathrm{E}\left[\ln(p(\boldsymbol{z}) \cdot p(\boldsymbol{l}|\boldsymbol{z}, \boldsymbol{\alpha}, \boldsymbol{\beta}))\right] \\ &= \mathrm{E}\left[\ln \prod_j \left(p(z_j) \prod_i p\left(l_{ij}|z_j, \alpha_i, \beta_j\right)\right)\right] \\ &= \sum_j \mathrm{E}\left[\ln p\left(z_j\right)\right] + \sum_{ij} \mathrm{E}\left[\ln p\left(l_{ij}|z_j, \alpha_i, \beta_j\right)\right]. \end{aligned} \tag{4}$$

In the above equation, the expectation is taken over z_j given the old parameter values $\boldsymbol{\alpha}^{\text{old}}$ $\boldsymbol{\beta}^{\text{old}}$ as estimated in the last M-step. Line two is derived by applying Bayes' rule to $\ln p(\boldsymbol{l}, \boldsymbol{z}|\boldsymbol{\alpha}, \boldsymbol{\beta})$, and line three is obtained because $\boldsymbol{z}$ is conditional independent of $\boldsymbol{\alpha}, \boldsymbol{\beta}$. Using gradient ascent, we find values of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ that locally maximize $Q(\boldsymbol{\alpha}, \boldsymbol{\beta})$, and the following provides the derivations of the gradients of $Q(\boldsymbol{\alpha}, \boldsymbol{\beta})$ with respect of α_i and β_j.

Let us define $p_j^k = p(z_j = k|\boldsymbol{l}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ [39]. Then we can expand this expectation as

$$\begin{aligned} Q(\boldsymbol{\alpha}, \boldsymbol{\beta}) = &\sum_j \sum_{k=0}^{l} p_j^k \ln p\left(z_j = k\right) \\ &+ \sum_{ij} \sum_{k=0}^{l} p_j^k \ln p\left(l_{ij}|z_j = k, \alpha_i, \beta_j\right). \end{aligned} \tag{5}$$

Based on (2), we can compute $p(l_{ij}|z_j = k, \alpha_i, \beta_j)$ as

$$p\left(l_{ij}|z_j = 1, \alpha_i, \beta_j\right) = \sigma\left(\alpha_i\beta_j\right)^{l_{ij}}\left(1 - \sigma\left(\alpha_i\beta_j\right)\right)^{1-l_{ij}} \tag{6}$$

and

$$p\left(l_{ij}|z_j = 0, \alpha_i, \beta_j\right) = \sigma\left(\alpha_i\beta_j\right)^{1-l_{ij}}\left(1 - \sigma\left(\alpha_i\beta_j\right)\right)^{l_{ij}} \tag{7}$$

where $\sigma(x) = [1/(1 + e^{-x})]$ is the logistic function. To avoid clutter, we represent $\sigma(\alpha_i\beta_j)$ simply as σ. Then, after expanding the summation over k into the two cases $z = 0$ and $z = 1$, we get

$$\begin{aligned} Q(\boldsymbol{\alpha}, \boldsymbol{\beta}) = &\sum_j \Big(p_j^1 \ln p\left(z_j = 1\right) + \Big(p_j^0 \ln p\left(z_j = 0\right) \\ &+ \sum_{ij} p_j^1\left[l_{ij} \ln \sigma + \left(1 - l_{ij}\right) \ln(1 - \sigma)\right] \\ &+ \sum_{ij} p_j^0\left[\left(1 - l_{ij}\right) \ln \sigma + l_{ij} \ln(1 - \sigma)\right]. \end{aligned} \tag{8}$$

Taking the first derivatives causes the first summation to vanish since it is constant with respect to $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ [39]. Using the fact that

$$\frac{d}{dx}\sigma(x) = \sigma(x)(1 - \sigma(x)) \tag{9}$$

we can differentiate $\boldsymbol{Q}$ to arrive at

$$\begin{aligned} \frac{\partial Q}{\partial \alpha_i} &= \sum_j p_j^1\left(l_{ij}(1 - \sigma)\beta_j - \left(1 - l_{ij}\right)\sigma\beta_j\right) \\ &\quad + \sum_j p_j^0\left(\left(1 - l_{ij}\right)(1 - \sigma)\beta_j - l_{ij}\sigma\beta_j\right) \\ &= \sum_j \Big(p_j^1 l_{ij} + p_j^0\left(1 - l_{ij}\right) - \Big(p_j^1 + p_j^0\Big)\sigma\Big)\beta_j \\ &= \sum_j \Big(p_j^1 l_{ij} + p_j^0\left(1 - l_{ij}\right) - \sigma\Big)\beta_j \\ &\quad \text{since} \quad p_j^0 + p_j^1 = 1. \end{aligned} \tag{10}$$

Similarly, we can derive

$$\frac{\partial Q}{\partial \beta_j} = \sum_i \Big(p_j^1 l_{ij} + p_j^0\left(1 - l_{ij}\right) - \sigma\Big)\alpha_i. \tag{11}$$

The gradient equation for $(\partial Q/\partial \alpha_i)$ has an intuitive interpretation: the first two items compute the empirical probability of the given label l_{ij} being correct given posterior probabilities of z_j from the previous E-step [39]. The σ that is subtracted is the model's current estimate of the probability that l_{ij} is correct given the current estimate of the re-id method's recognition ability and each gallery image's recognition difficulty. Hence, the likelihood function will locally

increase by increasing the recognition ability α_i, if the empirical estimate of the number of correct images recognized by the ith person re-id method is greater than its previous belief of correctness. Similar function applies to $(\partial Q/\partial \beta_j)$ with regards to image recognition difficulty. To find locally optimal values of the $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ parameter we set the gradients to zero. The resulting equations are nonlinear and thus need to be solved using iterative methods [39].

D. Priors on the Variables

The parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ in the Q function are updated by iterations. We set their initial values to 1.0 assuming that we have no ideas of the recognition ability of each base person re-id method and the difficulty of each gallery image. The prior probabilities $p(z_j)$ of $z_j, j = 1, \ldots, N$, which are used in (1) by E-step of the EM algorithm, are very important parameters for correctly inferring strong labels of the gallery images, therefore, we need to carefully define their values. Given a query image, if all the base person re-id methods $\boldsymbol{\Phi}_i, i = 1, \ldots, M$ rank a gallery image j highly in their rank lists, then there should be a very low probability that its strong label z_j equals zero, and vice verse. Generally speaking, the prior probability $p(z_j = 1)$ is highly related to the gallery image j's rank orders, and the rank consistency among the M rank lists generated by all the base person re-id methods. Based on this observation, we define $p(z_j)$ using the following equation:

$$p(z_j) = e^{-\left((\mu(\mathbf{x}_j)-1)^2+\lambda*\sigma(\mathbf{x}_j)\right)} \tag{12}$$

where $\mathbf{x}_j = \{x_{ij}\}_{i=1}^{M}$, and x_{ij} is the rank order given to gallery image j by the base person re-id method $\boldsymbol{\Phi}_i$, $\mu(\mathbf{x}_j)$ is the mean order, and $\sigma(\mathbf{x}_j)$ is the order variance of $\mathbf{x}_j$. Clearly, (12) gives high score to $p(z_j)$ if the mean order $\mu(\mathbf{x}_j)$ is close to one (i.e., gallery image j ranks highly in all the rank lists), and the order variance $\sigma(\mathbf{x}_j)$ is close to zero (i.e., the ranks of gallery image j among all the rank lists are highly consistent). We have observed from the experiments that the mean order $\mu(\mathbf{x}_j)$ contributes more to producing the correct strong label z_j, therefore we set smaller $\lambda(= 0.001)$ in our implementation.

IV. Experiments

A. Datasets

We use four popular person re-id benchmark datasets, *i*-LIDS, PRID2011, VIPeR, and CUHK01, for performance evaluations. All the datasets contain a set of persons, each of whom has several images captured by different cameras. The following is a brief description of these four datasets.

i-LIDS Dataset: It is constructed from video images shooting a busy airport arrival hall. It contains 479 images from 119 persons, which are normalized to 128×64 pixels. Each person has four images in average. These images are captured by nonoverlapping cameras, and are subject to large illumination changes and occlusions.

PRID2011 Dataset: Pedestrian images in this dataset were captured from two static surveillance cameras. These two cameras contain 385 and 749 different individuals, respectively, with 200 individuals appearing in both of these two cameras.

VIPeR Dataset: This dataset contain 1264 images of 632 different individuals. Each person has two images from two different camera views. The main challenges of this dataset are the huge variance in viewpoints, poses, and lighting conditions, and this dataset is one of the most widely used datasets for person re-id task.

CUHK01 Dataset: This is one relatively large dataset for the person re-id task. It contains 3884 images of 971 different individuals. Each person has four images captured from two different camera views, and each view with two images. Camera view A captures frontal or back views of a person while camera B captures the person's profile views.

B. Evaluation Protocol

We adopt the widely used single shot experimental setting for performance evaluations, which assumes that each individual has only one image in both the gallery and the query subsets in the test phase. For each benchmark dataset, all the individuals are randomly divided into two equal subsets so that the training and test sets contain half of available individuals with no overlap on person identities. In the test phase, the gallery set comprises only one image for each individual, and the rest person images are all left in the query set. Since the PRID2011 dataset contains 385 and 749 individuals in camera view A and B, respectively, and only 200 persons appear in both views, we randomly divide these 200 persons into two halves, and form the training and test sets using images of persons belonging to the first and second halves, respectively. During the test phase, the rest 649 images in camera view B are selected as the gallery images, and the camera view A images of the 100 persons in the test set are set as the query image. Therefore, there are 549 distracter images in the gallery set, which makes the person re-id performance on this dataset lower than those on other datasets. Finally, given a query image, each person re-id method outputs a rank list of all the gallery images in the dataset. If the rank list contains the matched image to the query image at kth position, then this query is considered as the right match of top k.

C. Evaluation Results

To evaluate our proposed fusion method, we have implemented six base person re-id methods: KISSME, kLFDA, kMFA, SiameseNet, TripletNet, and MultiParts. We choose these methods because their source codes are publicly available, and their methodologies belong to two different categories: 1) traditional approaches and 2) deep CNN-based approaches. By choosing methods from different categories, we want to explore complementary properties among these methods. For comparison purpose, we have also implemented three representative fusion methods which include the mean fusion (MeanFus), the majority voting (MajVote) [3], the adaptive weighted fusion (WFusion) [4], average distance fusion (AvgDisFus), and Sakrapee's feature-based reranking method (Sak's) [30]. With the MeanFus, we rank each gallery image by the mean order of all the base person re-id methods, while with the MajVote, we obtain the fusion results by voting with the binary labels generated by all the base

TABLE I
PERFORMANCE COMPARISONS BETWEEN THE PROPOSED FUSION METHOD AND BASE PERSON RE-ID METHODS ON PRID2011 AND I-LIDS DATASETS

PRID2011						i-LIDS					
Base Algo	Top1	Top10	Top20	Top50	Top100	Base Algo	Top1	Top5	Top10	Top15	Top20
KISSME	7.0	21.0	26.0	43.0	60.0	KISSME	28.3	55.7	68.9	77.3	83.4
kLFDA	10.0	27.0	34.0	47.0	65.0	kLFDA	33.7	59.3	71.7	80.2	86.5
kMFA	14.0	38.0	42.0	56.0	76.0	kMFA	38.0	65.1	77.4	84.4	89.2
SiameseNet	17.0	42.0	49.0	62.0	78.0	SiameseNet	41.3	67.3	80.3	85.1	89.2
TripletNet	19.0	33.0	43.0	57.0	74.0	TripletNet	51.6	73.9	83.4	91.4	94.6
MultiParts	22.0	43.0	55.0	67.0	78.0	MultiParts	60.4	82.7	90.7	93.4	95.1
Ours	**24.0**	**49.3**	**60.0**	**76.4**	**83.2**	Ours	**65.5**	**83.1**	**91.1**	**95.0**	**96.3**

TABLE II
PERFORMANCE COMPARISONS BETWEEN THE PROPOSED FUSION METHOD AND BASE PERSON RE-ID METHODS ON VIPER AND CUHK01 DATASETS

VIPeR						CUHK01					
Base Algo	Top1	Top5	Top10	Top15	Top20	Base Algo	Top1	Top5	Top10	Top15	Top20
KISSME	25.8	56.2	70.1	77.8	82.9	KISSME	27.0	52.0	71.3	72.7	77.7
kLFDA	31.2	65.8	79.6	86.7	90.6	kLFDA	29.3	61.1	74.6	84.7	88.0
kMFA	32.3	65.8	79.7	87.0	90.9	kMFA	30.7	62.3	75.5	84.9	88.2
SiameseNet	35.6	60.1	74.3	81.3	88.6	SiameseNet	37.0	62.8	78.7	85.3	87.7
TripletNet	37.3	62.3	76.3	81.6	87.3	TripletNet	47.3	72.1	81.3	84.0	87.6
MultiParts	43.8	69.5	79.7	81.0	90.2	MultiParts	53.7	81.3	91.0	93.3	95.3
Ours	**54.3**	**77.1**	**85.5**	**93.1**	**94.4**	Ours	**56.5**	**82.2**	**91.1**	**93.5**	**95.7**

TABLE III
PERFORMANCE COMPARISONS WITH OTHER FUSION METHODS ON PRID2011 AND ILIDS DATASETS

PRID2011						i-LIDS					
Algo	Top1	Top10	Top20	Top50	Top100	Algo	Top1	Top5	Top10	Top15	Top20
Sakrapee's[30]	20.9	−−	−−	−−	−−	Sakrapee's.[30]	61.2	−−	−−	−−	−−
MeanFus	17.0	39.0	49.0	66.0	83.0	MeanFus	57.0	74.4	84.9	91.4	94.1
AvgDisFus	17.3	39.2	49.8	67.5	83.1	AvgDisFus	57.4	74.1	85.2	92.4	94.7
wFusion[4]	21.0	45.0	53.1	74.2	81.3	wFusion[4]	62.8	81.0	90.4	93.7	94.5
MajVote[3]	16.0	36.0	45.0	64.0	78.0	MajVote[3]	57.4	73.7	84.7	90.4	93.1
Ours	**24.0**	**49.3**	**60.0**	**76.4**	**83.2**	Ours	**65.5**	**83.1**	**91.1**	**95.0**	**96.3**

methods, where the binary labels are generated in the same way as in our proposed method, the WFusion [4] method adaptively learns the weights for the base methods under a fusion framework to obtain improved performance, and the AvgDisFus computes the average distance between each gallery and query image obtained by different methods for sorting.

We used the proposed fusion method to fuse the above six person re-id methods, and evaluated the performances using the four datasets described in Section IV-A. Tables I and II show performance comparisons between the proposed fusion method and the six base person re-id methods on PRID2011 and *i*-LIDS, VIPeR, and CUHK01 datasets, respectively. We also used the four representative fusion approaches described above to fuse the same six person re-id methods, and compared their fusion results with those of our proposed method. The evaluation results are shown in Tables III and IV. Finally, Tables V and VI present the performance evaluation results of several state-of-the-art person re-id methods on the four benchmark dataset. For ease of comparisons, we have also included the performance scores of our proposed fusion method in these tables.

The performance evaluation results in the above tables can be summarized as follows.

1) Compared to all the six base person re-id methods, the proposed fusion method has achieved the best performances on all the four benchmark datasets, with respect to all the five evaluation metrics. With the Top1 error metric, the performance improvements range from 2% to 10.5% in comparison with the best base person re-id method.
2) Compared to the other four fusion methods, the proposed fusion method has also achieved the best performances on all the four datasets with all the five evaluation metrics. Our method is able to improve the performance accuracies by an average of 4% in comparison with the best performances achieved by one of these four fusion methods.
3) The proposed fusion method is superior to all the six state-of-the-art methods listed in Tables V and VI on all the four datasets, and with respect to all the five evaluation metrics. With the Top1 error metric, the performance improvements range from 2% to 6.5% in comparison with the best state-of-the-art method.

It is noteworthy that the performance accuracies achieved by some fusion methods, such as AvgFus and MajVote, are even worse than those achieved by some of the base person re-id methods on PRID2011, *i*-LIDS, and CUHK01

TABLE IV
PERFORMANCE COMPARISONS WITH OTHER FUSION METHODS ON VIPeR AND CUHK01 DATASETS

VIPeR						CUHK01					
Algo	Top1	Top5	Top10	Top15	Top20	Algo	Top1	Top5	Top10	Top15	Top20
Sakrapee's[30]	52.2	−−	−−	−−	−−	Sakrapee's[30]	54.5	−−	−−	−−	−−
MeanFus	49.8	74.9	84.4	91.4	93.6	MeanFus	51.7	79.7	88.6	91.3	93.1
AvgDisFus	51.4	75.1	84.3	92.2	94.1	AvgDisFus	51.3	79.4	89.7	91.1	93.0
wFusion[4]	51.7	74.3	83.2	90.8	93.4	wFusion[4]	53.2	80.7	88.9	91.4	93.3
MajVote[3]	47.8	71.5	82.7	92.0	94.2	MajVote[3]	50.7	78.3	87.0	90.3	93.3
Ours	**54.3**	**77.1**	**85.5**	**93.1**	**94.4**	Ours	**56.5**	**82.2**	**91.1**	**93.5**	**95.7**

TABLE V
PERFORMANCE COMPARISONS BETWEEN THE PROPOSED FUSION METHOD AND OTHER STATE-OF-THE-ART PERSON RE-ID METHODS ON PRID2011 AND I-LIDS DATASETS

PRID2011						i-LIDS					
Algo	Top1	Top10	Top20	Top50	Top100	Algo	Top1	Top5	Top10	Top15	Top20
Eim[17]	16.0	39.0	52.0	68.0	80.0	MFA[41]	38.0	65.1	77.4	84.4	89.2
Itml[8]	12.0	36.0	47.0	64.0	79.0	PCCA[41]	29.6	57.3	71.7	80.4	85.9
Mah[34]	16.0	41.0	51.0	64.0	76.0	PRDC[47]	37.8	63.7	75.1	82.8	88.4
Ding[9]	17.9	45.9	55.4	71.4	−−	Sak.[30]	50.3	−−	−−	−−	−−
Ska.[30]	17.9	−−	−−	−−	−−	Ding[9]	52.1	−−	−−	−−	−−
MultiParts[6]	22.0	47.0	55.0	76.0	83.0	MultiParts[6]	60.4	82.7	90.7	93.4	96.8
Ours	**24.0**	**49.3**	**60.0**	**76.4**	**83.2**	Ours	**65.5**	**83.1**	**91.2**	**95.0**	**97.1**

TABLE VI
PERFORMANCE COMPARISONS BETWEEN THE PROPOSED FUSION METHOD AND OTHER STATE-OF-THE-ART PERSON RE-ID METHODS ON VIPeR AND CUHK01 DATASETS

VIPeR						CUHK01					
Algo	Top1	Top5	Top10	Top15	Top20	Algo	Top1	Top5	Top10	Top15	Top20
Sal.[45]	30.2	52.3	66.0	73.4	79.2	Sal.[45]	28.5	46.3	57.2	64.1	−−
lmlf[46]	29.1	52.3	66.0	73.9	79.9	mFilter[46]	34.3	55.0	65.3	70.5	−−
Ding[9]	40.5	60.8	70.5	78.3	84.4	FPNN[25]	27.9	−−	−−	−−	−−
mFilter[46]	43.4	−−	−−	−−	−−	Ejaz[1]	47.5	−−	−−	−−	−−
Sak.[30]	45.9	−−	−−	−−	−−	Sak.[30]	53.4	76.4	84.4	−−	90.5
MultiParts[6]	47.8	74.7	84.8	89.2	91.1	MultiParts[6]	53.7	81.3	91.0	93.3	95.3
Ours	**54.3**	**77.1**	**85.5**	**93.1**	**94.4**	Ours	**56.5**	**82.2**	**91.1**	**93.5**	**95.7**

datasets (see Tables III and IV). The insight behind this phenomenon is that, the six base person re-id methods used in our performance evaluations have relatively large performance discrepancies, with the deep CNN-based methods having much higher performance accuracies than those traditional methods. When we fuse these methods using simple approaches, such as the average fusion and the MajVote which use the equal weights for all the base methods, bad performers with inferior performance accuracies will generate very negative effects to the fusion process, and drag the final fusion results even below those of good performers. In contrast, due to its method-aware and data-aware inference abilities, our proposed method automatically assigns higher weights to the base methods with better performance accuracies, and to the gallery images with higher labeling simplicities. These abilities have enabled the proposed method to produce better performances than any other existing individual person re-id and fusion methods. More importantly, the proposed fusion method can always make full use of the existing base person re-id methods, and obtain more complementary properties of these methods

D. Evaluations With All Possible Combinations

To better understand the characteristics of the proposed fusion method, we have used the proposed method to fuse different number of base models, and evaluated its performance accuracies using the VIPeR dataset as shown in Fig. 6. For a given number of base models $b \in \{1, 6\}$, we exhaustively enumerate all the possible combinations among the six base models, which is a total of $C_6^b = [6!/(b!(6-b)!)]$ combinations, and use the proposed fusion method to fuse each of them. The evaluation results are summarized in Fig. 4, where the horizontal and vertical axes correspond to the number of base models, and the Top1 matching accuracy, respectively. In this figure, each red line is the mean, and the blue box is the variance, and the black solid lines at the bottom and top of the blue box are the highest and lowest performance accuracies, respectively, of the fusion results for all the possible combinations of the given number of base models. The observations obtained from this experiment can be summarized as follows.

1) Fusing more base models tends to obtain better performance accuracies.
2) Fusing less base models tends to obtain unstable results with large variances.
3) There are large performance differences among different combinations of base models, especially when the number of base models is small. The near-bottom performances are obtained by fusing those bad performers among the base models, and vice verse, while the

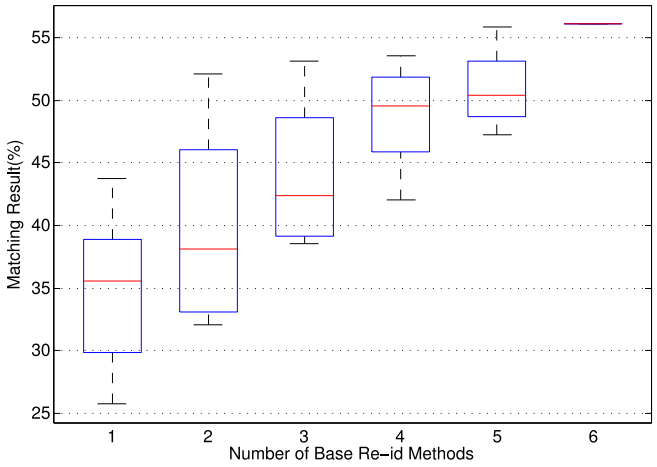


Fig. 4. Top 1 matching results on VIPeR dataset with different number of base fusion method.

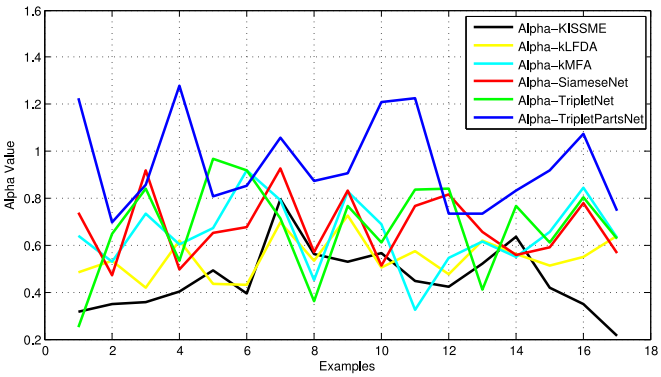


Fig. 5. Plots of inferred α_i's of the six base person re-id models. The horizontal axis denotes different query images, and the vertical axis shows the α_i's values.

middle-range performances are achieved by fusing those base models with large performance discrepancies.

Indeed, these observations are very much in line with our expectation, and can be easily explained by common sense.

E. Analysis of the Latent Parameters

The proposed fusion method is able to automatically infer the following two latent parameters: 1) the recognition ability α_i of each base person re-id method Φ_i and 2) the difficulty $(1/\beta_j)$ of each gallery image j to be correctly recognized. We studied the contributions of these latent parameters to the fusion results by setting one of them to value one and keeping the other intact (i.e., using the values inferred by the proposed method). It can be observed from Table VII that, setting all β_j's to one while keeping the inferred values for all α_i's gives better results than the opposite case, and that using the inferred values for both α_i's and β_j's produces the best outcomes. These observations manifest that α_i's play a more important role than β_j's to generate the fusion results.

To examine the correctness of the parameters α_i's, we have plotted them in Fig. 5, where the horizontal axis denotes

Fig. 6. Examples of the fusion result based on our referring algorithm.

different query images, and the vertical axis corresponds to the α_i's values of all the six base person re-id models. Interestingly, there is a strong consistency between the inferred parameter α_i and the recognition ability of the base model Φ_i. It can be observed from Fig. 5 that TriPNet has the highest α value, while KISSME has the lowest one. These two α values exactly match the person re-id performances of the two base models shown in Tables I and II. Therefore, it can be concluded that the parameter α_i inferred by our proposed fusion method does correspond to the recognition ability of the base person re-id model Φ_i.

TABLE VII
PERFORMANCES ON VIPeR DATASETS FOR PARAMETER ANALYSIS

Method	Top1	Top5	Top10	Top15	Top20
Only with α (β=1)	53.1	75.7	84.4	91.8	94.1
Only with β (α=1)	51.7	75.5	84.7	92.1	94.2
Ours	54.3	77.1	85.5	93.1	94.4

F. Merits of the Proposed Method

Based on all the above experiment results and method analysis, we have obtained the following merits of the proposed fusion method.

1) The proposed fusion approach for person re-id enables us to not only build upon but also make the best use of the existing, as well as future person re-id methods, to achieve much better re-id performance to satisfy the practical use.
2) When fusing multiple base re-id methods with various performance, the base methods with poor performance would probably have negative effects on the final decision by the fusion method. However, since the proposed fusion method considers the recognition ability of the base method, and the difficulty of the person image in the gallery set to be correctly recognized, we can better overcome this problem than some other fusion methods to obtain further improved re-id performance compared with all the base methods.

V. CONCLUSION

In this paper, we present an effective fusion method for person re-id. And we have proposed to tackle the person re-id as a rank fusion task, where only the predictions from the existing base re-id ranking results are offered and no ground truth information is required. We propose to used these base predictions to obtain the strong re-id results by fully making use of each base re-id model's strength, especially when these method complements each other well. During the fusion process, we define the recognition ability parameter for each re-id algorithm to measure its contribution to the input query image, and a labeling difficulty for each image in the gallery set. Then we adopt the GLAD model to simultaneously infer each re-id method's recognition accuracy, and the difficulty of each image in the gallery set, then finally infer the strong ranking results. The experiments on four public benchmark datasets have demonstrated that the proposed approach is superior compared with a number of state-of-the-art re-id methods. In the future work, we tend to extend this method to explore some other factors that may influence the reranking process and utilize more forthcoming re-id methods to further refine the re-id performance for satisfying the practical use.

REFERENCES

[1] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," in *Proc. CVPR*, vol. 5. Boston, MA, USA, 2015, pp. 3908–3916.

[2] S. Bak, E. Corvee, F. Brémond, and M. Thonnat, "Person re-identification using spatial covariance regions of human body parts," in *Proc. 7th IEEE Int. Conf. Adv. Video Signal Based Surveillance*, 2010, pp. 435–440.

[3] R. S. Boyer and J. S. Moore, *MJRTY—A Fast Majority Vote Algorithm*. Dordrecht, The Netherlands: Springer, 1991.

[4] X. Cao, Z. Tao, B. Zhang, H. Fu, and W. Feng, "Self-adaptively weighted co-saliency detection via rank constraint," *IEEE Trans. Image Process.*, vol. 23, no. 9, pp. 4175–4186, Sep. 2014.

[5] D. Chen, Z. Yuan, G. Hua, N. Zheng, and J. Wang, "Similarity learning on an explicit polynomial kernel feature map for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, USA, 2015, pp. 1565–1573.

[6] 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.*, Las Vegas, NV, USA, 2016, pp. 1335–1344.

[7] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino, "Custom pictorial structures for re-identification," in *Proc. BMVC*, vol. 1, 2011, pp. 1–11.

[8] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, "Information-theoretic metric learning," in *Proc. 24th Int. Conf. Mach. Learn.*, 2007, pp. 209–216.

[9] S. Ding, L. Lin, G. Wang, and H. Chao, "Deep feature learning with relative distance comparison for person re-identification," *Pattern Recognit.*, vol. 48, no. 10, pp. 2993–3003, 2015.

[10] P. Dollár, Z. Tu, H. Tao, and S. Belongie, "Feature mining for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–8.

[11] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 2360–2367.

[12] N. Gheissari, T. B. Sebastian, and R. Hartley, "Person reidentification using spatiotemporal appearance," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2, 2006, pp. 1528–1535.

[13] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. CVPR*, 2014, pp. 580–587.

[14] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *Computer Vision—ECCV*. Heidelberg, Germany: Springer, 2008, pp. 262–275.

[15] J. Han, D. Zhang, S. Wen, L. Guo, T. Liu, and X. Li, "Two-stage learning to predict human eye fixations via SDAEs," *IEEE Trans. Cybern.*, vol. 46, no. 2, pp. 487–498, Feb. 2016.

[16] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof, "Person re-identification by descriptive and discriminative classification," in *Image Analysis*. Heidelberg, Germany: Springer, 2011, pp. 91–102.

[17] M. Hirzer, P. M. Roth, and H. Bischof, "Person re-identification by efficient impostor-based metric learning," in *Proc. IEEE 9th Int. Conf. Adv. Video Signal Based Surveillance*, 2012, pp. 203–208.

[18] M. Hirzer, P. M. Roth, M. Köstinger, and H. Bischof, "Relaxed pairwise learned metric for person re-identification," in *Computer Vision—ECCV*. Heidelberg, Germany: Springer, 2012, pp. 780–793.

[19] W. Hu *et al.*, "Principal axis-based correspondence between multiple cameras for people tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 4, pp. 663–671, Apr. 2006.

[20] S. Khamis, C.-H. Kuo, V. K. Singh, V. D. Shet, and L. S. Davis, "Joint learning for attribute-consistent person re-identification," in *Proc. Comput. Vis. Workshops*, 2014, pp. 134–146.

[21] M. Köestinger, 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.*, 2012, pp. 2288–2295.

[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. NIPS*, 2012, pp. 1097–1105.

[23] W. Li and X. Wang, "Locally aligned feature transforms across views," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Portland, OR, USA, 2013, pp. 3594–3601.

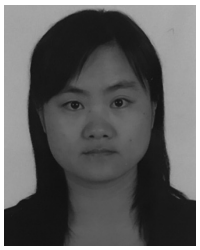
[24] W. Li, R. Zhao, and X. Wang, "Human reidentification with transferred metric learning," in *Proc. ACCV*, 2012, pp. 31–44.

[25] W. Li, R. Zhao, T. Xiao, and X. Wang, "DeepReID: Deep filter pairing neural network for person re-identification," in *Proc. CVPR*, 2014, pp. 152–159.

[26] 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.*, Boston, MA, USA, 2015, pp. 2197–2206.
[27] C. C. Loy, C. Liu, and S. Gong, "Person re-identification by manifold ranking," in *Proc. 20th IEEE Int. Conf. Image Process.*, 2013, pp. 3567–3571.
[28] B. Ma, Y. Su, and F. Jurie, "BiCov: A novel image representation for person re-identification and face verification," in *Proc. Brit. Mach. Vis. Conf.*, 2012, p. 11.
[29] A. Mignon and F. Jurie, "PCCA: A new approach for distance learning from sparse pairwise constraints," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 2666–2672.
[30] S. Paisitkriangkrai, C. Shen, and A. van den Hengel, "Learning to rank in person re-identification with metric ensembles," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, USA, 2015, pp. 1846–1855.
[31] U. Park, A. K. Jain, I. Kitahara, K. Kogure, and N. Hagita, "Vise: Visual search engine using multiple networked cameras," in *Proc. 18th Int. Conf. Pattern Recognit.*, vol. 3, 2006, pp. 1204–1207.
[32] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian, "Local fisher discriminant analysis for pedestrian re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Portland, OR, USA, 2013, pp. 3318–3325.
[33] R. Quan *et al.*, "Unsupervised salient object detection via inferring from imperfect saliency models," *IEEE Trans. Multimedia*, vol. 20, no. 5, pp. 1101–1112, May 2018.
[34] P. M. Roth, M. Hirzer, M. Köstinger, C. Beleznai, and H. Bischof, "Mahalanobis distance learning for person re-identification," in *Person Re-Identification*. London, U.K.: Springer, 2014, pp. 247–267.
[35] Y. Sun, Y. Chen, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," in *Proc. NIPS*, 2014, pp. 1988–1996.
[36] A. Toshev and C. Szegedy, "DeepPose: Human pose estimation via deep neural networks," in *Proc. CVPR*, 2014, pp. 1653–1660.
[37] X. Wang, G. Doretto, T. Sebastian, J. Rittscher, and P. Tu, "Shape and appearance context modeling," in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, 2007, pp. 1–8.
[38] 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, Dec. 2009.
[39] J. Whitehill, P. L. Ruvolo, T.-F. Wu, J. Bergsma, and J. R. Movellan, "Whose vote should count more: Optimal integration of labels from labelers of unknown expertise," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 2035–2043.
[40] Z. Wu, Y. Li, and R. J. Radke, "Viewpoint invariant human re-identification in camera networks using pose priors and subject-discriminative features," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 5, pp. 1095–1108, May 2015.
[41] F. Xiong, M. Gou, O. Camps, and M. Sznaier, "Person re-identification using kernel-based metric learning methods," in *Computer Vision–ECCV*. Cham, Switzerland: Springer, 2014, pp. 1–16.
[42] S. Yan *et al.*, "Graph embedding and extensions: A general framework for dimensionality reduction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 1, pp. 40–51, Jan. 2007.
[43] Y. Yang *et al.*, "Salient color names for person re-identification," in *Computer Vision—ECCV*. Cham, Switzerland: Springer, 2014, pp. 536–551.
[44] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Deep metric learning for person re-identification," in *Proc. ICPR*, 2014, pp. 34–39.
[45] R. Zhao, W. Ouyang, and X. Wang, "Person re-identification by salience matching," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 2528–2535.
[46] R. Zhao, W. Ouyang, and X. Wang, "Learning mid-level filters for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 144–151.
[47] W.-S. Zheng, S. Gong, and T. Xiang, "Person re-identification by probabilistic relative distance comparison," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 649–656.
[48] W.-S. Zheng, S. Gong, and T. Xiang, "Reidentification by relative distance comparison," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 3, pp. 653–668, Mar. 2013.

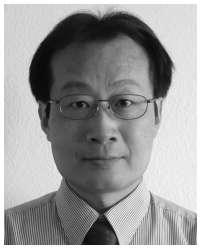
De Cheng received the B.S. degree in automation control from Xi'an Jiaotong University, Xi'an, China, in 2011, where he is currently pursuing the Ph.D. degree in artificial intelligence with the Institute of Artificial Intelligence and Robotics.

He visited Carnegie Mellon University, Pittsburgh, PA, USA, to do some cooperation research. His current research interests include pattern recognition, machine learning, and multimedia analysis.



Zhihui Li is currently pursuing the Ph.D. degree in artificial intelligence with the University of New South Wales, Sydney, NSW, Australia.

Her publications appear in proceedings of prestigious international conference like AAAI and IJCAI. Her current research interests include machine learning, data mining, and computer vision.



Yihong Gong (SM'10–F'18) received the B.S., M.S., and Ph.D. degrees in electronic engineering from the University of Tokyo, Tokyo, Japan, in 1987, 1989, and 1992, respectively.

In 1992, he joined Nanyang Technological University, Singapore, as an Assistant Professor with the School of Electrical and Electronic Engineering. From 1996 to 1998, he was a Project Scientist with the Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, USA. From 1999 to 2009, he was the Head of the Department of Information Analysis and Management, NEC Laboratories America, Inc., Cupertino, CA, USA. He is currently with Xi'an Jiaotong University, Xi'an, China, as a Professor. His current research interests include image and video analysis, multimedia database systems, and machine learning.

Dingwen Zhang received the B.E. degree in computer science from Northwestern Polytechnical University, Xi'an, China, in 2012, where he is currently pursuing the Ph.D. degree.

He visited Carnegie Mellon University, Pittsburgh, PA, USA, to do some cooperation research for two years. His current research interests include computer vision and multimedia processing, especially on saliency detection, co-saliency detection, and weakly supervised learning.