IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 31, 2021, accepted September 17, 2021, date of publication September 28, 2021, date of current version October 5, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3116064

A Few-Shot Learning Method Using Feature Reparameterization and Dual-Distance Metric Learning for Object Re-Identification

SHENG-HUNG FAN, MIN-HONG LIN, JUNG-YI JIANG, AND YAU-HWANG KUO, (Senior Member, IEEE)
Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan 70101, Taiwan
Corresponding author: Yau-Hwang Kuo (kuoyh@ismp.csie.ncku.edu.tw)

This work is supported in part by the Ministry of Science and Technology of Taiwan under Grants 109-2221-E-006-186-MY3.

ABSTRACT Many object re-identification (Re-ID) methods that depend on large-scale training datasets have been proposed in recent years. However, the performance of these methods degrades dramatically when insufficient training data are available. To address this challenging problem, we propose a few-shot object re-identification (FSOR) method that enhances the generalization and discrimination abilities of object Re-ID models trained on small datasets. This method applies two novel techniques: reparameterization for feature vectors and dual-distance metric learning. The reparameterization mechanism transforms the primary feature vector of each input image into a Gaussian distribution to enhance the robustness of the FSOR method when performing object Re-ID tasks. The dual-distance metric learning technique, called H&C learning, considers both the hard mining distance and the center-point distance between each query sample and each support set of different object identities. H&C learning extracts the characteristics of the entire training dataset more precisely than other approaches and thus improves the discriminative abilities of object Re-ID models. Extensive experiments on both person and vehicle Re-ID datasets, such as Market-1501, DukeMTMC-ReID, CUHK03, and VeRi-776, show that the FSOR method has improved performance and outperforms state-of-the-art methods when the amount of labeled training data is small.

INDEX TERMS Object re-identification, few-shot learning, metric learning, visual recognition.

I. INTRODUCTION

As the demand for intelligent video surveillance has increased, object Re-identification (Re-ID), which retrieves an object of interest from a large image gallery dataset across multiple nonoverlapping cameras, has become an important computer vision task. This task is challenging due to different camera viewpoints [1], varying image resolutions [2], illumination changes, unconstrained poses [3], image occlusion, and significant background changes. Generally, building an object Re-ID system for a specific scenario requires five main steps [4]. The data collection step involves the collection of video data from multiple nonoverlapping cameras, but these raw data are likely to contain considerably complex and noisy background clutter. The object extraction step extracts object bounding boxes from the collected video data through an object detection or tracking algorithm. The data annotation step labels the extracted object images; this is usually a time-consuming process. The model training step constructs a discriminative and robust Re-ID model using the annotated object images. The object retrieval step generates a ranked list of object images from a large-scale gallery dataset for a given query regarding an object of interest by sorting based on the similarity between the query image and each gallery image. Note that the data annotation and model training steps are invoked only during the learning phase.

Most existing object Re-ID models depend on large-scale labeled training data to learn how to distinguish between objects with different identities. Obtaining this large amount of labeled training data requires tedious data collection and time-consuming annotation processes, which lead to poor scalability in real-world Re-ID applications. It is challenging to annotate the identities of objects in a large-scale cross-camera dataset because many similar objects exist in the dataset and because indistinguishable object images are captured by the cameras under varying conditions. These factors

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

make scaling a Re-ID system into a large camera network difficult. Because the majority of Re-ID datasets provide few images for each individual object, deep learning-based models usually suffer from a lack of training data and performance degradation resulting from overfitting. Furthermore, some object IDs in the testing dataset likely will not appear in the training dataset in a few training data scenarios. Therefore, a method for training object Re-ID models should possess excellent generalizability, allowing the models to identify object IDs that are not included in the training dataset. To tackle these problems, one intuitive approach is to use transfer learning [5] to retrain an existing model for a different application with a new dataset. However, when applying transfer learning for object Re-ID, the constructed model can still easily overfit the new training data. Another approach is to use a few-shot learning scheme [6] that can rapidly generalize the trained model from only a few labeled samples for each target object class. However, an effective few-shot object Re-ID model must have sufficient generalization and discrimination abilities. Good generalizability can allow the model to avoid the overfitting problem when utilizing limited training data and enables the Re-ID model to identify objects that do not appear in the training set. Good discriminability allows the Re-ID model to learn discriminative features and handle drastic viewpoint changes with few training data.

The proposed few-shot object re-identification (FSOR) method is a few-shot learning approach that applies some novel techniques to construct object Re-ID models with superior generalization and discrimination abilities. First, a reparameterization approach, which is derived from the concept of the reparameterization trick proposed for variational auto-encoders (VAEs) [7], is adopted to transfer the primary feature vector of each input image extracted by a convolutional neural network (CNN) (a ResNet-50 model) into a Gaussian distribution. This approach avoids overfitting with the constructed Re-ID model and enhances its generalizability when re-identifying nontrained objects. Second, we propose a dual-distance metric learning approach that evaluates both the hard-sample distance and the center-point distance between the support dataset of each object identity and the query sample. This approach, called H&C learning, is useful for enhancing the discrimination ability of the constructed Re-ID model. In addition, we apply two data augmentation approaches to increase the richness and diversity of the training data for FSOR. Padding and random crop approaches are used to make the trained Re-ID model more adaptable to the positions of identified objects in the images. A random erasing approach [8] is used to increase the robustness of the trained Re-ID model for image object occlusion. We also adopt some training tricks to improve the learning results. A warming-up learning rate strategy [9] is adopted to bootstrap the FSOR model by dynamically changing the learning rate during training. A label smoothing strategy [10] is used to prevent the Re-ID model from overfitting the object IDs in the training dataset by changing the prediction logits term in the ID loss to reduce the weight of the ground truth label's ID prediction logits, q_i, which is the output of the neural network, can be computed by $\exp(p_i) / \sum_{l=1}^{K} \exp(p_l)$, where p_i denotes the predicted score for class i and K denotes the number of labels.

In summary, the main contributions of this paper include the following:

- The FSOR method can efficiently construct object Re-ID models without tedious data collection and time-consuming annotation processes.
- FSOR guarantees model discrimination and generalization abilities when performing object Re-ID tasks through a novel few-shot learning model that includes two performance-improving mechanisms: a reparameterization mechanism that causes the FSOR approach to be more adaptive to re-identification for objects not included in the training data and the H&C metric learning mechanism, which makes the FSOR method more discriminative.
- The superior performance of FSOR is confirmed with various widely-used person Re-ID and vehicle Re-ID datasets and a comparison with some state-of-the-art methods.

II. RELATED WORKS

Recent studies on object Re-ID have mostly focused on deep CNNs, which learn the identity-discriminative features of object images. The most commonly used feature representation methods can be classified into global and local representation schemes [11], [12]. Global schemes extract features that represent entire object images, while local schemes extract features that represent critical parts of object images.

The deep learning methods that use global features for object Re-ID can be roughly divided into two main types according to the loss functions that they use: classification loss or metric loss functions [13]. When using the classification loss function, a Re-ID model is trained with the same object identities as those in the image classification task. For example, ID-discriminative embedding (IDE) [14] combines both an identity model and a verification model to train a Re-ID model. CamStyle [15] uses a generative model to perform data augmentation, which changes the image style between different cameras. It also uses classification loss for the Re-ID model. Several learning methods that use metric loss functions have also been proposed to construct object Re-ID models. Wang *et al.* [16] proposed a network model to extract feature maps with multiple scales from different stages of the backbone network and utilized the acquired feature maps to obtain an advanced result. TriNet [17] proposed a batch selection method for hard triplet samples to train a person Re-ID model according to the triplet loss. This type of loss function learns the relationships between triplet samples, including an anchor sample, positive sample and negative sample, from a distance function that measures the similarity between a pair of samples [18]. In [19], a robust person Re-ID model was learned with a Fast-Approximated Triplet (FAT) loss that converts a point-wise triplet loss into a

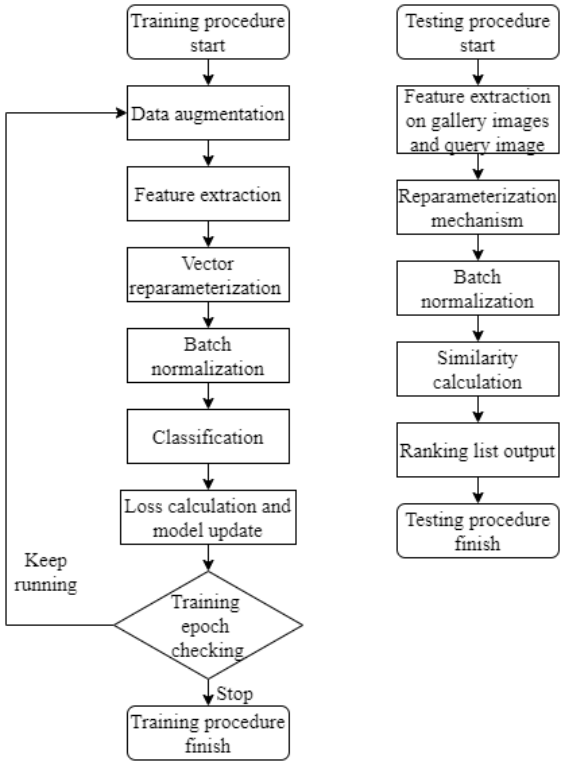

FIGURE 1. The flow chart of the proposed method.

point-to-set form. The Deep Meta Metric Learning (DMML) method [20] evaluates the hard mining distances of hard samples in each identity to construct an object Re-ID model. The BagTricks [21], [22] model was proposed, which combines a classification loss and deep metric loss to achieve better performance. The LiftedStructure [23] is a method that lifts the vector of pairwise distances within the batch to the matrix of pairwise distances. It helps to learn the state-of-the-art feature embedding by optimizing a novel structured prediction objective on the lifted problem. More recently, Sun *et. al* [24] proposed a circle loss to maximize the within-class similarity and minimize the between-class similarity. Proxy anchor [25] combines the advantages of pair-based and proxy-based loss. It can boost the speed of convergence and is robust against noisy labels and outliers. Khosla *et. al* [26] proposed a fully-supervised contrastive method to effectively leverage label information. The asymmetric weighted logistic metric learning (AWLML) [27] constructs a logistic metric-learning approach that uses an objective function with a positive semidefinite constraint to learn the metric matrix from a set of labeled samples. Then, an asymmetric weighted strategy is adopted to solve the unbalance problem between the number of target and background samples.

To avoid the need for a large, labeled training dataset, Xin *et al.* [28] proposed a self-paced multi-view clustering (SPMVC) method, which is a semi-supervised person Re-ID model trained with a small amount of labeled data and a large amount of unlabeled data. SPMVC performs the object Re-ID task using a heterogeneous set of CNNs

TABLE 1. The type of method, main idea, and research gap of related works.

Method	Type of method	Main idea	Research gap
Contrastive[40] Triplet[18] FAT[19] N-pair[41] Sturectured[23] SupCon[26] BagTricks[21]	Pairwise cost method	To develop an optimization algorithm based on the cost of pairwise distance.	In the few shot scenario, this type of method leads to the overfitting problem.
CircleLoss[24] ProxyAnchor[25] DMML[20]	Distribution method	To learn the whole distribution rather than the cost information of pairwise distance.	In the few shot scenario, the lack of data makes this type of method hard to learn the whole distribution.

initialized by the labeled training samples. Then, these models assign pseudo labels to the unlabeled training data step by step to further fine-tune all the constructed CNNs together with the original labeled training samples. In contrast, few-shot Re-ID models use only a labeled dataset with small amount of data for training. Few-shot learning [6] aims to enhance model generalizability and avoid overfitting while retaining a good discriminative capability. The existing few-shot learning methods can be roughly divided into model-based, metric-based, and optimization-based categories. For example, the memory-augmented neural network model [29], which uses external memory as short-term memory and slowly updated weights as long-term memory, is a model-based method. This model learns strategies for storing expressions in memory and learns how to use these expressions to make predictions. Metric-based methods learn the relationships between samples of different object classes by training an end-to-end few-shot classifier with a nonparametric scheme. In contrast, a parametric scheme must optimize tens of thousands of parameters in the neural network classifier; therefore, it will almost certainly overfit in situations with few data samples. Matching networks [30], prototypical networks [31] and relation networks [32] are some examples of metric-based methods. Unlike conventional transfer learning, optimization-based methods learn a beneficial common initialization for transfer learning, such as model-agnostic meta-learning (MAML) [33]. In summary, Table 1 lists the type of method, main idea, and research gap of existing methods. Our method is proposed to overcome the research gaps on the overfitting and discrimination abilities in existing methods.

III. LEARNING FRAMEWORK FOR FEW-SHOT OBJECT RE-IDENTIFICATION (FSOR)

As mentioned in sections I and II, there are several problems with object Re-ID in cases with few labeled data. When the amount of training data decreases, the model overfits these few data, resulting in low generalization ability. In addition, most existing object Re-ID models suffer from low discrimination ability. Our FSOR method aims to solve these

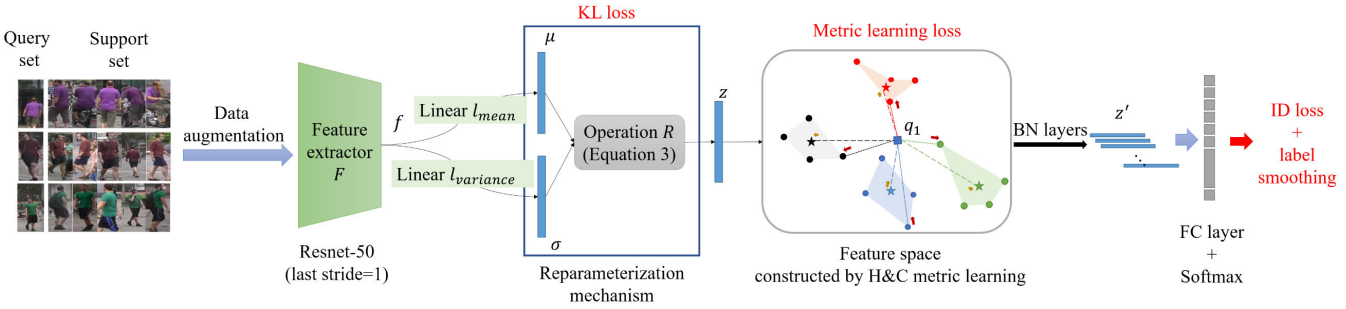

FIGURE 2. The learning framework of the proposed FSOR model.

problems to create object Re-ID models with superior generalization and discrimination abilities. The whole flow chart of FSOR is illustrated in Figure 1. The learning framework of FSOR is illustrated in Figure 2, and its detailed learning procedure is illustrated in Figure 3. The input images are first augmented using random erasing, padding, and random crop techniques. Next, the backbone network (ResNet-50 in this case) extracts primary feature vectors, and then a reparameterization mechanism is used to transform these primary feature vectors so that they conform to a Gaussian distribution. This process forms a continuous feature space that allows the object Re-ID model to be more generalizable and cover object images not included in the training dataset. During the learning phase, the training samples in each batch are divided into a query set and support sets containing different object identities. The H&C metric learning mechanism is first invoked to acquire the relationships between each query sample and the support sets. This learning mechanism possesses the ability to make the feature vectors of objects with the same identity closer while making those with different identities further apart. Then, a batch normalization layer is used to separate the feature vectors used for the metric loss and classification loss (ID loss) [22] because they are inconsistent in a single embedding space. The batch normalization layer optimizes these two losses in two different embedding spaces. Finally, a fully connected layer with a softmax function is implemented as a classifier to learn the association between each sample and its identity. To improve the learning efficiency of the model, the ID loss with label smoothing is used to predict the identity of each image. The KL loss, metric learning loss, and ID loss are all referred to when fine-tuning the parameters of the feature extractor F and the linear layers l_{mean} and $l_{variance}$. However, when fine-tuning the parameters of the fully connected layer, only the ID loss is referred to. The testing procedure will be introduced in part IV.

A. FEATURE VECTOR REPARAMETERIZATION

To enhance the generalization ability of the model, we use the concept of reparameterization trick proposed in variational auto-encoders (VAEs) [7], which force the feature distribution to follow the normal distribution, to make the model adapt to the data not seen in the training set. As shown

Learning procedure of FSOR for a batch of training data

Input:
A training data set D_{train}, which consists of a query set Q_{train} and a support set S_{train}, with a batch size of bs

Preprocessing:
Conduct data augmentation on D_{train} with the random erasing, padding and random crop techniques to obtain D'_{train} which consists of an augmented query set Q'_{train} and an augmented support set S'_{train}.

The overall FSOR procedure:
Execute the following operations for each $I \in D'_{train}$:

1. Send I to the backbone network F (ResNet-50) to extract the primary feature vector $\boldsymbol{f}$.
2. Send $\boldsymbol{f}$ to the linear layer l_{mean} to obtain its mean vector $\boldsymbol{\mu}$ according to Equation (1).
3. Send $\boldsymbol{f}$ to the linear layer $l_{variance}$ to obtain its variance vector $\boldsymbol{\sigma}$ according to Equation (2).
4. Send $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ to the reparameterization mechanism to obtain a new feature vector $\boldsymbol{z}$ according to Operation R (Equation (3)).
5. Send $\boldsymbol{z}$ to the batch normalization layer to acquire a new feature vector $\boldsymbol{z}'$.
6. Send $\boldsymbol{z}'$ to the fully connected layer and a softmax function to acquire the probability vector $\boldsymbol{p}$.

Execute the following operations for all generated $\boldsymbol{z}'$ vectors, which consist of $\boldsymbol{z}'_q$ query vectors and $\boldsymbol{z}'_s$ support vectors:

1. Calculate the KL divergence loss for each $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ according to Equation (4a), and then aggregate the calculated KL divergence loss of the entire batch using Equation (4b).
2. Remove the outliers from the support sets using Equation (11).
3. Calculate the H&C metric loss for each $\boldsymbol{z}'_q$ with respect to all $\boldsymbol{z}'_s$ vectors according to Equation (10).
4. Calculate the ID loss for each $\boldsymbol{p}$ to verify its ID using cross-entropy loss with label smoothing.
5. Update the parameters of the FSOR model based on the KL loss, H&C metric learning loss, and ID loss values.

FIGURE 3. The detailed learning procedure of FSOR.

in Figure 2, the FSOR method uses a ResNet-50 [34] model pretrained with ImageNet as the backbone network for feature extraction. This network receives a 256 × 128 input image and outputs a 2048-dimensional feature vector. To enrich the feature granularity, the last spatial down-sampling operation of the ResNet-50 backbone is removed, that is, the last stride is reduced from 2 to 1. This removal increases the spatial resolution of each feature map from 8 × 4 to 16 × 8. For each feature vector $\boldsymbol{f}$ acquired from the ResNet-50 backbone, the proposed reparameterization mechanism (R) invokes two independent linear layers to generate the $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ vectors

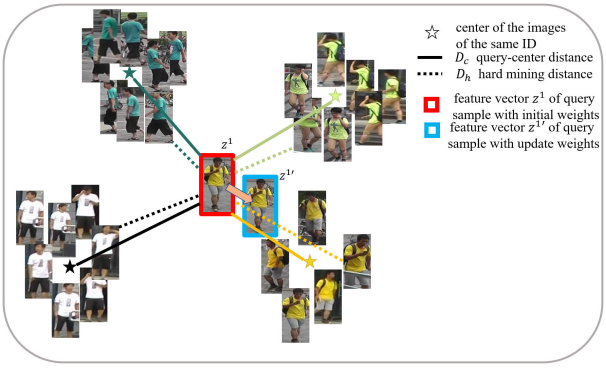

FIGURE 4. The H&C metric learning method. The query-center distance is useful for conforming to the property of the whole support set, and the hard mining distance is useful for making the query sample closer to the "hard" sample, which can improve the convergence speed of the model.

using (1) and (2), respectively:

$$\boldsymbol{\mu} = l_{mean}(\boldsymbol{f}) = \mathrm{w}_{\mu}\boldsymbol{f} + b_{\mu}, \tag{1}$$
$$\boldsymbol{\sigma} = l_{variance}(\boldsymbol{f}) = \mathrm{w}_{\sigma}\boldsymbol{f} + b_{\sigma}, \tag{2}$$

where w_{μ} and b_{μ} are the trainable parameter and bias of the linear layer l_{mean} used to generate $\boldsymbol{\mu}$, respectively, while w_{σ} and b_{σ} denote the trainable parameter and bias of the linear layer $l_{variance}$ used to generate $\boldsymbol{\sigma}$, respectively. Using $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$ and an additional noise vector $\boldsymbol{v}$ sampled from a normal distribution, a new feature vector $\boldsymbol{z}$ is generated by (3):

$$R(\boldsymbol{f}) = \boldsymbol{z} = \boldsymbol{\mu} + exp(\boldsymbol{\sigma}) \times \boldsymbol{v}, \tag{3}$$

where $\boldsymbol{\mu}$ and $\exp(\boldsymbol{\sigma})$ denote the mean and variance of a Gaussian distribution devoted to $\boldsymbol{f}$, respectively. The exponential operation of $\boldsymbol{\sigma}$ is invoked to ensure a non-negative variance. All the $\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{v}$, and $\boldsymbol{z}$ vectors have 2,048 dimensions. To make all the $\boldsymbol{z}$ vectors conform to a Gaussian distribution, we attempt to minimize the Kullback-Leibler (KL) divergence between the $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ vectors according to the KL divergence loss L_{KL}, also used in VAEs, as shown in (4a) and (4b),

$$L_{KL}^{i} = -\frac{1}{2}\sum_{j=1}^{J}\left(\left(1+\boldsymbol{\sigma}_{\boldsymbol{ij}}\right) - \left(\boldsymbol{\mu}_{\boldsymbol{ij}}\right)^2 - exp\left(\boldsymbol{\sigma}_{\boldsymbol{ij}}\right)\right), \tag{4a}$$
$$L_{KL} = \frac{1}{bs}\sum_{i=1}^{bs} L_{KL}^{i}, \tag{4b}$$

where bs denotes the number of training samples and J is the dimensionality of $\boldsymbol{z}$, $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$.

B. H&C METRIC LEARNING

To enhance the discriminative ability of the FSOR method on the object Re-ID task, we develop the H&C metric learning method, which learns a distance metric that can precisely determine the similarities between objects. This method uses the negative log likelihood (NLL) loss, which simultaneously evaluates both the hard mining distance and the query-center distance.

1) HARD MINING DISTANCE

The hard mining distance is used to find hard samples in each batch to produce substantial gradients from very few data points. Using hard samples rather than randomly selected samples for model training can speed up the convergence speed as mentioned in [20]. This is because the model can obtain more useful information and be guided to put in more effort to efficiently reduce the loss value when it is trained with hard samples. Thus, the convergence speed of model learning can be significantly accelerated. For this reason, the hard sample mining is an important aspect of several metric learning methods. For each query sample, the furthest support sample with the same ID is defined as a positive hard sample, while the closest support samples with different IDs are defined as negative hard samples. Then, the hard mining distance between a query sample and the support set of a specific ID $D_h\left(q_j^m, S^n\right)$ can be computed as follows:

$$D_h\left(q_j^m, S^n\right) = \begin{cases} \max\limits_{i}\left(dis\left(q_j^m, s_i^n\right)\right), & \text{for } m = n \\ \min\limits_{i}\left(dis\left(q_j^m, s_i^n\right)\right), & \text{for } m \neq n, \end{cases} \tag{5}$$

where q_j^m denotes the query sample with ID m, S^n denotes the support set of ID n, $s_i^n \in S^n$, n denotes the ID index, and i denotes the index of a support sample in S^n. The distance function $dis(x, y)$ estimates the Euclidean distance between feature vectors x and y as follows:

$$dis(x, y) = d\left(\boldsymbol{z}_{\boldsymbol{x}}, \boldsymbol{z}_{\boldsymbol{y}}\right) = d\left(R\left(F\left(x\right)\right), R\left(F\left(y\right)\right)\right), \tag{6}$$

where R is the function in (3) and F is a trainable ResNet-50 feature extractor.

2) QUERY-CENTER DISTANCE

The query-center distance is a set-based distance from the query sample to the center point of the support set of a specific ID. Because the center point represents the properties of the entire support set, the query-center distance is useful for learning the overall relationship between a query sample and the support set for a specific ID. We define the mean of all the samples in the support set as its center point; this approach regards each support sample as having the same influence on the query sample.

The query-center distance $D_c\left(q_j^m, S^n\right)$ is computed as follows:

$$D_c\left(q_j^m, S^n\right) = d(R(F\left(q_j^m\right)), \boldsymbol{c^n}) \tag{7}$$

where $\boldsymbol{c^n}$ denotes the center feature vector of ID n in each batch and is computed as $\boldsymbol{c^n} = \frac{1}{M}\sum_{i=1}^{M}\mathrm{R}(\mathrm{F}(s_i^n))$ where M denotes the amount of images of ID n in each batch.

In both the hard mining distance and query-center distance schemes, a query sample is assigned the same ID as that of the point closest to it.

3) DUAL-DISTANCE METRIC LEARNING

The H&C metric learning method combines the advantages of the hard mining distance and query-center distance, as shown

Inference Algorithm of FSOR

Input:

A query image I_q, and a set of gallery images GA (I_g^i denotes the i-th gallery image of GA)

Transformation of I_q and each I_g^i: (denote I_q or I_g^i as the input I)

1. Send I to the backbone network F to extract its primary feature vector $\boldsymbol{f}$.
2. Send $\boldsymbol{f}$ to the linear layers l_{mean} and $l_{variance}$ to obtain its mean vector $\boldsymbol{\mu}$ and variance vector $\boldsymbol{\sigma}$ according to Equations (1) and (2), respectively.
3. Send $(\boldsymbol{\mu}, \boldsymbol{\sigma})$ to the reparameterization mechanism to obtain a new feature vector $\boldsymbol{z}$ according to Operation R (Equation (3)).
4. Send z to the batch normalization layer to obtain its $\boldsymbol{z}'$ (denote $\boldsymbol{z}'$ as $\boldsymbol{z_q}'$ for input I_q or $\boldsymbol{z_g^{i\prime}}$ for input I_g^i).

Generate a ranked list of gallery images for the query image I_q:

1. Calculate the Euclidean distance between $\boldsymbol{z_q}'$ and each $\boldsymbol{z_g^i}'$.
2. Sort the evaluated gallery images by the Euclidean distances calculated in the previous step in increasing order; the resulting list is the output of the inference process.

FIGURE 5. The detailed inference procedure of FSOR.

in Figure 4. The hard mining distance improves the convergence speed, and the query-center distance learns the overall relationships between the query samples and the support sets.

The hard mining distance learns from the hard samples to distinguish similar samples with different IDs or dissimilar samples with the same ID. However, the hard samples cannot represent the properties of the entire set of support samples, so this approach might ignore the influence of other samples in the support set. To address this problem, the H&C method simultaneously learns the overall effect of all the support samples using the query-center distance and learns from the extreme samples based on the hard mining distance. The loss functions for learning with the hard mining distance L_{HM} and the query-center distance L_{CM} are as follows:

$$\boldsymbol{L_{HM}} = \sum_{j=1}^{n_q}\sum_{n=1}^{N}\boldsymbol{\alpha_n}\left(-log\frac{-\boldsymbol{D_h}\left(\boldsymbol{q_j^m}, \boldsymbol{S^n}\right)}{\sum_{n'=1}^{N}-\boldsymbol{D_h}\left(\boldsymbol{q_j^m}, \boldsymbol{S^{n'}}\right)}\right),$$
$$\boldsymbol{\alpha_n} = \begin{cases}\boldsymbol{0}, & \boldsymbol{n \neq m}\\ \boldsymbol{1}, & \boldsymbol{n = m},\end{cases} \tag{8}$$

$$\boldsymbol{L_{CM}} = \sum_{j=1}^{n_q}\sum_{n=1}^{N}\boldsymbol{\alpha_n}\left(-log\frac{-\boldsymbol{D_c}\left(\boldsymbol{q_j^m}, \boldsymbol{S^n}\right)}{\sum_{n'=1}^{N}-\boldsymbol{D_c}\left(\boldsymbol{q_j^m}, \boldsymbol{S^{n'}}\right)}\right),$$
$$\boldsymbol{\alpha_n} = \begin{cases}\dfrac{\boldsymbol{\varepsilon}}{\boldsymbol{N}}, & \boldsymbol{n \neq m}\\ \boldsymbol{1}-\dfrac{\boldsymbol{N-1}}{\boldsymbol{N}}\boldsymbol{\varepsilon}, & \boldsymbol{n = m},\end{cases} \tag{9}$$

where n_q is the number of query samples, q_j^m denotes the j^{th} one among the n_q query samples with ID m, N is the number of IDs in each batch, and n denotes the ID index form 1 to N. In addition, ε is a hyper-parameter in the query-center distance-based label smoothing process to prevent overfitting when learning with the query-center distance. Label smoothing is not necessary for the hard mining distance process because the hard samples are clearly distinguished from the query samples.

In summary, the total loss of the H&C metric learning function L_{ML} is as follows:

$$L_{ML} = \lambda_{hm}L_{HM} + \lambda_{cm}L_{CM}, \tag{10}$$

where λ_{hm} and λ_{cm} are hyper-parameters.

In addition, we remove the outliers from each support set to avoid noise samples being selected as hard samples, which might lead to an incorrect gradient. We calculate the mean and standard deviation of the distances between all support samples and the center point. Then, each support samples whose distance from the central point is greater than a threshold, as formulated in (11), is ignored when selecting the hard samples:

$$dis\left(s_i^n, \boldsymbol{c^n}\right) > mean(R\left(F\left(S^n\right)\right))+\delta \times std(R(F\left(S^n\right))), \tag{11}$$

where $dis\,(x, y)$ is the distance function in (5) and δ is a hyper-parameter.

C. OVERALL TRAINING LOSS

During FSOR training, the hybrid loss function is formulated as follows:

$$L_{total} = \lambda_{kl}L_{KL} + \lambda_{ml}L_{ML} + \lambda_{id}L_{ID}, \tag{12}$$

where λ_{kl}, λ_{ml} and λ_{id} are the hyper-parameters denoting the weights of the three partial losses. In (12), the KL divergence loss (L_{KL}) in the reparameterization approach is utilized to make the extracted feature vectors conform to a Gaussian distribution. The H&C metric learning loss (L_{ML}) allows the FSOR method to learn a precise distance measure between two feature vectors. Finally, the label smoothing ID loss (L_{ID}) acts as a classification loss for the FSOR model.

IV. THE FSOR INFERENCE PROCESS

Figure 5 illustrates the detailed inference procedure of FSOR, and Figure 6 depicts the inference process of FSOR, which retrieves and sorts the input gallery images according to their similarity scores with respect to the query image. During the inference phase, the feature vectors of the query image and all the gallery images are extracted and reparametrized by the same mechanism used during the learning phase. Then, the similarity between the query image and each gallery image is estimated by the Euclidean distance between their feature vectors, as formulated in (6). Finally, a ranked list of candidate gallery images is obtained according to their similarity scores with respect to the query image. In Figure 6, the query image is shown in the red box, and the gallery images that have the same ID as that of the query image are shown in the blue boxes.

V. EXPERIMENTAL RESULTS AND ANALYSIS

In this section, the performance of FSOR is evaluated for both person and vehicle Re-ID. In addition, we embed the FSOR method into two existing person Re-ID models to

TABLE 2. The details of the four Re-ID datasets used to evaluate the performance of the FSOR approach.

	Market-1501	DukeMTMC-ReID	CUHK03-labeled	VeRi-776
Number of training IDs	751	702	767	576
Number of testing IDs	750	702	700	200
Number of training images	12,936	16,522	7,368	37,778
Number of query images	3,368	2,228	1,400	1,678
Number of gallery images	19,732	17,661	5,328	11,579
Number of cameras	6	8	2	20

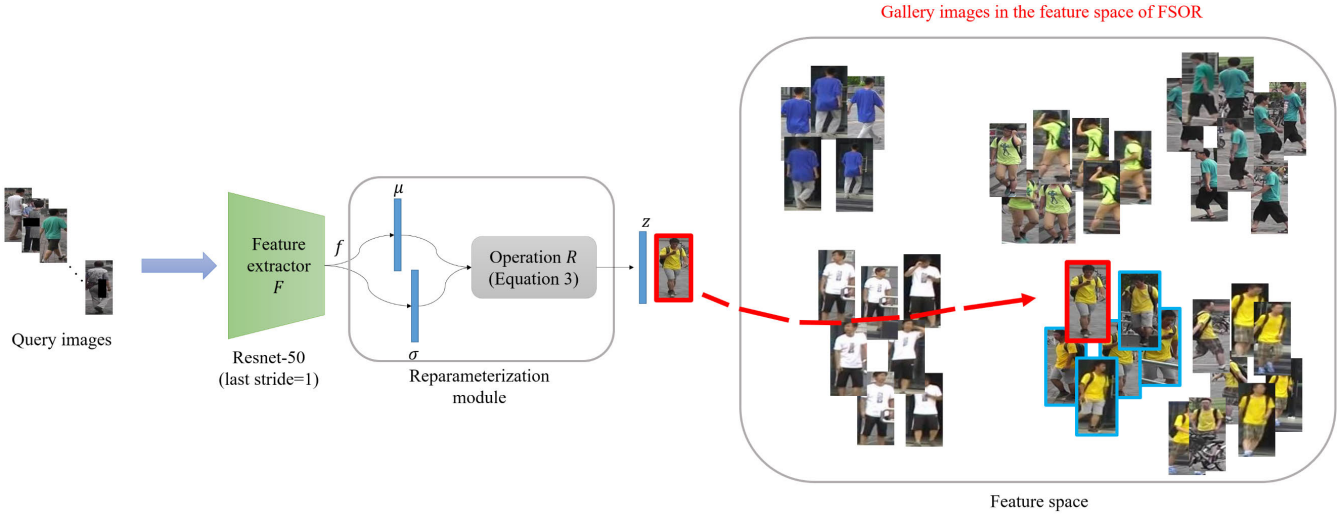

FIGURE 6. The inference process of FSOR.

demonstrate its ability to improve the performance of existing Re-ID models in situations with insufficient training data.

A. DATASET

These experiments use three datasets for person Re-ID and one dataset for vehicle Re-ID. Table 2 lists detailed information regarding these four datasets. Figure 7 shows some examples of training samples, which are selected from the Market-1501, DukeMTMC-ReID, CUHK03 and VeRi-776 datasets.

Market-1501 [35] is a person Re-ID dataset that contains 32,668 images of 1,501 identities captured by 6 cameras. It is divided into a training set with 12,936 images of 751 identities and a testing set with 19,732 gallery images of 750 identities and 3,368 hand-drawn query images of 750 identities.

DukeMTMC-ReID (Duke Multi-Tracking Multi-Camera ReIDentification) [36] is a subset of the DukeMTMC dataset for image-based person Re-ID. This dataset contains 34,183 images of 1,404 identities captured by 8 cameras. These images are divided into a training set with 16,522 images of 702 identities and a testing set with 17,661 gallery images of 702 identities and 2,228 hand-drawn query images of 702 identities.

CUHK03 (Chinese University of Hong Kong Re-identification) [37] is a person Re-ID dataset derived from two camera viewpoints. We use the CUHK03-labeled set in our experiments; it contains 12,696 images of 1,467 identities. The dataset is divided into a training set with 7368 images of 767 identities and a testing set with 5,328 gallery images of 700 identities and 1,400 hand-drawn query images of 700 identities.

VeRi-776 (Vehicle Re-identification) [38] is a vehicle Re-ID dataset covering a 1.0 km^2 area over 24 hours. Each vehicle is captured by 2∼18 cameras with different viewpoints, illumination conditions, resolutions, and occlusions. The VeRi-776 dataset contains 49,357 images of 776 different vehicles captured by 20 cameras. It is divided into a training set with 37,778 images of 576 vehicles and a testing set with 11,579 gallery images and 1,678 query images of 200 vehicles.

B. EVALUATION METRICS

As with most existing Re-ID methods, for our experiments with FSOR, we adopt two popular performance evaluation metrics: the cumulative matching characteristic (CMC) curve and mean average precision (mAP). The CMC metric checks the position of the first matching gallery image in the ranked list for each query image and obtains the rank-k accuracy. The rank-k accuracy indicates the probability of correct matching results appearing in the top k in the ranking list. For example,

TABLE 3. Description of the training data used in the 5-shot setting.

Number of Training Samples	
Market-1501	DukeMTMC-ReID
3710/12936 (29%)	3510/16522 (21%)
CUHK03	VeRi-776
3835/7368 (52%)	2880/37778 (7.6%)

the rank-1 accuracy equals 1 when the label of the first image in the sorted gallery images matches the label of the query image. The mAP metric reflects the positions of all the gallery images that belong to the same object identity as that of the query image as a whole.

C. IMPLEMENTATION DETAILS

We implement all our experiments in PyTorch and use ResNet-50 pretrained with ImageNet as the backbone network of the feature extractor [34]. When training the FSOR model, we use a 5-shot setting for few-shot learning. In the 5-shot setting, we select at most 5 samples for each ID from the training set of each dataset. As summarized in Table 3, we select 3,710, 3,510, 3,835 and 2,880 images from the Market-1501, DukeMTMC-ReID, CUHK03, and VeRi-776 datasets, respectively, as training samples. All the training and testing images input into the FSOR model are resized to "256 × 128". For H&C metric learning, the batch size is 80, each batch contains 16 IDs (n_q), and there are 5 samples for each ID (N). Then, we divide the 5 samples of each ID into two parts: 4 samples from the support set (S), and 1 sample acts as the query set (Q). In addition, the warmup learning rate at epoch t is determined as follows:

FIGURE 7. Some training samples in the (a) Market-1501, (b) DukeMTMC-ReID, (c) CUHK03, and (d) VeRi-776 datasets.

$$lr(t) = \begin{cases} 3.5 \times 10^{-4} \times \dfrac{t}{10} & if\ t \le 10 \\ 3.5 \times 10^{-4} & if\ 10 < t \le 40 \\ 3.5 \times 10^{-5} & if\ 40 < t \le 70 \\ 3.5 \times 10^{-6} & if\ 70 < t \le 120, \end{cases} \tag{13}$$

When using the 5-shot setting, all the person Re-ID datasets and the vehicle Re-ID dataset share the same experimental settings described above.

TABLE 4. Rank-1 accuracy and mAP comparisons with state-of-the-art methods.

Experiment with a 5-shot setting	Market-1501		DukeMTMC-ReID		CUHK03-labeled		VeRi	
	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
IDE [14]	70.42±0.55	56.34±0.68	31.62±0.75	16.16±0.36	17.88±0.61	16.26±0.58	47.60±0.75	14.34±0.19
CamStyle [15]	70.42±0.55	56.34±0.68	56.86±0.92	34.08±0.49	-	-	-	-
TriNet [17]	75.68±0.41	56.60±0.32	62.90±2.19	44.28±2.77	50.40±2.67	49.20±1.94	68.52±2.13	41.84±1.59
FAT [19]	81.12±0.48	56.28±0.85	67.00±0.53	45.80±0.48	51.30±0.72	47.18±0.80	70.76±0.49	35.56±0.40
DMML [20]	73.56±1.52	51.94±0.97	20.32±0.91	9.30±0.32	16.30±0.84	16.16±0.59	27.90±2.64	10.80±0.67
BagTricks [21]	77.30±0.12	54.98±0.30	67.02±0.80	51.22±0.59	44.14±1.00	41.68±0.49	66.80±0.78	38.30±0.31
LiftedSturectured[23]	12.02±2.21	6.32±1.58	6.78±0.71	3.14±0.38	3.48±0.95	4.70±0.83	22.13±1.15	11.45±0.61
CircleLoss[24]	77.60±0.45	60.94±0.64	71.28±0.35	52.74±0.50	60.38±0.83	58.98±0.69	68.32±1.15	42.32±0.46
ProxyAnchor[25]	44.54±3.83	24.56±2.99	22.28±4.03	12.60±3.01	32.30±2.89	33.28±2.55	43.92±4.54	18.78±0.64
SupCon[26]	70.88±0.82	53.52±0.68	63.66±1.00	45.46±0.61	57.04±0.73	54.90±0.79	56.52±1.55	34.96±0.27
SPMVC [28]	80.1	62.8	70.7	50.2	-	-	-	-
FSOR (ours)	**86.64±0.24**	**70.2±0.4**	**76.42±0.36**	**59.78±0.23**	**74.4±0.44**	**72.64±0.4**	**74.2±0.65**	**46.62±0.33**

TABLE 5. A comparison between FSOR and some baseline metric learning methods.

Experiment with a 5-shot setting	Market-1501		DukeMTMC-ReID		CUHK03-labeled		VeRi	
	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
Only ID loss[22]	65.06±1.25	40.24±1.37	56.62±0.97	39.84±1.00	49.28±0.82	47.14±0.67	58.72±0.43	27.04±0.46
Contrastive[40]	82.30±0.48	62.18±0.3	66.34±0.50	48.36±0.31	65.64±0.57	62.74±0.19	62.48±0.68	34.10±0.31
Triplet[18]	84.10±0.54	64.28±0.81	74.90±0.25	57.64±0.43	69.86±0.98	67.60±0.78	72.52±0.28	44.60±0.19
N-pair[41]	85.22±0.82	67.34±0.59	74.06±0.79	57.34±0.60	72.54±0.37	70.18±0.04	71.70±0.71	42.80±0.36
LiftedSturectured[23]	82.40±0.60	65.12±0.53	55.12±0.41	41.26±0.36	73.04±0.97	71.36±0.53	53.98±0.56	30.30±0.31
CircleLoss[24]	85.54±0.89	69.46±1.65	75.66±1.20	59.08±0.84	73.22±1.19	70.02±1.09	69.74±1.30	44.76±0.47
ProxyAnchor[25]	84.82±0.46	66.50±0.51	73.66±0.32	56.98±0.43	68.76±0.54	67.16±0.21	73.56±1.02	44.72±0.30
SupCon[26]	83.04±0.75	63.44±1.15	71.50±0.87	54.88±0.89	70.32±0.58	67.80±0.72	68.70±0.72	42.70±0.53
H&C (ours)	**86.64±0.24**	**70.2±0.4**	**76.42±0.36**	**59.78±0.23**	**74.4±0.44**	**72.64±0.4**	**74.2±0.65**	**46.62±0.33**

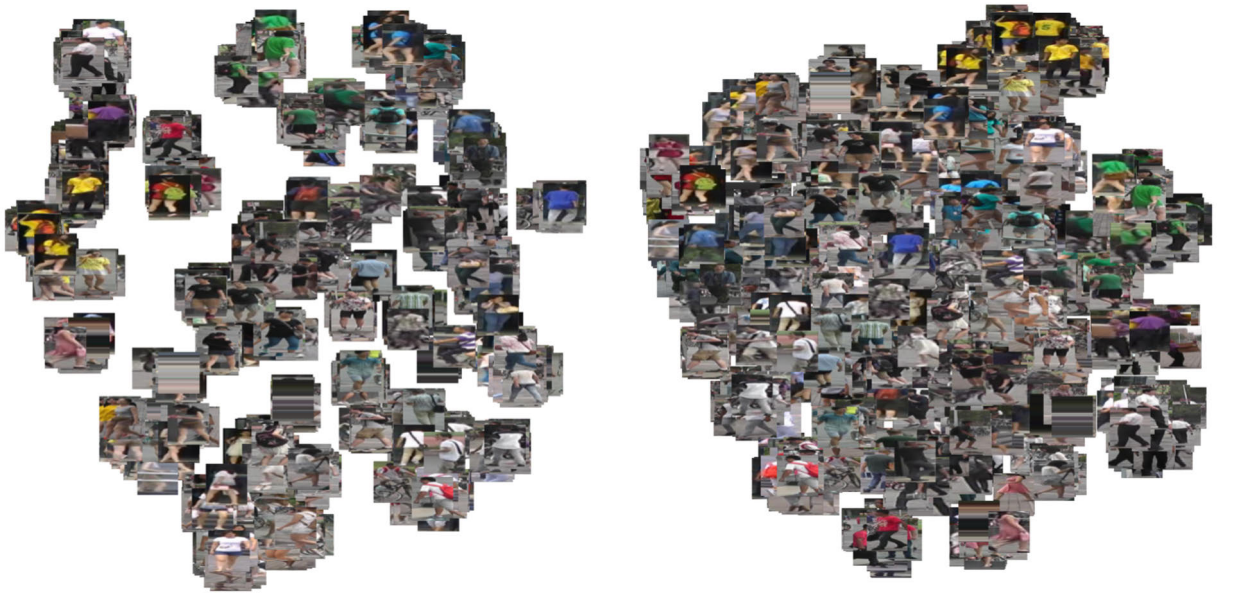

FIGURE 8. Visualization results of feature distribution with and without H&C metric learning.

D. RE-IDENTIFICATION METHOD COMPARISON

1) COMPARISON WITH STATE-OF-THE-ART METHODS

Table 4 shows a comparison between the FSOR method and some state-of-the-art methods in terms of the mAP and rank-1 accuracy when training data are scarce. We repeat the experiments 5 times and show the mean and std of the results that are close to each other for different random seeds. To verify the generalization ability of our FSOR approach, we perform experiments not only on the person Re-ID dataset but also on the vehicle Re-ID dataset. From this table, we can observe that the performances of most existing object Re-ID methods degrade substantially in the experiment under the 5-shot setting. However, the FSOR model achieves the best performance when the training data are scarce, even better than that of SPMVC, which is a semi-supervised method that uses similar amounts of labeled data as those in our method. The experiments of CamStyle on CUHK03-labeled and VeRi are lack because of that the CamStyle needs some specifically generated data to train the model, but the author of CamStyle did not offer them. It makes that the experiments on CamStyle cannot be conducted. The experiments of SPMVC on CUHK03-labeled and VeRi are lack because the method does not provide open-source code. We therefore can only show the results that have been showed in the published paper.

2) COMPARISON WITH BASELINE METRIC LEARNING METHODS

To validate the superiority of H&C metric learning when it is used in the construction of object Re-ID models, Table 5 shows a comparison between the H&C method and some baseline metric learning methods. To ensure a fair comparison, only the metric learning loss differs for all the

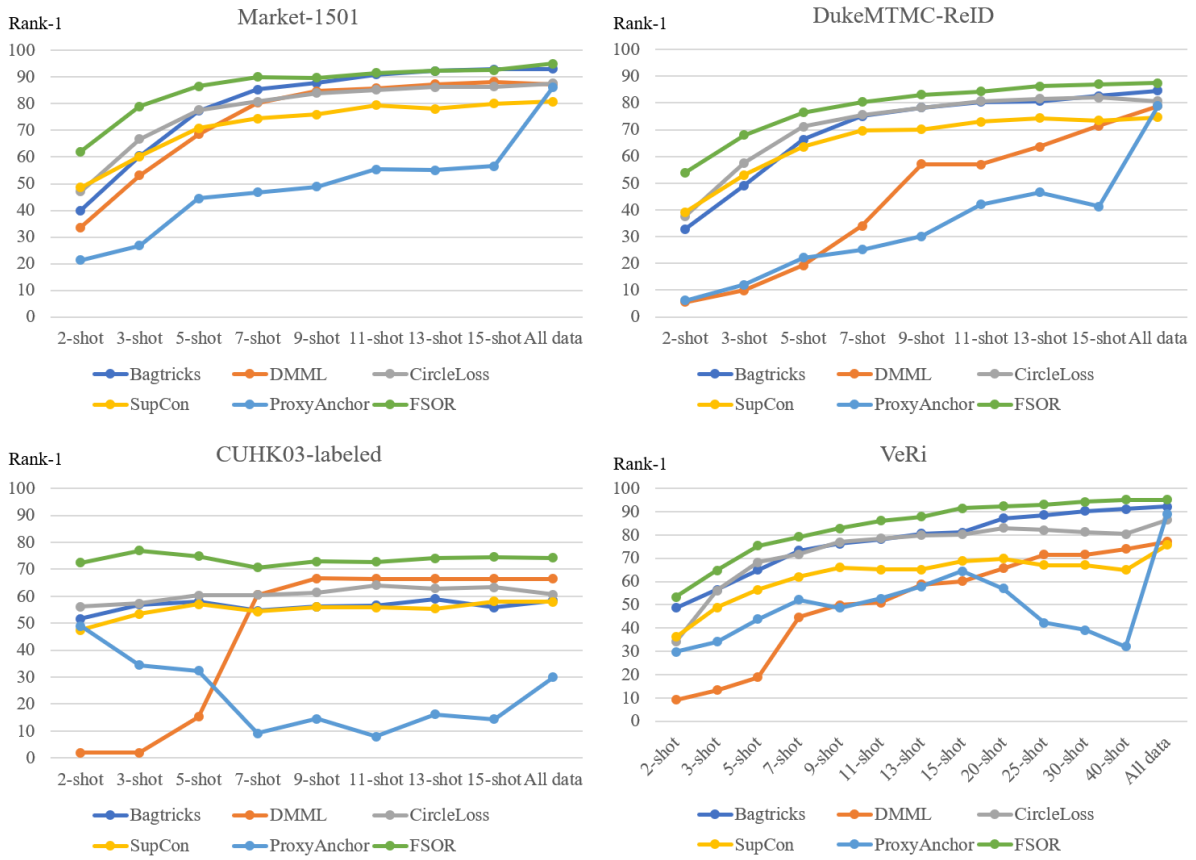

FIGURE 9. The rank-1 accuracy with difference number of training samples per ID on each dataset with different methods.

TABLE 6. Result of an ablation experiment on the Market-1501 and DUKEMTMC-ReID datasets.

5-shot setting	Market-1501		DukeMTMC-ReID		CUHK03-labeled		VeRi	
	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
Baseline	68.10±0.57	45.08±0.36	58.94±0.89	39.50±1.14	53.26±1.71	51.16±1.56	59.54±0.82	33.08±0.32
+ Data Augmentation Methods	71.98±2.11	49.82±2.62	61.28±0.73	44.22±1.45	51.94±0.66	50.94±0.67	64.16±0.94	35.90±0.48
+ Reparameterization Mechanism	84.16±0.90	66.60±1.38	73.72±1.56	57.48±2.16	71.72±1.12	69.60±1.20	72.32±0.66	45.04±0.78
+ H&C Metric Learning	**86.28±0.27**	**69.92±0.62**	**76.52±0.56**	**59.94±0.21**	**74.82±0.29**	**72.88±0.18**	**74.04±0.42**	**46.68±0.38**

methods compared in this experiment. From the results listed in Table 5, we can observe that the H&C method achieves the best performance, outperforming the other metric learning methods in terms of its discriminative ability with few labeled training samples. We also visualize the feature distribution to verify the discrimination ability of the model trained with H&C metric learning on the Market-1501 dataset in Figure 8. The model trained with H&C metric learning makes the distances between feature vectors belonging to the same classes closer and the distance between feature vectors in different classes further than those yielded by the model trained without H&C metric learning.

3) ABLATION STUDY AND PARAMETER ANALYSIS

To verify the effectiveness of each component in the FSOR model, we perform an ablation experiment with a 5-shot setting on the same four datasets, as shown in Table 6. First, we build a baseline model that employs the label smoothing ID loss and metric learning based on the hard mining distance. Then, the data augmentation, reparameterization, and H&C metric learning mechanisms are added step by step to investigate the effect of each. The experimental results of the ablation study show that the reparameterization mechanism greatly improves both the rank-1 accuracy and mAP scores by more than 9% on all experimental datasets. The improvement on CUHK03 is much bigger than that on the other sets. This phenomenon may result from CUHK03 being the smallest among all testing datasets. This means that its data distribution is sparser than that of other datasets. Our reparameterization mechanism can greatly improve the situation.

In addition, the influences of different numbers of samples per ID are analyzed. As shown in Figure 9, the performance

TABLE 7. Notation table.

Notation	Description	Notation	Description
q_i	prediction logits of class i	I_q	query image
p_i	prediction score for class i	I_g^i	gallery image with index i
p_l	prediction score for class l	I	input image
K	number of labels	$z_q{}'$	final feature vector of query image
F	feature extractor	$z_g^i{}'$	final feature vector of gallery image
f	feature vector outputted from F	q_j^m	query sample with ID m and index j
l_{mean}	linear layer to obtain the mean vector	S^n	the support set of ID n
$l_{variance}$	linear layer to obtain the variance vector	s_i^n	sample of S^n with index i and ID n
μ	mean vector	m	ID of query sample
σ	variance vector	n	ID of support sample
R	reparameterization operation	x	image
z	new feature vector outputted from R	y	image
z'	new feature vector generated by the batch normalization layer	z_x	final feature vector of x
z'_q	final feature vector of query sample	z_y	final feature vector of y
z'_s	final feature vector of support sample	d	function to calculate distance
D_{train}	training dataset	c^n	center feature vector of ID n in each batch
Q_{train}	query set of training dataset	M	the amount of image of ID n in each batch.
S_{train}	support set of training dataset	N	the number of IDs in each batch
bs	batch size in training	L_{HM}	loss function calculated with hard mining distance
D'_{train}	training dataset after data augmentation	n_q	number of query samples
Q'_{train}	query set of training dataset after data augmentation	L_{CM}	loss function calculated with query-center distance
S'_{train}	support set of training dataset after data augmentation	ε	hyper-parameter in the query-center distance-based label smoothing
p	probability vector	L_{ML}	H&C metric learning function
$\mathbf{w}_\mu$	weights for getting mean vector	λ_{hm}	hyper-parameter for hard mining distance
b_μ	bias for getting mean vector	λ_{cm}	hyper-parameter for query-center distance
$\mathbf{w}_\sigma$	weights vector for getting variance vector	$std\ (input)$	function to calculate the standard deviation of $input$
b_σ	bias vector for getting variance vector	$mean\ (input)$	function to calculate the mean of $input$
v	noise vector in reparameterization mechanism	δ	hyper-parameter for standard deviation
D_c	query-center distance	L_{total}	hybrid loss function for FOSR training
D_h	hard mining distance	λ_{kl}	hyper-parameter for Kullback-Leibler (KL) divergence
L_{KL}	Kullback-Leibler (KL) divergence of whole batch	λ_{ml}	hyper-parameter for metric learning
L_{KL}^i	Kullback-Leibler (KL) divergence of one sample with index i	λ_{id}	hyper-parameter for label smoothing ID loss
σ_{ij}	value of variance vector with index i at dimension j	L_{ID}	label smoothing ID loss
μ_{ij}	value of mean vector with index i at dimension j	t	current epoch
J	dimensionality		

of FSOR improves as the number of training samples per ID increases. Furthermore, when the FSOR model uses only approximately half of the training data for training (the 9-shot setting in Market-1501 uses 48.6% of the training data, the 11-shot setting in DukeMTMC-ReID uses 46.6% of the training data, the 5-shot setting in CUHK03 uses 52.04% of the training data, and the 30-shot setting in VerRi-776 uses 44.0% of the training data), it achieves rank-1 accuracy scores close to those of other existing models using all the training data. In this experiment, the result of proxy anchor method which records proxy information to help the training is unstable because similar training samples with different labels may mislead the proxy information in this method.

VI. CONCLUDING REMARKS

In this paper, a novel few-shot object Re-ID method, FSOR, is presented; it efficiently constructs object Re-ID models without the need for tedious data collection and

time-consuming annotation processes. Moreover, it guarantees the discrimination and generalization abilities of object Re-ID models with an efficient few-shot learning model that employs a reparameterization mechanism and a dual-distance metric learning approach, named H&C metric learning. The reparameterization mechanism makes the constructed object Re-ID model more generalizable and adaptive, allowing it to re-identify objects not covered in the training data. The proposed H&C metric learning enhances the discrimination ability of the constructed model by combining the advantages of query-center distance and hard-mining distance. According to our experimental results, both of the reparameterization and H&C metric learning can increase more than 17% mAP in average. In addition, we employ several simple but effective techniques, such as data augmentation, a warmup learning rate, and label smoothing, during the construction and operation processes of the FSOR model.

The extensive experimental results and comparisons show that FSOR effectively improves model performances on object Re-ID tasks when the amount of training data is small. Our method even outperforms a semi-supervised Re-ID method when only a few labeled training data are available and without a large number of unlabeled data. The experimental results of the ablation study show that the reparameterization and H&C metric learning schemes significantly improve the performances of object Re-ID models. We also observe from the experimental results of the metric learning method comparison that the proposed H&C metric learning technique is most suitable for model training when the amount of training data is insufficient for satisfying other approaches.

APPENDIX

See Table 7.

REFERENCES

[1] S. Karanam, Y. Li, and R. J. Radke, "Person re-identification with discriminatively trained viewpoint invariant dictionaries," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4516–4524.

[2] Y. Wang, L. Wang, Y. You, X. Zou, V. Chen, S. Li, G. Huang, B. Hariharan, and K. Q. Weinberger, "Resource aware person re-identification across multiple resolutions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8042–8051.

[3] Y.-J. Cho and K.-J. Yoon, "Improving person re-identification via pose-aware multi-shot matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1354–1362.

[4] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. H. Hoi, "Deep learning for person re-identification: A survey and outlook," 2020, *arXiv:2001.04193*. [Online]. Available: http://arxiv.org/abs/2001.04193

[5] M. Shu, "Deep learning for image classification on very small datasets using transfer learning," M.S. thesis, Dept. Comput. Eng., Iowa State Univ., Ames, IA, USA, 2019.

[6] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni, "Generalizing from a few examples: A survey on few-shot learning," *ACM Comput. Surv.*, vol. 53, no. 3, pp. 1–34, Jul. 2020.

[7] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2013, pp. 1–14.

[8] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," 2017, *arXiv:1708.04896*. [Online]. Available: http://arxiv.org/abs/1708.04896

[9] X. Fan, W. Jiang, H. Luo, and M. Fei, "SphereReID: Deep hypersphere manifold embedding for person re-identification," *J. Vis. Commun. Image Represent.*, vol. 60, pp. 51–58, Apr. 2019.

[10] 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. (CVPR)*, Jun. 2016, pp. 2818–2826.

[11] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 480–496.

[12] 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.

[13] D. Yi, Z. Lei, and S. Z. Li, "Deep metric learning for practical person re-identification," in *Proc. Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2014, pp. 34–39.

[14] Z. Zheng, L. Zheng, and Y. Yang, "A discriminatively learned CNN embedding for person re-identification," 2016, *arXiv:1611.05666*. [Online]. Available: http://arxiv.org/abs/1611.05666

[15] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang, "CamStyle: A novel data augmentation method for person re-identification," *IEEE Trans. Image Process.*, vol. 28, no. 3, pp. 1176–1190, Mar. 2019.

[16] C. Wang, L. Song, G. Wang, Q. Zhang, and X. Wang, "Multi-scale multi-patch person re-identification with exclusivity regularized softmax," *Neurocomputing*, vol. 382, pp. 64–70, Mar. 2020.

[17] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," 2017, *arXiv:1703.07737*. [Online]. Available: http://arxiv.org/abs/1703.07737

[18] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 815–823.

[19] Y. Yuan, W. Chen, Y. Yang, and Z. Wang, "In defense of the triplet loss again: Learning robust person re-identification with fast approximated triplet loss and label distillation," 2019, *arXiv:1912.07863*. [Online]. Available: http://arxiv.org/abs/1912.07863

[20] G. Chen, T. Zhang, J. Lu, and J. Zhou, "Deep meta metric learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9547–9556.

[21] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang, "Bag of tricks and a strong baseline for deep person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 1487–1495.

[22] H. Luo, W. Jiang, Y. Gu, F. Liu, X. Liao, S. Lai, and J. Gu, "A strong baseline and batch normalization neck for deep person re-identification," *IEEE Trans. Multimedia*, vol. 22, no. 10, pp. 2597–2609, Oct. 2020.

[23] H. O. Song, Y. Xiang, S. Jegelka, and S. Savarese, "Deep metric learning via lifted structured feature embedding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4004–4012.

[24] Y. Sun, C. Cheng, Y. Zhang, C. Zhang, L. Zheng, Z. Wang, and Y. Wei, "Circle loss: A unified perspective of pair similarity optimization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6398–6407.

[25] S. Kim, D. Kim, M. Cho, and S. Kwak, "Proxy anchor loss for deep metric learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3238–3247.

[26] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2020, pp. 1–18.

[27] Y. Dong, W. Shi, B. Du, X. Hu, and L. Zhang, "Asymmetric weighted logistic metric learning for hyperspectral target detection," *IEEE Trans. Cybern.*, early access, May 26, 2021, doi: 10.1109/TCYB.2021.3070909.

[28] X. Xin, X. Wu, Y. Wang, and J. Wang, "Deep self-paced learning for semi-supervised person re-identification using multi-view self-paced clustering," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 2631–2635.

[29] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap, "One-shot learning with memory-augmented neural networks," 2016, *arXiv:1605.06065*. [Online]. Available: http://arxiv.org/abs/1605.06065

[30] O. Vinyals, C. Blundell, T. Lillicrap, and D. Wierstra, "Matching networks for one shot learning," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2016, pp. 3630–3638.

[31] J. Snell, K. Swersky, and R. S. Zemel, "Prototypical networks for few-shot learning," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 4080–4090.

[32] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. S. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1199–1208.

[33] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Jul. 2017, pp. 1126–1135.
[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[35] 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. (ICCV)*, Dec. 2015, pp. 1116–1124.
[36] 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. Workshop Benchmarking Multi-Target Tracking*, 2016, pp. 17–35.
[37] W. Li, R. Zhao, T. Xiao, and X. Wang, "DeepReID: Deep filter pairing neural network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 152–159.
[38] X. Liu, W. Liu, T. Mei, and H. Ma, "A deep learning-based approach to progressive vehicle reidentification for urban surveillance," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Oct. 2016, pp. 869–884.
[39] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.
[40] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2, Jun. 2006, pp. 1735–1742.
[41] K. Sohn, "Improved deep metric learning with multi-class N-pair loss objective," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2016, pp. 1857–1865.

SHENG-HUNG FAN received the M.S. degree in computer science and information engineering from the National Cheng Kung University, in 2017, where he is currently pursuing the Ph.D. degree in computer science and information engineering. His research interests include computer vision, machine learning, and deep learning.

MIN-HONG LIN received the B.S. degree in biomedical science from the National Chung Cheng University, in 2018. He is currently pursuing the M.S. degree in computer science and information engineering with the National Cheng Kung University. His research interests include computer vision, machine learning, and deep learning.

JUNG-YI JIANG received the M.S. and Ph.D. degrees in electrical engineering from Sun Yat-sen University, Taiwan, in 2004 and 2011, respectively. He currently holds a postdoctoral position with the Center for Research of E-life Digital Technology, National Cheng Kung University, Taiwan. His main research interests include machine learning, data mining, and information retrieval.

YAU-HWANG KUO (Senior Member, IEEE) received the Ph.D. degree in computer engineering from the National Cheng Kung University (NCKU), Tainan, Taiwan, in 1988.

He has served as the Dean for the College of Science, National Chengchi University. He is currently a Distinguished Professor with the Department of Computer Science and Information Engineering, NCKU. During his career, he has been persistently active in academia, education accreditation, and government policy planning. He has published more than 336 papers and 42 patents. His research interests include artificial intelligence, artificial intelligence-based privacy preserving, intelligent data analytics, and 5G AIoT. He has served as the Director for the Computer Center, Ministry of Education, the Computer Science and Information Engineering Program on the National Science Council (NSC), and the Engineering and Technology Promotion Center (ETPC), NSC. From 1999 to 2000, he was elected as the President of the Taiwanese Artificial Intelligence Association. He has also served as an editor for several international journals and consulted for several research institutes and high-tech companies.