Simulation Modelling Practice and Theory 119 (2022) 102568

Contents lists available at ScienceDirect

Simulation Modelling Practice and Theory

journal homepage: www.elsevier.com/locate/simpat

Mutual prediction learning and mixed viewpoints for unsupervised-domain adaptation person re-identification on blockchain

Shuang Li [a,b], Fan Li [a,b], Kunpeng Wang [c,*], Guanqiu Qi [d,*], Huafeng Li [a,b]

[a] Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, Yunnan, PR China
[b] Key Laboratory of Artificial Intelligence in Yunnan Province, Kunming University of Science and Technology, Kunming 650500, Yunnan, PR China
[c] School of Information Engineering, Southwest University of Science and Technology, Mianyang 621010, PR China
[d] Computer Information Systems Department, State University of New York at Buffalo State, Buffalo, NY 14222, USA

ARTICLE INFO

Keywords:
Person re-identification
Mutual prediction learning
Unsupervised domain adaptation
Reasoning imagination
Blockchain

ABSTRACT

In addition to the domain shift between different datasets, the diversity of pedestrian appearance (physical appearance and postures) caused by different camera views also affects the performance of person re-identification (re-ID). Since existing methods tend to extract the shared information of the same pedestrian across multiple images, the above diversity issue has not been effectively alleviated. In addition, while making full use of pedestrian image data and realizing its value, there are also risks of privacy leakage and data loss. Therefore, this paper proposes the mutual prediction learning (MPL) and mixed viewpoints for unsupervised domain adaptation (UDA) person re-ID on blockchain. This method enables the network to first obtain the ability of MPL on multi-view polymorphic features and further acquire the reasoning imagination to alleviate the ambiguity caused by morphological differences. In the process of MPL, the training samples are first divided into different groups and each group has two sets. Then the corresponding identity classifiers of every two sets are integrated and applied to the cross-prediction of polymorphic features. Finally, the joint distribution alignment of domain- and identity-level features is achieved. Furthermore, an adversarial mechanism of mixed viewpoints is proposed to improve the accuracy of identity matching. The domain-invariant salient features are extracted and fused with the polymorphic features obtained by global average pooling (GAP) after domain alignment. Thanks to blockchain technology, the pedestrian image data of the data owner is also protected. Comparative experimental results confirm the effectiveness of the proposed solution in person re-ID. The related source codes will be available at: https://github.com/lhf12278/MPL-MV.

1. Introduction

Person re-ID aims to match people across multiple pedestrian images captured by non-overlapping cameras [1–4]. Since person re-ID has been widely used in intelligent monitoring, tracking criminal suspects, and finding missing people, more and more researchers are joining the related research [5–10]. Some excellent supervised person re-ID methods have been proposed. However, due to the domain shift between source and target datasets, the performance of existing supervised person re-ID models slumps significantly,

* Corresponding authors.
E-mail addresses: kwang@swust.edu.cn (K. Wang), qig@buffalostate.edu (G. Qi).

https://doi.org/10.1016/j.simpat.2022.102568

Available online 6 May 2022

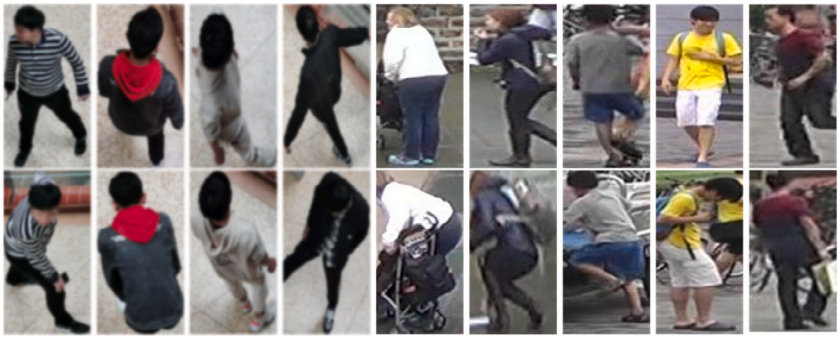

Fig. 1. The morphological differences of the same person across different camera views. The images shown in each column contain the same person captured from two different camera views.

when they are directly applied to real-world scenes. Therefore, UDA person re-ID has attracted considerable attention, and a lot of effective algorithms have been proposed [11–16].

Although UDA person re-ID has made great progress, existing methods still face great challenges in real-world applications due to the ambiguity of pedestrian appearance. Moreover, most of existing UDA person re-ID methods only focus on the influence of the domain shift between source and target domains. Actually, the morphological inconsistency of the same pedestrian across different camera views as another key factor also affects the performance of person re-ID. As shown in Fig. 1, the appearance of pedestrians (such as physical appearance and postures) varies greatly from different perspectives. Existing posture-invariant feature learning methods tend to extract the shared features across multiple images that are not affected by any pedestrian posture. Although these methods can alleviate the morphological differences caused by different postures, the complementary information in different forms of pedestrians is ignored, resulting in the extracted features cannot fully describe the appearance of pedestrians. If the feature extraction network can have a certain prediction learning ability, morphological features from the input single-view images can be predicted, which can effectively alleviate the adverse effects on the recognition performance caused by morphological differences.

In addition, pedestrian image data is one of the important privacy information of pedestrians. Data breaches can cause unnecessary damage, so people are reluctant to share pedestrian image data. Lack of sufficient data often leads to poor performance of deep learning models [17]. As an emerging technology, blockchain has the characteristics of decentralization, tamper resistance, immutability, non-repudiation and traceability [18]. It has been widely used in medical data sharing and protection [19–21], cloud data protection [22–24] and so on. Therefore, blockchain technology can also provide data protection for deep learning-based person re-ID, and promote the development of person re-ID methods.

This paper proposes an MPL and mixed viewpoints for UDA person re-ID (MPLMV) on blockchain, which cannot only provide technical support for preventing the leakage of pedestrian image data, but also alleviate the ambiguity of features through the inference and prediction of multi-view and polymorphic features. To reduce computational complexity, the proposed solution divides the training samples into different groups according to camera views. The grouping of the training samples ensures the mutual prediction between different camera views can be realized. Each group has two sample sets to ensure that the MPL can be applied to two sets of camera views. The classifiers corresponding to two sample sets in the same group are first integrated, and then a cross-classification mechanism ensures the feature encoder realizes the mutual prediction of the complementary features from the same pedestrians between two different sets. This design not only achieves the mutual prediction of polymorphic features, but also realizes the joint distribution alignment of domain- and identity-level features. Therefore, the discrepancy between different domains is effectively eliminated. In addition, an adversarial mechanism of mixed viewpoints is proposed to exploit the distinctive features of pedestrians to further enhance the discriminability of the obtained features. The domain-invariant salient features are first extracted, and then merged with polymorphic features obtained by GAP after domain alignment.

Overall, this paper has three main contributions as follows.

- An MPL is proposed to alleviate the issues of identity matching caused by the morphological differences of pedestrians from different perspectives. This method can simultaneously realize the joint distribution alignment of domain- and identity-level features during the implementation of MPL.
- An adversarial mechanism of mixed viewpoints is proposed to extract and fuse the domain-invariant salient features with global features to describe pedestrian appearance. This mechanism effectively improves the discriminability of discriminators and promotes the extraction of domain-invariant salient features.
- A blockchain-based UDA person re-ID method is proposed. Thanks to blockchain technology, data users cannot directly access pedestrian image data, which effectively ensures the safety of pedestrian image data of data owners. In addition, the proposed MPLMV neither requires any pseudo-label prediction nor relies existing additional model assistance, so it has important practical significance for person re-ID.

The rest of this paper is organized as follows. Section 2 discusses existing work; Section 3 presents the proposed method in detail; Section 4 analyzes the comparative experimental results; and Section 5 concludes this paper.

2. Related work

Since the proposed solution belongs to UDA person re-ID, this section discusses three typical types of UDA-person re-ID methods, including clustering and pseudo-label prediction (CPLP)-based methods [25–29], AMA-based methods [30–32], and domain adaptation feature learning (DAFL)-based methods [33–38].

2.1. CPLP-based UDA person re-ID

CPLP usually uses a clustering algorithm to predict the pseudo labels of target samples, and further optimizes the re-ID model in a supervised way. Existing clustering methods tend to introduce noisy labels. Therefore, some pseudo-label refinery methods were proposed [26,28,39]. Particularly, Yang et al. [26] designed an asymmetric co-teaching framework to select data that may have clean labels to resist noisy labels. Ge et al. [39] presented an unsupervised framework with mutual mean-teaching (MMT). Zhao et al. [40] developed a noise resistible mutual-training method to suppress the noise marked in pseudo labels. Luo et al. [41] first separated the target data according to camera perspectives, and then predicted and refined the sample labels across the camera views so that the confident labels became more reliable. Zeng [28] integrated hierarchical clustering with hard-batch triplet loss to improve the pseudo-label prediction of existing methods. To solve the non-convergence issues of existing cluster-based pseudo-label prediction methods in practical applications, Ji et al. [42] proposed an attention-driven two-stage clustering method. Since each sample in the target datasets participating in the model training has a positive sample, which provides a great help for the correct prediction of pseudo labels. However, it is common that pedestrians are only captured by a single camera in real-world monitoring scenes. Due to the interference from the isolated pedestrians captured by a single camera, the performance of CPLP-based UDA person re-ID methods may drop dramatically, when they are applied to real-world scenes.

2.2. AMA-based UDA person re-ID

Additional model assistance (AMA)-based methods usually first use the additionally pre-trained models, such as style transfer models and pedestrian pose extraction models, to enable the person re-ID models to overcome the ambiguity of pedestrian appearance features caused by image style differences or pedestrian posture differences. Then, they further improve the performance of person re-ID models. Particularly, in PTGAN [30], SPGAN [31] and ATNet [32], the labeled source-domain samples are first transferred to target domain, and then the transferred source-domain samples are used in supervised training. However, these methods ignore the intra-domain changes of target domain, resulting in the recognition performance cannot be further improved. Zhong et al. [43] did a comprehensive study on the intra-domain changes of target domain, and proposed to assign three basic invariances, sample invariance, camera invariance, and neighborhood invariance to the person re-ID model. To alleviate the impact of the inconsistency of pedestrian postures and the domain discrepancy on the recognition performance, Li et al. [44] applied a posture extraction model to extract the posture information of pedestrians, and used the adversarial generation mechanism to obtain the invariant features of postures after domain alignment. Although AMA-based UDA person re-ID methods can achieve good recognition performance, they need to use an additional model to perform the specific preprocessing on the training data in the model training process, which seriously affects their application efficiency in practical scenes.

2.3. DAFL-based UDA person re-ID

The domain-adaptation based feature extraction is usually dedicated to learn the transferable features to resolve the domain discrepancy. In particular, Wang et al. [35] developed a deep learning based person re-ID method to transfer the labeled information of existing datasets to a new unlabeled target domain. Wu et al. [36] proposed a consistency loss of camera-aware similarity to learn the distribution of consistent pairwise similarity for matching within and across cameras. Yang et al. [33] proposed a patch-based unsupervised learning framework to learn features from patches rather than from the entire image. Qi et al. [34] proposed an adaptive framework of unsupervised camera-aware domain to solve the domain differences between different camera views in cross-domain person re-ID. DAFL-based person re-ID has high training efficiency and strong scalability, so they are in line with the needs of real-world scenes. However, due to the lack of the fine-tuning of pseudo-label prediction and the additional model assistance, they only show relatively low recognition performance.

3. Proposed method

3.1. Blockchain and MPLMV workflow

In order to prevent pedestrian image data from leaking pedestrian privacy during use, this paper proposes a blockchain-based UDA person re-ID method. Considering the characteristics of decentralization, tamper resistance, and traceability, blockchain can be used as a database to store pedestrian image data. As shown in Fig. 2, the proposed method consists of a blockchain database (built by ethereum) and MPLMV. MPLMV is trained by acquiring pedestrian image data from a blockchain database. The users of the blockchain database are categorized into data owners and data users. Data owners can upload pedestrian image data to the blockchain database. Each data owner only has permission to view and download his/her own data. Data users first request

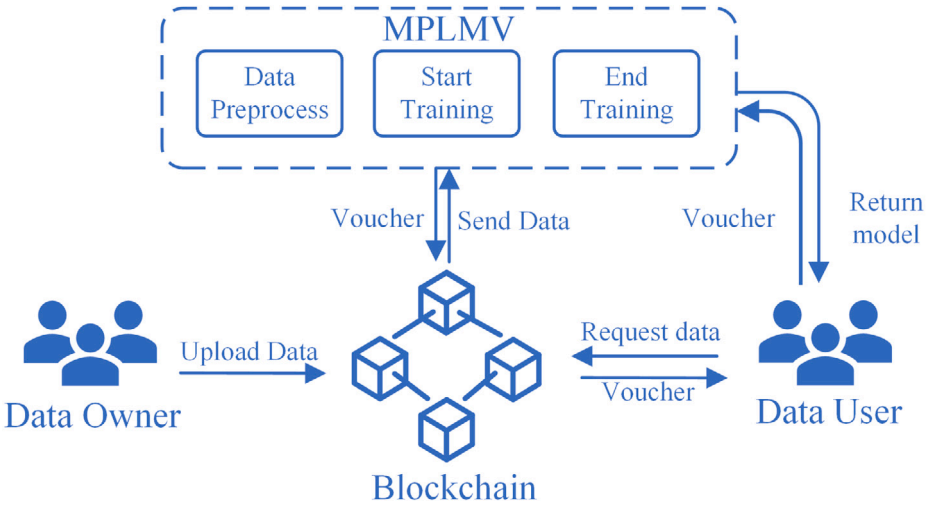

Fig. 2. The Blockchain and MPLMV workflow.

pedestrian image data from the blockchain database to obtain vouchers for using the data, then call the model for training through the vouchers, and finally obtain the trained model.

Specifically, data users select the corresponding type of pedestrian image data in the blockchain database according to their own needs to complete the transactions. The smart contract of the blockchain database sends the vouchers of the corresponding pedestrian data to the data users. MPLMV then fetches data from the blockchain database with vouchers sent by the data users and starts training. After training, the final model is sent to the data users. In the whole process, data users can get their own satisfactory models, and data owners do not need to worry about the leakage of pedestrian image data.

3.2. MPLMV

Let $S = \{(x_{s,i}, y_{s,i}, c_{s,i})|_{i=1}^{n_s}\}$ be the labeled source-domain sample set, where the subscript s means the corresponding sample is from the source-domain sample set S, n_s is the total number of pedestrian images in the sample set S, and $x_{s,i}$ is the ith image, $y_{s,i}$ and $c_{s,i}$ denote the corresponding identity label and camera label of $x_{s,i}$, respectively. Suppose there are K identities in S, $y_{s,i} \in \{1, 2, \ldots, K\}$. Similarly, the defined target-domain data $\boldsymbol{T} = \{(x_{t,i}, c_{t,i})|_{i=1}^{n_t}\}$ contains n_t pedestrian images. Fig. 3 shows the overall process of the proposed solution. The whole framework consists of mutual prediction learning (MPL), domain-invariant salient features learning (DISFL), and alignment and fusion of domain-invariant salient features and multi-view polymorphic features (AFDSF+MPF). MPL is used to make the feature encoder $\boldsymbol{E}_1$ have a certain reasoning ability. So, it can predict and extract polymorphic features from the input single-view pedestrian images and avoid extracting the common features. The extraction of domain-invariant salient features is applied to compensate the information loss caused by GAP in multi-view polymorphic feature extraction. In the whole process, the source-domain samples are used to train the multi-view polymorphic feature encoder $\boldsymbol{E}_1$, domain-invariant salient feature encoder $\boldsymbol{E}_2$, and pedestrian ID classifier $\boldsymbol{W}_{id1}$ with both cross entropy loss and triplet loss as follows.

$$\boldsymbol{L}_{id} = -\frac{1}{n_b}\sum_{i=1}^{n_b} q_{s,i}\log(\boldsymbol{W}_{id1}(\boldsymbol{E}_1(x_{s,i}))) + q_{s,i}\log(\boldsymbol{W}_{id1}(\boldsymbol{E}_2(x_{s,i}))), \tag{1}$$

$$\begin{aligned}\boldsymbol{L}_{tri} = &\frac{1}{n_b}\sum_{i=1}^{n_b}[m + \|\boldsymbol{E}_1(x_{s,i}) - \boldsymbol{E}_1(x_{s,i}^p)\|_2 - \|\boldsymbol{E}_1(x_{s,i}) - \boldsymbol{E}_1(x_{s,i}^n)\|_2]_+ \\ &+ [m + \|\boldsymbol{E}_2(x_{s,i}) - \boldsymbol{E}_2(x_{s,i}^p)\|_2 - \|\boldsymbol{E}_2(x_{s,i}) - \boldsymbol{E}_2(x_{s,i}^n)\|_2]_+,\end{aligned} \tag{2}$$

where m is set to 0.03 empirically, n_b represents the batch size, $x_{s,i}^n$ and $x_{s,i}^p$ are the hard-negative and hard-positive samples of $x_{s,i}$ respectively, $[\cdot]_+ = max(\cdot, 0)$ is the hinge loss, $q_{s,i} \in \mathbb{R}^{K\times 1}$ is a one-hot vector, and only the element at $y_{s,i}$ is 1.

3.2.1. Mutual prediction learning

The encoder $\boldsymbol{E}_1$ obtained by minimizing the loss in Eqs. (1) and (2) does not have the ability to extract the domain-invariant features. It is possible to introduce an adversarial mechanism like [34] to make $\boldsymbol{E}_1$ have the ability to extract the domain-invariant features. However, this method tends to extract the common information of the same pedestrian from different perspectives, which ignores the complementarity of polymorphic features from different perspectives and thereby reduces the discriminability of pedestrian appearance features. In fact, due to the differences from different perspectives, the same pedestrian often shows

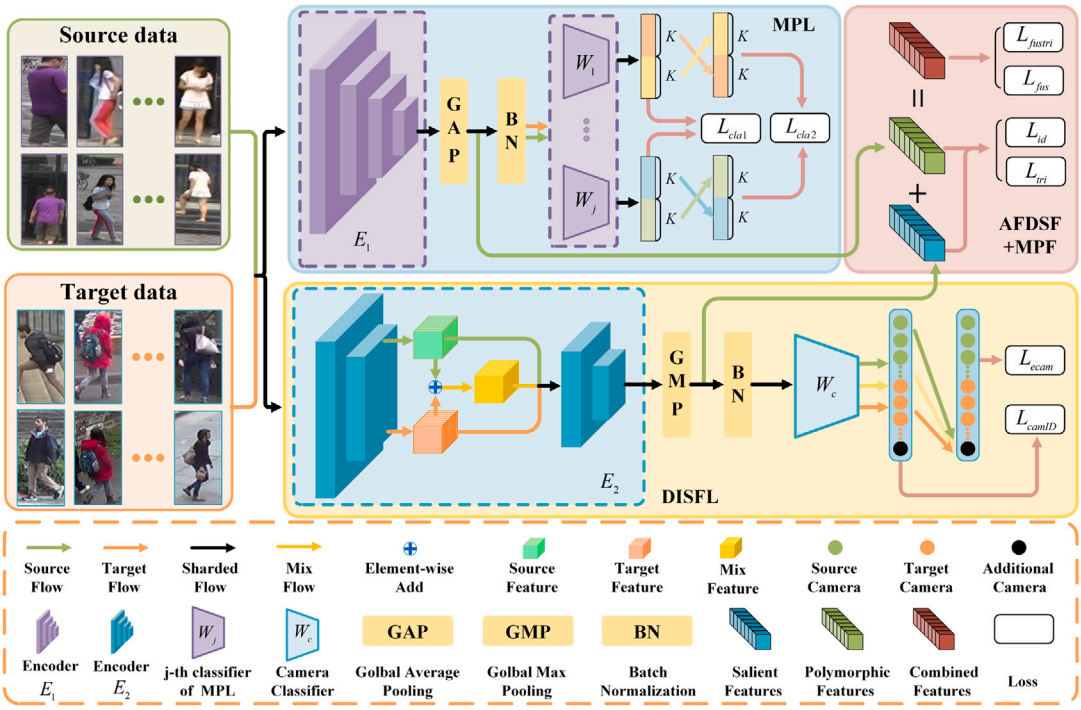

Fig. 3. The proposed MPLMV. It consists of MPL, DISFL and AFDSF+MPF. MPL is used to prompt the encoder E_1 to extract the multi-view polymorphic features. The DIFSL is achieved by using the mixed perspectives at domain level. In AFDSF+MPF, the multi-view polymorphic and domain-invariant salient features are aligned and combined to obtain the complete descriptions of pedestrian appearance.

the inconsistent appearance features from different perspectives. If $\boldsymbol{E}_1$ can predict the corresponding features related to other perspectives from a single-view pedestrian image, it is conducive to the extraction of relatively complete pedestrian features. Inspired by the latest advances in UDA [45–47], this paper proposes a new adversarial learning mechanism to achieve the MPL among different perspectives.

This method first divides both source- and target-domain samples participating in training into the approximately equal parts $G_s^j = \{G_{s,1}^j, G_{s,2}^j\}$ and $G_t^j = \{G_{t,1}^j, G_{t,2}^j\}$ according to the corresponding camera IDs. $G_{s/t,1}^j$ and $G_{s/t,2}^j$ represent pedestrian image sets of the first and second sample sets in the jth group, respectively. According to the principle of the approximate equality, the camera numbers in the first and second sets are $\left\lfloor \frac{V_{s/t}}{2} \right\rfloor$ and $\left\lceil \frac{V_{s/t}}{2} \right\rceil$ respectively, where $\lfloor\cdot\rfloor$ and $\lceil\cdot\rceil$ represent rounding down and up, and $V_{s/t}$ represents the number of cameras in the source/target domain. As shown in Fig. 4, to realize the MPL between two camera views, each camera ID of $\lceil\cdot\rceil - 1$ in $G_{s/t,1}^j$ or $G_{s/t,2}^j$ is exchanged with any one of the other set to form a new group. In this way, the samples in source and target domains should have $\left\lfloor \frac{V_s}{2} \right\rfloor$ and $\left\lfloor \frac{V_t}{2} \right\rfloor$ groups, respectively. Finally, G_s^j and G_t^j are randomly combined to form a group. The final number of groups is $\max\left\{\left\lfloor \frac{V_s}{2} \right\rfloor, \left\lfloor \frac{V_t}{2} \right\rfloor\right\}$.

To realize the domain alignment at identity level, both camera classifier and pedestrian identity classifier are integrated into one classifier, which is conducive to the joint alignment of identity and domain information [45]. Since each group corresponds to an integrated classifier, a total number $n_w = \max\left\{\left\lfloor \frac{V_s}{2} \right\rfloor, \left\lfloor \frac{V_t}{2} \right\rfloor\right\}$ of the integrated classifiers is needed in the MPL. The output of the jth classifier $\boldsymbol{W}_j$ is the identity probability of the jth sample group. Assuming that the number of source-domain identities categorized into the first and second sets is at most K in the jth group, so the output dimension of the integrated classifier $\boldsymbol{W}_j$ is $2K$, where the former K dimension represents the identity categories of the first set, and the last K dimension represents the identity categories of the second set. In the MPL, $\boldsymbol{W}_j$ is optimized by minimizing the loss in Eq. (3).

$$\begin{aligned} \boldsymbol{L}_{cla1} = &- \frac{1}{n_b} \sum_{j=1}^{n_w} \sum_{i=1}^{n_b} q_{s,i}^j \log(\boldsymbol{W}_j(\boldsymbol{E}_1(x_{s,i})) \\ &+ q_{t,i}^j \log([p_{t,i}^{j,1} \boldsymbol{W}_j(\boldsymbol{E}_1(x_{t,i})), p_{t,i}^{j,2} \boldsymbol{W}_j(\boldsymbol{E}_1(x_{t,i}))])), \end{aligned} \tag{3}$$

where $q_{t,i}^j$ is the label of $x_{t,i}$ in the jth group. When $x_{t,i} \in G_{t,1}^j$, $p_{t,i}^{j,1} = [\mathbf{1}_K^T, \mathbf{0}_K^T]$ and $q_{t,i}^j = [1, 0]$, where $\mathbf{1}_K \in \mathbb{R}^{K\times 1}$ is a K-dimensional one vector, $\mathbf{0}_K \in \mathbb{R}^{K\times 1}$ is a K-dimensional 0 vector, and T is the transpose of a vector. In $p_{t,i}^{j,1}$, the superscript 1 indicates $p_{t,i}^{j,1}$ corresponds to the first k-dimension of $\boldsymbol{W}_j$ output, and the superscript 2 in $p_{t,i}^{j,2}$ indicates $p_{t,i}^{j,2}$ corresponds the last k-dimension of

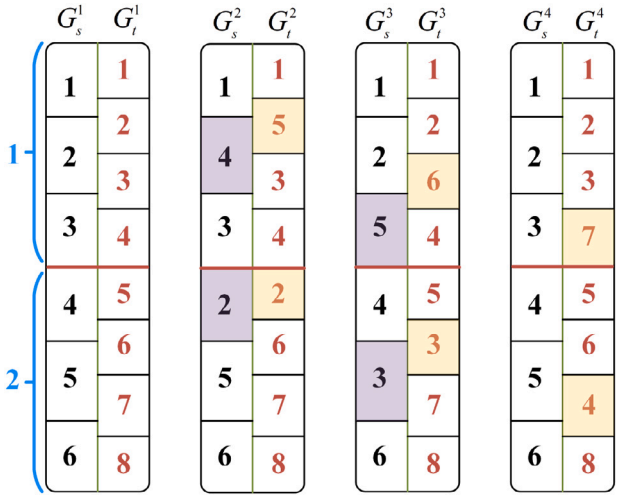

Fig. 4. Illustration of group division. It assumes that both target and source domains have six and eight cameras respectively. In each group $\{G_s^j, G_t^j\}$, the cameras (such as the purple block 4) of source domain in the first set are first exchanged with any one (such as the purple block 2) of source domain in the second set, and then the same operation is applied to the cameras of target domain. Finally, a new group is formed.

$\boldsymbol{W}_j$ output. When $x_{t,i} \in G_{t,2}^j$, $p_{t,i}^{j,2} = [\mathbf{0}_K^T, \mathbf{1}_K^T]$ and $q_{t,i}^j = [0, 1]$. $q_{s,i}^j$ is the label of $x_{s,i}$ in the jth group. When $x_{s,i} \in G_{s,1}^j$, $q_{s,i}^j = [\mathbf{1}_i^T, \mathbf{0}_K^T]$, where $\mathbf{1}_i \in \mathbb{R}^{K\times 1}$ as a K-dimensional one-hot vector represents the true label vector of $x_{s,i}$, and only the element at $y_{s,i}$ is 1. When $x_{s,i} = G_{s,2}^j$, $q_{s,i}^j = [\mathbf{0}_K^T, \mathbf{1}_i^T]$.

With the strong joint classification ability of $\boldsymbol{W}_j$, the proposed solution optimizes the encoder $\boldsymbol{E}_1$ by minimizing the loss function shown in Eq. (4), which prompts the encoder $\boldsymbol{E}_1$ to predict and learn the features of another set of samples, so that the encoder $\boldsymbol{E}_1$ has the MPL ability as follows.

$$\begin{aligned} \boldsymbol{L}_{cla2} = &-\frac{1}{n_b}\sum_{j=1}^{n_w}\sum_{i=1}^{n_b} \tilde{q}_{s,i}^{j,2}\log(\boldsymbol{W}_j(\boldsymbol{E}_1(x_{s,i}^1))) + \tilde{q}_{s,i}^{j,1}\log(\boldsymbol{W}_j(\boldsymbol{E}_1(x_{s,i}^2))) \\ &+ q_{t,i}^{j,1}\log([p_{t,i}^{j,1}\boldsymbol{W}_j(\boldsymbol{E}_1(x_{t,i}^2)), p_{t,i}^{j,2}\boldsymbol{W}_j(\boldsymbol{E}_1(x_{t,i}^2))]) \\ &+ q_{t,i}^{j,2}\log([p_{t,i}^{j,1}\boldsymbol{W}_j(\boldsymbol{E}_1(x_{t,i}^1)), p_{t,i}^{j,2}\boldsymbol{W}_j(\boldsymbol{E}_1(x_{t,i}^1))]), \end{aligned} \tag{4}$$

where $x_{s,i}^1 \in G_{s,1}^j$, $x_{s,i}^2 \in G_{s,2}^j$, $x_{t,i}^1 \in G_{t,1}^j$, $x_{t,i}^2 \in G_{t,2}^j$, $\tilde{q}_{s,i}^{j,2} = [\mathbf{0}_K^T, \mathbf{1}_i^T]$, $\tilde{q}_{s,i}^{j,1} = [\mathbf{1}_i^T, \mathbf{0}_K^T]$, $q_{t,i}^{j,1} = [1, 0]$, $q_{t,i}^{j,2} = [0, 1]$.

The above process is shown in Fig. 5. After obtaining the optimized encoder $\boldsymbol{E}_1$, the loss of Eq. (3) is minimized to further optimize the classifier $\boldsymbol{W}_j$. This can ensure that the two sets of samples in one group are correctly classified. In the optimization of $\boldsymbol{E}_1$, if the network tends to extract the common features of the same pedestrian from different perspectives, the classifier $\boldsymbol{W}_j$ is not able to correctly classify them. This issue can be solved by minimizing the loss function shown in Eq. (4). In addition, since the classifier $\boldsymbol{W}_j$ of a specific group can simultaneously identify two sets of samples in the same group, and perform the mutual prediction of category between two sets of this group, the features in the two sets achieve the joint alignment at both domain and identity levels. Therefore, the features extracted by $\boldsymbol{E}_1$ are domain-invariant and can be used to realize the cross-domain recognition.

3.2.2. Domain-invariant salient feature learning

To prevent the excessive information loss, GAP is used to extract features in the MPL. GAP keeps the global information, but weakens the salient features on the feature map, which causes the distinguishing information of pedestrians fails to play a role in the matching of pedestrian identity. To solve this issue, a domain-invariant salient feature extraction framework is proposed. Similar to the traditional domain-invariant feature extraction methods [34], adversarial learning is used between the camera classification $\boldsymbol{W}_c$ and feature encoder $\boldsymbol{E}_2$ to ensure the domain invariance of the learned features. In this process, the distinguishing ability of $\boldsymbol{W}_c$ is the key to obtain the domain-invariant features.

To improve the distinguishing ability of $\boldsymbol{W}_c$, inspired by mixup [48–50], the output features of the first two layers of $\boldsymbol{E}_2$ are first extracted, and then randomly mixed. The mixed features are fed into the subsequent feature extraction layer of $\boldsymbol{E}_2$ to train $\boldsymbol{W}_c$. Specifically, the mixed features are expressed as follows.

$$\boldsymbol{f}_{mix,i} = \gamma \boldsymbol{E}_{2(l_2)}(x_{s,i}) + (1-\gamma)\boldsymbol{E}_{2(l_2)}(x_{t,i}), \tag{5}$$

where γ is the ratio of the source-domain feature map to the mixed feature map $\boldsymbol{f}_{mix,i}$, γ is set to 0.25, 0.5, or 0.75 empirically, $\boldsymbol{E}_{2(l_2)}$ is the output feature map of the middle layer of the encoder $\boldsymbol{E}_2$. The corresponding camera label of the mixed feature $\boldsymbol{f}_{mix,i}(\gamma, x_{s,i}, x_{t,i})$ is as follows.

$$c_{mix,i} = \gamma c_{s,i} + (1-\gamma)c_{t,i}, \tag{6}$$

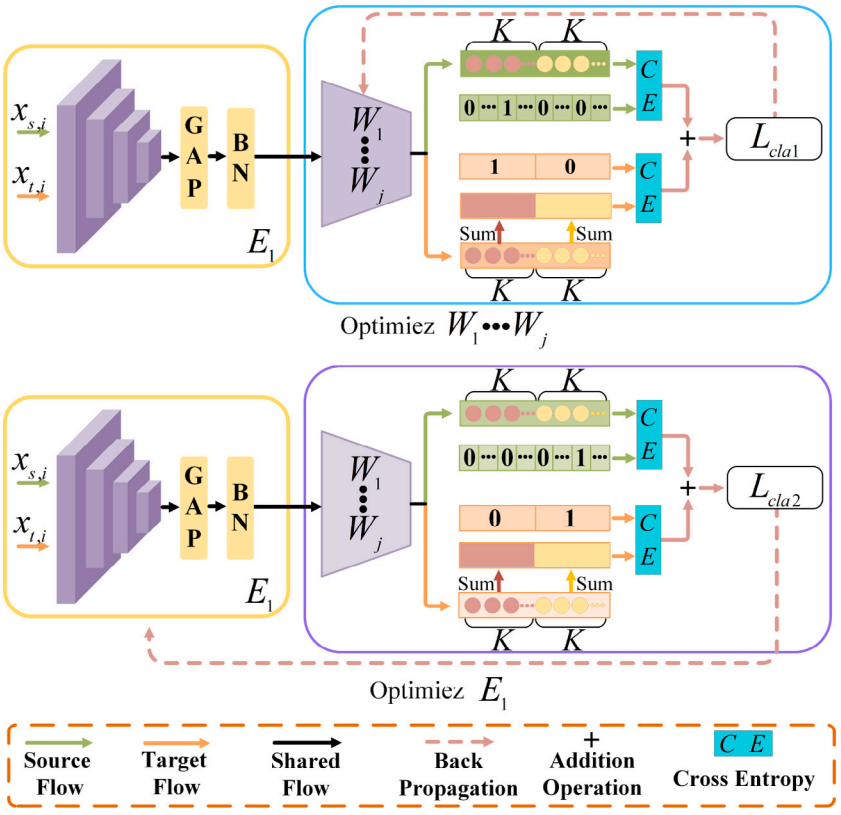

Fig. 5. Mutual prediction learning module. It shows the adversarial learning process of encoder E_1 and classifier W_j. In the first stage, the classifier W_j is optimized by L_{cla1}. In the second stage, W_j is fixed and encoder E_1 is optimized by L_{cla2}.

where $c_{s,i}$ and $c_{t,i}$ are the camera labels of $x_{s,i}$ and $x_{t,i}$, respectively. Both $c_{s,i}$ and $c_{t,i}$ are one-hot vector.

In the above process, the output dimension of W_c is $n_e = n_{s,c} + n_{t,c} + 1$, where $n_{s,c}$ is the number of cameras in source domain, and $n_{t,c}$ is the number of cameras in target domain. In the adversarial learning of E_2 and W_c, the features of source- and target-domain, and mixed samples are applied to the supervised training of W_c. In this process, the used loss function is shown in Eq. (7).

$$\begin{aligned} L_{camID} = &- \frac{1}{n_b}\sum_{i=1}^{n_b} c_{s,i}\log(W_c(f_{s,i}^2)) + c_{t,i}\log(W_c(f_{t,i}^2)) \\ &+ c_{mix,i}\log(W_c(f_{mi,i})), \end{aligned} \tag{7}$$

where $f_{t,i}^2 = E_2(x_{t,i})$, $f_{s,i}^2 = E_2(x_{s,i})$, and $f_{mi,i}$ is the output of subsequent feature extraction layer of E_2. When W_c is updated, the parameters are fixed to further optimize E_2. The extracted features of source- and target-domain, and mixed samples can be classified into an additional class by W_c to realize the domain alignment of these features. The loss function used in this process is shown in Eq. (8).

$$\begin{aligned} L_{ecam} = &- \frac{1}{n_b}\sum_{i=1}^{n_b} c_{n_e}\log(W_c(f_{s,i}^2)) + c_{n_e}\log(W_c(f_{t,i}^2)) \\ &+ c_{n_e}\log(W_c(f_{mi,i})), \end{aligned} \tag{8}$$

where c_{n_e} denotes the label of the additional class and c_{n_e} is a one-hot vector.

3.2.3. Alignment and fusion of salient and multi-view polymorphic features

Multi-view polymorphic features and domain-invariant salient features describe pedestrians from different aspects of their appearance, so these two types of features have a certain complementarity. Consequently, the combination of features is conducive to improving the accuracy of pedestrian identity matching. However, these two types of features are extracted by different encoders, and there is no feature alignment between them. So, it is difficult to directly add or concatenate them to use. To address this issue, the fusion of features $f_{s,i}^{mer} = E_1(x_{s,i}) + E_2(x_{s,i})$ by minimizing the losses in Eq. (9)(10).

$$L_{fus} = -\frac{1}{n_b}\sum_{i=1}^{n_b} q_{s,i}\log(W_{id2}(f_{s,i}^{mer})), \tag{9}$$

Table 1
The settings of different person Re-ID datasets in performance comparison. Ped: Number of pedestrians; Img: Number of images; Cam: Number of cameras.

Datasets	Ped	Training		Gallery (Testing)		Probe (Testing)		Cam
		Ped	Img	Ped	Img	Ped	Img	
Market1501	1501	751	12936	750	19732	750	3368	6
Duke	1812	702	16522	1110	17661	702	2228	8
MSMT17	4101	1041	32621	3060	82161	3060	11659	15
Market1501(S1)	1367	617	3197	750	19732	750	3368	6
Duke(S2)	1255	553	5300	702	17661	702	2228	8
MSMT17 (S2)	2831	1790	15356	1041	29721	1041	2900	15
PRID2011	749	400	500	349	349	100	100	2
GRID	1025	525	650	500	500	125	125	6

$$L_{fustri} = \frac{1}{n_b}\sum_{i=1}^{n_b}[m + \|\boldsymbol{f}_{s,i}^{mer} - \boldsymbol{f}_{s,i}^{p,mer}\|_2 - \|\boldsymbol{f}_{s,i}^{mer} - \boldsymbol{f}_{s,i}^{n,mer}\|_2]_+, \tag{10}$$

where $\boldsymbol{W}_{id2}$ is an identity classifier, $\boldsymbol{f}_{s,i}^{n,mer}$ and $\boldsymbol{f}_{s,i}^{p,mer}$ are the hard-negative and hard-positive samples of $\boldsymbol{f}_{s,i}^{mer}$.

3.2.4. Network structure and optimization

Network structure: ResNet50 [51] pre-trained on ImageNet [52] is used as the backbone of $\boldsymbol{E}_1$ and $\boldsymbol{E}_2$. The GAP layer of $\boldsymbol{E}_2$ is replaced by the global maximum pooling layer. $\boldsymbol{W}_j$ contains a dropout operation and four fully-connected layers. The dropout parameter is 0.5. The channels of the fully-connected layers are $1000, 1000, 1000, and\ 2K$, respectively. The first three fully-connected layers are followed by BN layer and ReLu layer. $\boldsymbol{W}_c$ is composed of six fully-convolutional layers and a fully-connected layer. The channels of the fully-convolutional layers are $1024, 512, 256, 128, 64, and\ 32$, respectively. Each fully-convolutional layer is followed by the BN layer and ReLu layer. The channels of the fully-connected layer are n_e.

Optimization: In the training, the total objective is shown as follows.

$$\begin{aligned} L_{full} = {} & L_{id}(\boldsymbol{E}_1, \boldsymbol{E}_2, \boldsymbol{W}_{id1}) + L_{tri}(\boldsymbol{E}_1, \boldsymbol{E}_2) + L_{fustri}(\boldsymbol{E}_1, \boldsymbol{E}_2) \\ & + L_{fus}(\boldsymbol{E}_1, \boldsymbol{E}_2, \boldsymbol{W}_{id2}) + \xi_1(L_{cal1}(\boldsymbol{W}_j) + L_{cal2}(\boldsymbol{E}_1)) \\ & + \xi_2(L_{camID}(\boldsymbol{W}_c) + L_{ecam}(\boldsymbol{E}_2)), \end{aligned} \tag{11}$$

where $\xi_1, \xi_2 \geq 0$ as the weights to balance the importance of the related loss items. Encoders $\boldsymbol{E}_1$ and $\boldsymbol{E}_2$ participate in the whole training process involving the optimization of $L_{id}(\boldsymbol{E}_1, \boldsymbol{E}_2, \boldsymbol{W}_{id1})$, $L_{tri}(\boldsymbol{E}_1, \boldsymbol{E}_2)$, $L_{fus}(\boldsymbol{E}_1, \boldsymbol{E}_2, \boldsymbol{W}_{id2})$ and $L_{fustri}(\boldsymbol{E}_1, \boldsymbol{E}_2)$. In the 70th epoch, the classifier $\boldsymbol{W}_c$ is added to the training process. $L_{camID}(\boldsymbol{W}_c)$ and $L_{ecam}(\boldsymbol{E}_2)$ are minimized to assist $\boldsymbol{E}_2$ in extracting the domain-invariant salient features. At the 80th epoch, the classifier $\boldsymbol{W}_j$ is added to the training process. An MPL is performed by minimizing $L_{cal1}(\boldsymbol{W}_j)$ and $L_{cal2}(\boldsymbol{E}_1)$ to assist $\boldsymbol{E}_1$ in extracting the multi-view polymorphic features.

4. Experiments

4.1. Datasets and evaluation protocol

This paper uses eight challenging public datasets, Market1501 [53], DukeMTMC-reID (Duke) [54], MSMT17 [55], Market1501 (S1), Duke (S2), MSMT17 (S2), PRID2011 [56], and GRID [57] to test the performance of the proposed method. The detailed settings of these datasets used in experiments are shown in Table 1, and the details of each dataset are given as follows.

Market1501 consists of 32,668 images of 1,501 pedestrians captured by six non-overlapping cameras. The training set contains 12,936 images of 751 pedestrians, and each pedestrian appears in at least two camera views. The gallery set contains 19,732 images of the remaining 750 pedestrians. In the testing phase, 3,368 images of 750 pedestrians in the probe set are used to match all pedestrian images in the gallery set.

Duke contains 36,411 pedestrian images captured by eight non-overlapping cameras. The training set contains 16,522 images of 702 pedestrians, and each pedestrian is captured by at least two cameras. The gallery set contains 17,661 images of 1,110 pedestrians, in which 408 pedestrian images are interference images. The probe set contains 2,228 images of 702 identities.

MSMT17 is a large-scale dataset for person re-ID. It consists of 126,441 images of 4,101 pedestrians captured by 15 non-overlapping cameras. The training set contains 32,621 images of 1,041 pedestrians, and each pedestrian appears in at least two camera views. The testing set contains 93,820 images of 3,060 pedestrians, in which 11,659 images are selected as the probe set and the remaining 82,161 images are used as the gallery set.

Market1501 (S1) uses the pedestrian samples in the training set of Market1501 to simulate a real-world scene based on the assumption that the pedestrians from adjacent camera views are not exactly same. As shown in Fig. 6, it assumes that the number of pedestrians from each camera view moving to an adjacent camera is 25% of the total pedestrians from the camera view on a straight road, which means that the pedestrians from the adjacent camera views are not completely overlapped. According to this ratio, the training set contains 3,197 images of 617 pedestrians. The testing set still follows the original protocol of Market1501. The probe and gallery sets contain 3,368 and 19,732 images, respectively.

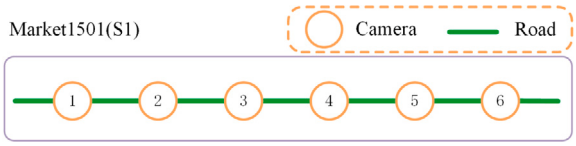

Fig. 6. Real-world scene setting 1 (S1) simulated by Market1501 dataset samples. The data of Market1501 dataset is captured by six cameras. It assumes that six cameras are installed at six different intersections and the number of pedestrians from each camera perspective moving to an adjacent camera perspective is about 25% of the total pedestrians from the camera perspective.

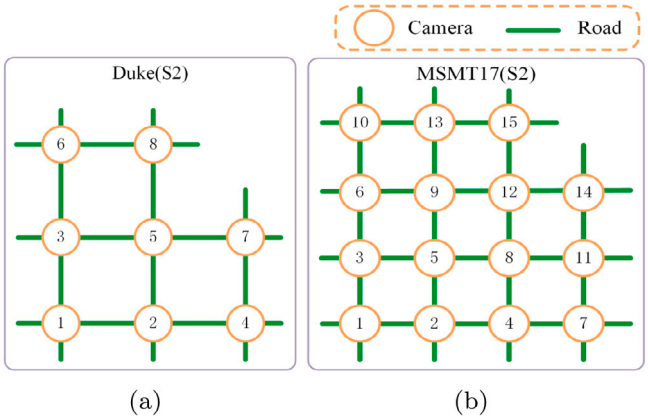

Fig. 7. Real-world scene setting 2 (S2) simulated by Duke and MSMT17 dataset samples. The subfigure (a) shows a monitoring network scene simulated using the Duke dataset samples collected by eight cameras, and the subfigure (b) shows a monitoring network scene simulated using the MSMT17 dataset samples collected by 15 cameras. It assumes that the scenes shown in subfigures (a) and (b) are the street networks composed of multiple intersections. Each intersection has a camera that can monitor all pedestrians from different directions. The number of pedestrians from each camera perspective moving to an adjacent camera perspective is about 25% of the total pedestrians from the camera perspective.

Duke (S2) uses the training samples of Duke dataset to simulate the video surveillance networks in a real-world urban street scene as shown in Fig. 7(a). It assumes that the number of pedestrians from each camera view moving to an adjacent camera is 25% of the total pedestrians from the camera view, which is same as the settings of Market1501(S1). It means that the pedestrians from the adjacent cameras are not exactly same. When the camera interval increases, the number of overlapping pedestrians decreases, which is in line with the real-world street scenes. According to this assumption, the training set of the newly constructed dataset Duke (S2) contains 5,300 images of 553 pedestrians. The testing set still follows the original protocol of Duke. The main difference from Duke is that some pedestrians are only captured by one camera installed at the boundary of the video surveillance networks (such as cameras 1,2,4,7,3,6,8).

MSMT17 (S2) uses the testing samples of MSMT17 to simulate the appearance of pedestrians from the cameras installed at different intersections in the real-world urban street monitoring scenes as shown in Fig. 7(b). Since the testing set of the original MSMT17 has more data samples than its training set, a new training set is constructed by applying the settings of S2 to the testing set of the original MSMT17, and the original training set is directly used as the testing set according to the protocol of [58].

PRID2011 is composed of 1,134 images of 934 pedestrians captured by cameras A and B. 385 and 749 pedestrians appeared under the views of cameras A and B, respectively. But, only 200 pedestrians appeared under both camera views at the same time. In this paper, 100 pairs of pedestrian images that appeared under both camera views at the same time and 300 interference images are selected as the training set, the remaining images of 100 pedestrians captured by camera A are used as the probe set, and the remaining images of 100 pedestrians captured by camera B and 249 interference images are used as the gallery set.

GRID contains 250 pedestrian image pairs captured by eight non-overlapping cameras and 775 interference images. Two images of each pedestrian image pair were captured by different cameras. 125 pedestrian image pairs and 400 interference images are randomly selected for training. The remaining 125 pedestrian image pairs are used for testing, of which 125 images with different pedestrian identities are used as the probe set, and the remaining 125 pedestrian images and 375 interference images are used as the gallery set.

Evaluation Metrics: Cumulative Matching Characteristics (CMC)[59] and mean Average Precision (mAP)[60] as two objective evaluation indicators are used to evaluate the performance of the proposed model and comparative models.

4.2. Implementation details

In the training phase, the image size is set to 256 × 128. Similar to Luo's solution [61], data enhancement in this paper is achieved through random cropping, flipping, and color dithering. In the experiments, the batch size n_b is set to 16, all the networks

Table 2
The performance comparisons of different methods on Duke → Market1501 and Market1501 → Duke. The CMC and mAP rates (%) of each method are reported. "–" means no data reported and the bold values indicate the best results.

Settings	Duke→ Market1501				Market1501→ Duke			
	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP
Methods	Methods with CPLP							
HHL [64]	62.2	78.8	84.0	31.4	46.9	61.0	66.7	27.2
PCB-PAST [27]	78.4	–	–	54.6	72.4	–	–	54.3
UDA-TP [65]	75.8	89.5	93.2	53.7	68.4	80.1	83.5	49.0
ACT [26]	80.5	–	–	60.6	72.4	–	–	54.5
MEB-NET [66]	89.9	96.0	97.5	76.0	79.6	88.3	92.2	66.1
MMT [39]	87.7	94.9	96.9	71.2	78.0	88.8	92.5	65.1
SPCL [67]	90.3	96.2	97.7	76.7	82.9	90.1	92.5	68.8
DG-Net++ [68]	82.1	90.2	92.7	61.7	78.9	87.8	90.4	63.8
Methods	Methods with MAM							
CamSty [69]	58.5	78.2	84.3	27.4	48.4	62.5	68.9	25.1
SBSGAN [70]	58.5	–	–	27.3	53.5	–	–	30.8
ATNet [32]	55.7	73.2	79.4	25.6	45.1	59.5	64.2	24.9
ECN [43]	75.1	87.6	91.6	43.0	63.3	75.8	80.4	40.4
PDA-Net [44]	75.2	86.3	90.2	47.6	63.2	77.0	82.5	45.1
LVRP [71]	63.9	81.1	86.4	33.9	36.3	54.0	61.6	17.9
Methods	Methods without CPLP and MAM							
ECN [43]	58.0	69.9	75.6	27.7	39.7	53.0	58.1	23.6
CaNE [72]	57.2	73.0	80.0	27.4	–	–	–	–
DG-Net++ [68]	52.2	70.7	77.0	28.6	53.2	68.7	73.8	36.3
CBN [73]	72.7	–	–	43.0	58.7	–	–	38.2
SNR [74]	66.7	–	–	33.9	55.1	–	–	33.6
MMCL [75]	66.6	–	–	35.3	58.0	–	–	40.2
CAC [76]	69.4	82.8	87.3	36.9	57.5	71.2	75.3	37.0
AADFL [77]	71.8	85.9	90.1	39.6	64.1	77.2	**81.4**	43.1
TALMVR [11]	73.1	70.7	–	40.0	63.5	76.6	–	41.3
Proposed	**75.3**	**88.1**	**91.9**	**43.1**	**65.4**	**77.8**	81.1	**43.3**

use the Adam optimizer, and the weight decay is set to 0.0005. The initial learning rates of $\boldsymbol{E}_1$ and $\boldsymbol{E}_2$ are set to 0.0002, and a total of 150 epochs is performed on model training. When the source domain is Market1501, the learning rates of $\boldsymbol{W}_j$ and $\boldsymbol{W}_c$ are set to 0.0002. When the source domain is Duke, the initial learning rates of $\boldsymbol{W}_j$ and $\boldsymbol{W}_c$ are set to 0.00012 and 0.0003, respectively. When the source domain is MSMT17, the initial learning rate is set to 0.0002. In the 0∼10th epoch, the learning rate is adjusted linearly by the warm-up strategy [62]. At the 20th and 70th epochs, the learning rate is decreased by 10%. Hyperparameters ξ_1 and ξ_2 are set to 1 in the experiments, and all the experiments in this paper are implemented in the pytorch[63] framework on a single 2080TI GPU with i9-9900K 3.6 GHz CPU and 64 GB RAM. During the testing process, the cosine similarity is used to match pedestrian identities.

4.3. Comparison to the state-of-the-art methods

To verify the effectiveness, the proposed solution is compared with CPLP-based, MAM-based, and DAFL-based methods on Market1501 → Duke and Duke → Market1501. A → B means that datasets A and B are the source and target domains, respectively. As shown in Table 2, the proposed method are compared with CPLP-based methods HHL [64], PCB-PAST [27], UDA-TP [65], ACT [26], MEB-NET [66], MMT [39], SPCL [67] and DG-Net++[68], MAM-based methods CamSty [69], SBSGAN [70], ATNet [32], ECN [43], PDA-Net [44] and LVRP [71], and DAFL-based methods ECN (without additional model assistance) [43], CaNE [72], DG-Net++ (without self-training) [68], CBN [73], SNR [74], MMCL [75], CAC [76] and AADFL [77]. According to Table 2, the proposed method is significantly better than other similar methods without using any pseudo-label prediction and relying on any additional model assistance. Specifically, compared with the sub-optimal DAFL-based methods, Rank-1 and mAP obtained by the proposed method is 2.2% and 3.1% (1.3% and 0.2%) higher than the second best one TALMVR (AADFL) on Duke → Market1501 (Market1501 → Duke), respectively. The reason is that the proposed method can extract multi-view and polymorphic features by MPL, which makes the learned features more discriminative.

As shown in Table 2, ECN obtained 75.1% /43.0% (63.3% /40.4%) Rank-1/mAP recognition accuracy on Duke → Market1501 (Market1501 → Duke). PDA-Net achieved 75.2%/47.6% and 63.2%/45.1% Rank-1/mAP recognition accuracy on the same tasks. However, the performance of this type of methods depends on additional models, such as ECN relies on the style transfer model, PDA-Net relies on the style transfer model [78,79] and pose estimation model [80]. The introduction of additional models seriously affects the efficiency of the corresponding recognition algorithms. Due to open access of source codes, only the complete source codes of ECN can be downloaded. For the remaining solutions, either partial source codes are available or there is no access to source codes. So, the proposed solution is only compared with ECN on Market1501 → Duke and Market1501→ MSMT17. As shown in Table 3, on Market1501 → Duke, the training time of the proposed method is about 11 h, while the ECN training model takes

Table 3
The performance comparisons between the proposed method and ECN.

Settings	Market1501 → Duke		Market1501 → MSMT17	
Methods	Rank-1	Time/h	Rank-1	Time/h
ECN [43]	63.3	≈ 90	25.3	≈ 420
Proposed	65.4	≈ 11	39.4	≈ 20

Table 4
The performance comparisons of different methods on Market1501 → MSMT17, Duke → MSMT17, MSMT17 → Market1501, and MSMT17 → Duke. The CMC and mAP rates (%) of each method are reported. "–" means no data reported and the bold values indicate the best results.

Settings	Market1501→ MSMT17				Duke→ MSMT17			
	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP
PTGAN [30]	10.2	–	24.4	2.9	11.8	–	27.4	3.3
ECN [43]	25.3	36.3	42.1	8.5	30.2	41.5	46.8	10.2
CAC [76]	29.3	40.2	45.9	10.5	37.0	49.9	55.6	13.3
AADFL [77]	30.5	42.6	48.8	11.4	38.6	50.8	56.1	14.0
TALMVR [11]	30.9	43.5	–	11.2	39.0	51.5	–	14.2
Proposed	**39.4**	**51.6**	**56.8**	**15.1**	**44.2**	**56.8**	**61.9**	**16.4**
Settings	MSMT17→ Market1501				MSMT17→ Duke			
	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP
CASCL [36]	65.4	80.6	86.2	35.5	59.3	73.2	77.8	37.8
CaNE [72]	59.1	75.4	–	30.3	60.7	74.7	–	39.1
CAC [76]	72.7	85.1	88.8	41.0	68.0	80.3	84.3	47.4
TALMVR [11]	74.6	87.6	–	43.0	68.4	81.0	–	49.0
Proposed	**75.8**	**87.9**	**92.0**	**44.5**	**69.9**	**81.3**	**85.3**	**48.8**

about 90 h. The reason is that each camera style transfer in the dataset requires a special training. The more cameras are involved in the dataset, the longer training time is (such as the large-scale dataset MSMT17 shown in Table 1). In contrast, the proposed method does not use any additional model assistance to improve its performance, so its efficiency is higher. In addition, without using any additional model assistance, the performance of the proposed method is more competitive than AMA-based methods, so the proposed method can meet the needs of practical applications effectively.

To further verify the effectiveness and scalability, the performance of the proposed method is evaluated on Market1501 → MSMT17, Duke → MSMT17, MSMT17 → Duke, and MSMT17 → Market1501. Since the data size of Market1501 and Duke is much smaller than MSMT17, the extended experiment is quite in line with actual needs. As shown in Table 4, the recognition accuracy of Rank-1 and mAP obtained by the proposed method outperforms that of AADFL and TALMVR. On the tasks, the proposed method also shows better performance than CASCL, PTGAN, ECN and CaNE. The effectiveness of the proposed method is further verified.

The CPLP-based methods can achieve higher recognition performance by refining the model with pseudo labels. Particularly, the Rank-1 and mAP recognition accuracy of SPCL reach 90.3%/82.9% and 76.7%/68.8% on Duke → Market1501/Market1501 → Duke, respectively. The performance of SPCL is far better than the proposed method, because each target-domain sample of SPCL participating in the training has the corresponding positive samples. However, a certain number of isolated pedestrians may appear in the real-world scenes, which are captured by only one camera in a local area of camera networks. The negative samples composed of these isolated pedestrians may play a negative role in improving the performance of CPLP-based methods. In contrast, the proposed method is not restricted by any isolated pedestrians, so it has higher practical significance. To test the proposed method, Market1501 (S1), Duke (S2), and MSMT17 (S2) that simulate the real-world street scenes are selected as the target domains, and Market1501 and Duke are used as the source domains, respectively. In the comparative experiments, the results of all comparative methods were obtained by using the source codes published by the original authors. The hyperparameter settings of both the proposed method and comparative methods employed the original data used in the corresponding papers under the old protocols, and these parameters were no longer adjusted under the new protocol dataset.

The performance of CPLP-based methods and the proposed method is further verified. According to Table 5, UDA-TP, ACT, MMT, MEB-NET, and SPCL that have good performance on the original protocols show lower performance. When Duke was used as the source domain, the Rank-1 and mAP recognition accuracy of MMT only reached 59.7%/39.0% and 33.7%/17.3% on the Market1501(S1)/MSMT17(S2) dataset, respectively. However, the proposed method reached 71.2%/50.8% Rank-1 recognition accuracy and 39.6%/21.9% mAP recognition accuracy on the Market1501(S1)/MSMT17(S2) dataset, which is far higher than the performance of MMT. When Market1501 was used as the source domain and Duke(S2)/MSMT17(S2) was employed as the target domain, the proposed method obtained 63.2%/45.0% Rank-1 recognition accuracy and 42.0%/19.9% mAP recognition accuracy, which is better than the methods based on clustering pseudo-label prediction. This further confirms the effectiveness of the proposed method and its superiority to the CPLP-based methods.

In addition, compared with the original protocols of Market1501 and Duke, Market1501 (S1) and Duke (S2) reduced the number of samples in the training dataset under the new protocols and the isolated pedestrian samples appeared at the same time, when the testing protocols remain unchanged. After reducing the size of training dataset, all the methods showed the varying degrees of performance degradation. As shown in Table 6, in comparison with the performance on the original datasets, the Rank1 and mAP

Table 5
The performance comparisons of different methods on Duke → Market1501(S1), Market1501 → Duke(S2), Duke → MSMT17(S2) and Market1501 → MSMT17(S2). The CMC and mAP rates (%) of each method are reported. "–" means no data reported and the bold values indicate the best results.

Settings	Duke→ Market1501(S1)				Market1501→ Duke(S2)			
	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP
UDA-TP [65]	56.6	72.2	77.5	30.6	42.9	57.2	63.2	27.7
ACT [26]	51.5	67.0	72.5	26.0	30.3	42.3	48.9	18.0
MEB-NET [66]	57.3	73.0	79.1	33.4	44.2	59.1	65.4	30.7
MMT-500 [39]	59.7	75.1	80.9	33.7	45.4	61.0	67.6	30.8
MMT-700 [39]	57.3	73.5	80.2	32.0	45.2	60.5	67.0	30.5
MMT-900 [39]	58.4	74.3	80.3	32.5	43.9	61.0	67.1	30.8
SPCL [67]	14.1	26.1	33.0	5.6	13.2	21.5	25.3	5.5
Proposed	**71.2**	**85.7**	**89.9**	**39.6**	**63.2**	**75.1**	**79.3**	**42.0**
Settings	Duke→ MSMT17(S2)				Market1501→ MSMT17(S2)			
	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP
UDA-TP [65]	23.3	35.1	41.0	9.8	14.5	25.0	31.2	6.0
ACT [26]	15.2	24.2	29.4	6.4	9.5	18.3	24.1	4.0
MEB-NET [66]	33.9	46.5	52.2	15.9	26.1	37.3	43.5	12.0
MMT-500 [39]	35.8	47.6	53.4	15.3	27.4	39.7	45.7	11.6
MMT-1000 [39]	36.5	50.4	55.9	16.3	30.7	43.9	50.9	13.7
MMT-1500 [39]	39.0	51.0	57.8	17.3	33.2	45.2	51.9	14.5
MMT-2000 [39]	37.8	51.7	57.5	17.2	33.9	47.5	54.9	15.4
SPCL [67]	19.8	31.7	37.7	8.8	18.8	30.4	36.6	9.0
Proposed	**50.8**	**62.6**	**67.7**	**21.9**	**45.0**	**57.9**	**63.2**	**19.9**

Table 6
Comparison of the performance changes (%) of different methods from Duke → Market1501 and Market1501 → Duke to Duke → Market1501(S1) and Market1501 → Duke(S2), respectively. The CMC and mAP rates (%) of each method are reported. "–" means no data reported and the bold values indicate the best results.

Settings	Duke→ Market1501 And Market1501(S1)				Market1501→ Duke And Duke(S2)			
	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP
UDA-TP [65]	19.2	17.3	15.7	23.1	25.5	22.9	20.3	21.3
ACT [26]	29.0	–	–	34.6	42.1	–	–	36.5
MEB-NET [66]	32.6	23.0	18.4	42.6	35.4	29.2	26.8	35.4
MMT-500 [39]	28.0	19.8	16.0	37.5	31.4	27.0	24.6	32.3
MMT-700 [39]	29.5	21.1	16.7	37.0	32.8	28.3	25.5	34.6
MMT-900 [39]	28.4	20.6	16.3	33.7	33.5	27.1	25.4	32.3
SPCL [67]	76.2	70.1	64.7	71.1	69.7	68.6	67.2	63.3
Proposed	**4.1**	**2.4**	**2.0**	**3.5**	**2.2**	**2.7**	**1.8**	**1.3**

recognition accuracy of the CPLP-based methods dropped more than 19% on both Duke → Market1501(S1) and Market1501 → Duke(S2). For the CPLP-based methods, Rank1 dropped by 19.2% at the lowest and 76.2% at the highest, and mAP dropped by 23.1% at the lowest and 71.1% at the highest on Duke → Market1501 (S1). The Rank1/mAP recognition accuracy of the proposed method only dropped by 4.1%/3.5%.

Moreover, for the CPLP-based methods, Rank-1 dropped by 25.5% at the lowest and 69.7% at the highest, and mAP dropped by 21.3% at the lowest and 63.3% at the highest on Market1501 → Duke (S2), while Rank-1 and mAP recognition accuracy of the proposed method only dropped by 2.2% and 1.3%, respectively. In contrast, the performance of the CPLP-based methods was considerably reduced. In addition to the decrease in the number of training samples, the isolated pedestrian images only appearing in one camera view as another important reason cause the significant performance degradation of the CPLP-based methods, which indicates the performance of the CPLP-based methods is easily affected by unmatched samples. The proposed method does not rely on any matched samples, so its performance is less affected. So, this also confirms that the proposed method has more robust practical applicability.

Since the size of the data collected by the monitoring system in real-world scenes is often quite large, this undoubtedly brings great challenges to the matching of pedestrian identities in person re-ID models. In order to improve the matching efficiency of pedestrian images, an image can be extracted from each pedestrian image sequence to participate in the matching of pedestrian images. In this case, each pedestrian has only one image under each camera view. To test the performance of both the proposed method and CPLP-based methods in this protocol, GRID and PRID2011 datasets containing interference images are selected as the target domains, and Market1501 and Duke datasets are used as the source domains, respectively. According to Table 7, UDA-TP, ACT, MMT, and SPCL that perform well on large-scale datasets show lower performance.

When Market1501 was used as the source domain, the Rank-1 and mAP recognition accuracy of MMT only reached 32.0% (31.0%) and 40.3% (39.8%) on GRID (PRID2011), respectively. However, the Rank-1 and mAP recognition accuracy of the proposed method reached 49.3% (53.9%) and 57.4% (62.5%), respectively. When Duke was used as the source domain, the Rank-1 and mAP recognition accuracy of MMT only reached 32.8% (25.0%) and reached 41.3% (33.9%) on the GRID (PRID2011) dataset, respectively. The proposed method obtained 40.6% (54.8%) Rank-1 and 47.9% (63.7%) mAP on GRID (PRID2011), which is far

Table 7
Comparative experiment results of different methods on PRID2011 and GRID datasets. The CMC and mAP rates (%) of each method are reported and the bold values indicate the best results.

Settings	Market1501→ GRID				Duke→ GRID			
	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP
UDA-TP [65]	28.0	50.4	59.2	38.8	27.2	44.8	55.2	35.6
ACT [26]	14.4	31.2	42.4	23.4	13.6	25.6	30.4	20.4
MMT [39]	32.0	48.0	53.6	40.3	32.8	49.6	55.2	41.3
SPCL [67]	13.6	24.0	35.2	20.2	8.0	20.8	33.6	15.7
Proposed	**49.3**	**65.2**	**73.0**	**57.4**	**40.6**	**55.0**	**63.0**	**47.9**
Settings	Market1501→ PRID2011				Duke→ PRID2011			
	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP
UDA-TP [65]	12.0	23.0	35.0	19.5	22.0	47.0	55.0	33.3
ACT [26]	14.0	26.0	38.0	22.2	13.0	31.0	40.0	21.9
MMT [39]	31.0	48.0	57.0	39.8	25.0	41.0	54.0	33.9
SPCL [67]	4.0	10.5	16.5	8.6	16.7	31.1	39.6	7.6
Proposed	**53.9**	**72.7**	**80.0**	**62.5**	**54.8**	**75.7**	**82.4**	**63.7**

Table 8
The ablation study of the proposed method. The CMC and mAP rates (%) of each method are reported.

Methods	Duke → Market1501				Market1501→ Duke			
	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP
B(Baseline)	64.8	78.6	83.7	32.9	46.3	62.4	68.7	27.0
B+MPL	72.1	85.5	90.5	40.8	62.3	75.5	79.8	41.6
B+DISFL	71.9	86.2	90.3	39.6	60.5	73.0	77.6	38.0
B+MPL+DISFL	**75.3**	**88.1**	**91.9**	**43.1**	**65.4**	**77.8**	**81.1**	**43.3**

higher than the performance of MMT. There are two main reasons to cause the poor performance of the CPLP-based methods. The target-domain samples involved in the model training contain interference images, and each pedestrian has only one image under each camera view, which bring great challenges for the correct prediction of pseudo-labels based on clustering methods. In contrast, the proposed method shows better performance, which further confirms the effectiveness of the proposed method and its superiority to the clustering pseudo-label prediction methods.

4.4. Ablation study

The proposed method consists of three parts, MPL, DIFSL and AFDSF+MPF. In this section, the role of each part is analyzed to prove its effectiveness.

Effectiveness of MPL In the proposed method, ResNet50 with average pooling is used as the baseline (B), and both cross-entropy loss and triplet loss are applied to the optimization process. B+MPL is obtained by adding MPL to the baseline and compared with the baseline to verify the effectiveness of MPL. According to Table 8, B+MPL improves the recognition accuracy of Rank-1 from 64.8% (46.3%) to 72.1%(62.3%) and mAP from 32.9% (27.0%) to 40.8%(41.6%) on Duke → Market1501 (Market1501 → Duke), respectively. The above results show that multi-view MPL makes the network be able to extract multi-view and polymorphic features from single-view pedestrian images.

Effectiveness of DISFL. In AFDSF+MPF, the fused salient features are provided by DISFL. Here it only verifies the performance of the domain-invariant salient feature learning branch for the subsequent ablation study. As shown in Table 8, the Rank-1 accuracy of domain-invariant salient feature learning is 71.9% and 60.5% and the recognition accuracy of mAP reaches 39.6% and 38.0% on Duke → Market1501 and Market1501 → Duke, respectively. The above results show that DISFL has extracted the significant identity information of pedestrians.

Effectiveness of AFDSF+MPF. To use the discriminative features extracted by DISFL, the multi-view polymorphic features extracted by MPL are fused with the domain-invariant salient features after alignment, which is called B+MPL+DISFL in Table 8. Compared with MPL, the accuracy of Rank-1/mAP increased by 3.2%/2.3% (3.1%/1.7%) on Duke → Market1501 (Market1501 → Duke). Compared with DISFL, the accuracy of Rank-1/mAP increased by 3.4%/3.5% (4.9%/5.3%) on Duke → Market1501 (Market1501 → Duke). The above results show effectiveness of AFDSF+MPF.

4.5. Parameter analysis

This paper involves two hyperparameters ξ_1 and ξ_2. In the influence analysis of the two hyperparameters, a parameter is fixed to analyze the influence of another parameter on the experimental performance. During this process, all the experiments were conducted on Duke → Market1501 and Market1501 → Duke.

The influence of ξ_1. In Eq. 12, the ξ_1 mainly adjusts the role of L_{cla1} and L_{cla2}. Fig. 8(a) and (c) show the influence of ξ_1 on Rank-1 and mAP when the values of ξ_1 are different on Duke → Market1501 and Market1501 → Duke. Specifically, on Rank-1,

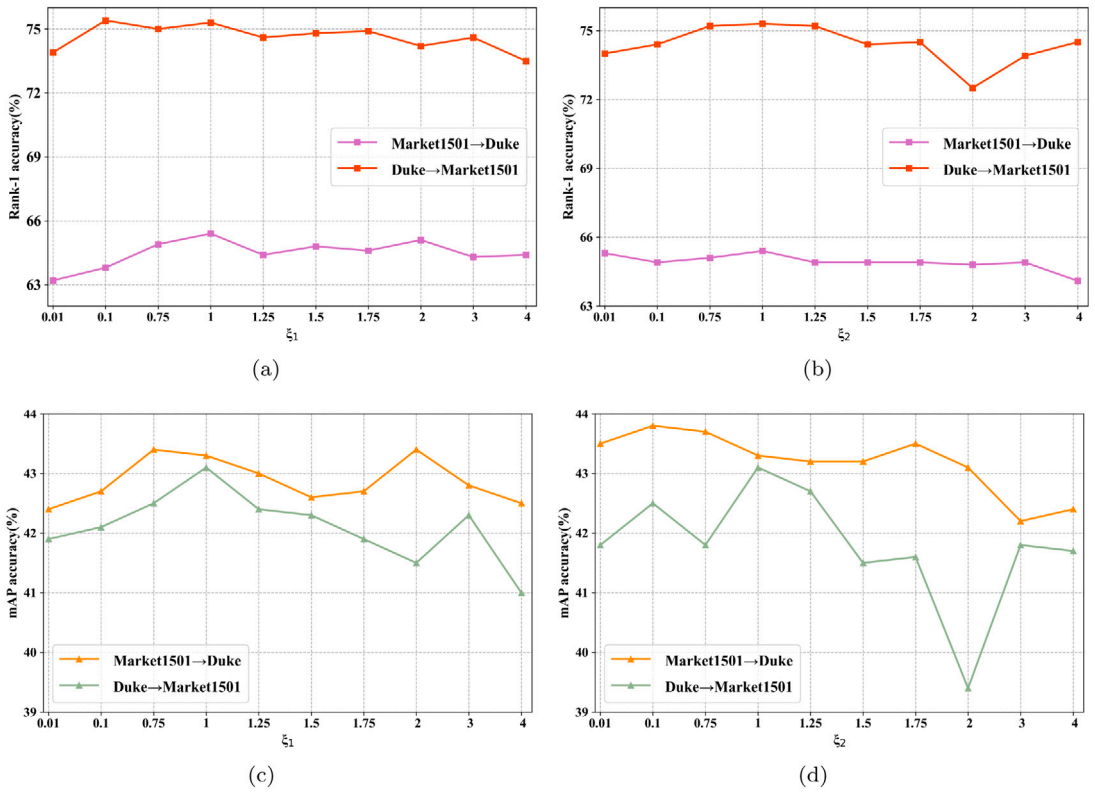

Fig. 8. Effect analysis on hyperparameters ξ_1 and ξ_2.

the developed method reaches the optimal performance at $\xi_1 = 1$, while on mAP, the developed method achieves the pleasing performance when $\xi_1 \in [0.75, 1]$. When $\xi_1 \in [1, 2]$, the recognition accuracy of Rank-1 and mAP dropped. Therefore, $\xi_1 = 1$ is a good choice.

The influence of ξ_2. In Eq. 12, the ξ_2 adjusts the role of L_{cameID} and L_{Ecam}. The ξ_1 is fixed, and the value of ξ_2 is selected within the range of $[0.01, 4]$. On Duke → Market1501 and Market1501 → Duke, the changes of Rank-1 and mAP with different values of ξ_2 are shown in Fig. 8(b) and (d). When $\xi_2 = 1$, the proposed method can obtain the best performance on both tasks. So, it is reasonable to set ξ_2 to 1.

5. Conclusion

This paper proposes a novel UDA person re-ID method, which consists of blockchain, MPL, DISFL, and AFDSF+MPF. In the proposed method, blockchain provides pedestrian image data security services. MPL prompts the networks to have a certain ability to predict and extract multi-view polymorphic features of pedestrians, and also plays a positive role in extracting the complete appearance features of pedestrians. The discriminative features of pedestrians are effectively extracted by the learning of domain-invariant salient features. After aligning and fusing with the multi-view polymorphic features, the description ability of features is effectively improved, which promotes the improvement of recognition performance. The proposed method requires neither any clustering and pseudo-label prediction nor any additional model assistance, so it has considerable practical significance. Comparative experimental results confirm its effectiveness and superiority to the similar state-of-the-art methods.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 61966021, Grant 2019FA-045, Sichuan Science and Technology Program under Grant 2021YFG0315.

References

[1] K. Navaneet, R.K. Sarvadevabhatla, S. Shekhar, R.V. Babu, A. Chakraborty, Operator-in-the-loop deep sequential multi-camera feature fusion for person re-identification, IEEE Trans. Inf. Forensics Secur. 15 (2019) 2375–2385.
[2] G. Qi, G. Hu, X. Wang, N. Mazur, Z. Zhu, M. Haner, EXAM: A Framework of learning extreme and moderate embeddings for person re-ID, J. Imaging 7 (1) (2021) 6.
[3] Q. Wang, W. Min, Q. Han, Z. Yang, X. Xiong, M. Zhu, H. Zhao, Viewpoint adaptation learning with cross-view distance metric for robust vehicle re-identification, Inform. Sci. 564 (2021) 71–84.
[4] Z. Zhu, Y. Luo, S. Chen, G. Qi, N. Mazur, C. Zhong, Q. Li, Camera style transformation with preserved self-similarity and domain-dissimilarity in unsupervised person re-identification, J. Vis. Commun. Image Represent. 80 (2021) 103303.
[5] A. Khatun, S. Denman, S. Sridharan, C. Fookes, End-to-end domain adaptive attention network for cross-domain person re-identification, IEEE Trans. Inf. Forensics Secur. 16 (2021) 3803–3813.
[6] M. Ye, J. Shen, L. Shao, Visible-infrared person re-identification via homogeneous augmented tri-modal learning, IEEE Trans. Inf. Forensics Secur. 16 (2021) 728–739.
[7] C.-X. Ren, B. Liang, P. Ge, Y. Zhai, Z. Lei, Domain adaptive person re-identification via camera style generation and label propagation, IEEE Trans. Inf. Forensics Secur. 15 (2019) 1290–1302.
[8] J. Wang, L. Yuan, H. Xu, G. Xie, X. Wen, Channel-exchanged feature representations for person re-identification, Inform. Sci. 562 (2021) 370–384.
[9] Y. Zhang, B. Ma, Y. Feng, M. Li, Pmt-net: Progressive multi-task network for one-shot person re-identification, Inform. Sci. 568 (2021) 133–146.
[10] C. Zhong, X. Jiang, G. Qi, Video-based person re-identification based on distributed cloud computing, J. Artif. Intell. Technol. 1 (2) (2021) 110–120.
[11] H. Li, N. Dong, Z. Yu, D. Tao, G. Qi, Triple adversarial learning and multi-view imaginative reasoning for unsupervised domain adaptation person re-identification, IEEE Trans. Circuits Syst. Video Technol. (2021).
[12] J. Pang, D. Zhang, H. Li, W. Liu, Z. Yu, Hazy re-ID: An interference suppression model for domain adaptation person re-identification under inclement weather condition, in: 2021 IEEE International Conference on Multimedia and Expo (ICME), IEEE, 2021, pp. 1–6.
[13] Y. Li, S. Chen, G. Qi, Z. Zhu, M. Haner, R. Cai, A GAN-based self-training framework for unsupervised domain adaptive person re-identification, J. Imaging 7 (4) (2021) 62.
[14] H. Li, Z. Kuang, Z. Yu, J. Luo, Structure alignment of attributes and visual features for cross-dataset person re-identification, Pattern Recognit. 106 (2020) 107414.
[15] H. Li, S. Yan, Z. Yu, D. Tao, Attribute-identity embedding and self-supervised learning for scalable person re-identification, IEEE Trans. Circuits Syst. Video Technol. 30 (10) (2019) 3472–3485.
[16] C. Zhong, G. Qi, N. Mazur, S. Banerjee, D. Malaviya, G. Hu, A domain adaptive person re-identification based on dual attention mechanism and camstyle transfer, Algorithms 14 (12) (2021) 361.
[17] J. Weng, J. Weng, J. Zhang, M. Li, Y. Zhang, W. Luo, Deepchain: Auditable and privacy-preserving deep learning with blockchain-based incentive, IEEE Trans. Dependable Secure Comput. 18 (5) (2019) 2438–2455.
[18] R. Belchior, A. Vasconcelos, S. Guerreiro, M. Correia, A survey on blockchain interoperability: Past, present, and future trends, ACM Comput. Surv. 54 (8) (2021) 1–41.
[19] B. Wang, Z. Li, Healthchain: A privacy protection system for medical data based on blockchain, Future Internet 13 (10) (2021) 247.
[20] U. Chelladurai, S. Pandian, A novel blockchain based electronic health record automation system for healthcare, J. Ambient Intell. Humaniz. Comput. 13 (1) (2022) 693–703.
[21] V.K. Chattu, et al., A review of artificial intelligence, big data, and blockchain technology applications in medicine and global health, Big Data Cogn. Comput. 5 (3) (2021) 41.
[22] C. Yang, L. Tan, N. Shi, B. Xu, Y. Cao, K. Yu, Authprivacychain: A blockchain-based access control framework with privacy protection in cloud, IEEE Access 8 (2020) 70604–70615.
[23] K. Gai, J. Guo, L. Zhu, S. Yu, Blockchain meets cloud computing: A survey, IEEE Commun. Surv. Tutor. 22 (3) (2020) 2009–2030.
[24] X. Liang, S. Shetty, D. Tosh, C. Kamhoua, K. Kwiat, L. Njilla, Provchain: A blockchain-based data provenance architecture in cloud environment with enhanced privacy and availability, in: 2017 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID), IEEE, 2017, pp. 468–477.
[25] Y. Zhai, S. Lu, Q. Ye, X. Shan, J. Chen, R. Ji, Y. Tian, AD-Cluster: Augmented discriminative clustering for domain adaptive person re-identification, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9021–9030.
[26] F. Yang, K. Li, Z. Zhong, Z. Luo, X. Sun, H. Cheng, X. Guo, F. Huang, R. Ji, S. Li, Asymmetric co-teaching for unsupervised cross-domain person re-identification, in: AAAI Conference on Artificial Intelligence(AAAI), 2020, pp. 12597–12604.
[27] X. Zhang, J. Cao, C. Shen, M. You, Self-training with progressive augmentation for unsupervised cross-domain person re-identification, in: IEEE International Conference on Computer Vision(ICCV), 2019, pp. 8222–8231.
[28] K. Zeng, M. Ning, Y. Wang, Y. Guo, Hierarchical clustering with hard-batch triplet loss for person re-identification, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13657–13665.
[29] X. Jin, C. Lan, W. Zeng, Z. Chen, Global distance-distributions separation for unsupervised person re-identification, in: European Conference on Computer Vision (ECCV), 2020, pp. 735–751.
[30] L. Wei, S. Zhang, W. Gao, Q. Tian, Person transfer gan to bridge domain gap for person re-identification, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 79–88.
[31] W. Deng, L. Zheng, Q. Ye, G. Kang, Y. Yang, J. Jiao, Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 994–1003.
[32] J. Liu, Z. Zha, D. Chen, R. Hong, M. Wang, Adaptative transfer network for cross-domain person re-identification, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7202–7211.
[33] Q. Yang, H. Yu, A. Wu, W. Zheng, Patch-based discriminative feature learning for unsupervised person re-identification, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3633–3642.
[34] L. Qi, L. Wang, J. Huo, L. Zhou, Y. Shi, Y. Gao, A novel unsupervised camera-aware domain adaptation framework for person re-identification, in: IEEE International Conference on Computer Vision(ICCV), 2019, pp. 8080–8089.
[35] J. Wang, X. Zhu, S. Gong, W. Li, Transferable joint attribute-identity deep learning for unsupervised person re-identification, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2275–2284.
[36] A. Wu, W.-S. Zheng, J.-H. Lai, Unsupervised person re-identification by camera-aware similarity consistency learning, in: IEEE International Conference on Computer Vision (ICCV), 2019, pp. 6922–6931.
[37] U.P.R. identification via Softened Similarity Learning, Yutian lin and lingxi xie and yu wu and chenggang yan and qi tian;, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3390–3399.
[38] Y. Huang, Z. Zha, X. Fu, R. Hong, L. Li, Real-world person re-identification via degradation invariance learing, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 14084–14094.

[39] Y. Ge, D. Chen, H. Li, Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identifiification, in: International Conference on Learning Representations (ICLR), 2020.
[40] F. Zhao, S. Liao, G.-S. Xie, J. Zhao, K. Zhang, L. Shao, Unsupervised domain adaptation with noise resistible mutual-training for person re-identification, in: European Conference on Computer Vision (ECCV), 2020, pp. 526–544.
[41] C. Luo, C. Song, Z. Zhang, Generalizing person re-identification by camera-aware invariance learning and cross-domain mixup, in: European Conference on Computer Vision (ECCV), 2020, pp. 224–241.
[42] Z. Ji, X. Zou, X. Lin, X. Liu, T. Huang, S. Wu, An attention-driven two-stage clustering method for unsupervised person re-identification, in: European Conference on Computer Vision (ECCV), 2020, pp. 20–36.
[43] Z. Zhong, L. Zheng, Z. Luo, S. Li, Y. Yang, Invariance matters: Exemplar memory for domain adaptive person re-identification, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 598–607.
[44] Y. Li, C. Lin, Y. Lin, Y.F. Wang, Cross-dataset person re-identification via unsupervised pose disentanglement and adaptation, in: IEEE International Conference on Computer Vision (ICCV), 2019, pp. 7919–7929.
[45] H. Tang, K. Jia, Discriminative adversarial domain adaptation, in: AAAI Conference on Artificial Intelligence (AAAI), 2020, pp. 5940–5947.
[46] Y. Du, Z. Tan, Q. Chen, X. Zhang, Y. Yao, C. Wang, Dual adversarial domain adaptation, ArXiv Preprint, 2020.
[47] K. Saito, K. Watanabe, Y. Ushiku, T. Harada, Maximum classifier discrepancy for unsupervised domain adaptation, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3723–3732.
[48] H. Zhang, M. Cisse, Y.N. Dauphin, D. Lopez-Paz, Mixup: Beyond empirical risk minimization, in: International Conference on Learning Representations (ICLR), 2018.
[49] V. Verma, A. Lamb, C. Beckham, A. Najafi, I. Mitliagkas, D. Lopez-Paz, Y. Bengio, Manifold mixup: Better representations by interpolating hidden states, in: International Conference on Machine Learning (ICML), 2019, pp. 6438–6447.
[50] H. Tang, Z. Li, Z. Peng, J. Tang, Blockmix: meta regularization and self-calibrated inference for metric-based meta-learning, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 610–618.
[51] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.
[52] J. Deng, W. Dong, R. Socher, L. Li, K. Li, F. Li, Imagenet: A large-scale hierarchical image database, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 248–255.
[53] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, Q. Tian, Scalabel person re-identification: A benchmark, in: IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1116–1124.
[54] E. Ristani, F. Solera, R. Zou, R. Cucchiara, C. Tomasi, Performance measures and a data set for multi-target, multi-camera tracking, in: European Conference on Computer Vision (ECCV), 2016, pp. 17–35.
[55] L. Wei, S. Zhang, W. Gao, Q. Tian, Person transfer gan to bridge domain gap for person re-identification, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 79–88.
[56] M. Hirzer, C. Beleznai, P.M. Roth, H. Bischof, Person re-identification by descriptive and discriminative classification, in: Scandinavian Conference on Image Analysis, 2011, pp. 91–102.
[57] C.C. Loy, C. Liu, S. Gong, Person re-identification by manifold ranking, in: IEEE International Conference on Image Processing (ICIP), 2013, pp. 3567–3571.
[58] H. Li, K. Xu, G. Lu, Y. Xu, J. Li, Z. Yu, D. Zhang, Dual-stream reciprocal disentanglement learning for domain adaption person re-identification, Arxiv, 2021.
[59] R. Zhao, W. Ouyang, X. Wang, Person re-identification by salience matching, in: ICCV, 2013, pp. 2528–2535.
[60] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, Q. Tian, Scalable person re-identification: A benchmark, in: ICCV, 2015, pp. 1116–1124.
[61] H. Luo, Y. Gu, X. Liao, S. Lai, W. Jiang, Bag of tricks and a strong baseline for deep person re-identification, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshop, 2019.
[62] X. Fan, W. Jiang, H. Luo, M. Fei, Spherereid: Deep hypersphere manifold embedding for person re-identification, J. Vis. Commun. Image Represent. 60 (2019) 51–58.
[63] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al., Pytorch: An imperative style, high-performance deep learning library, NIPS 32 (2019).
[64] Z. Zhong, L. Zheng, S. Li, Y. Yang, Generalizing a person retrieval model hetero-and homogeneously, in: European Conference on Computer Vision (ECCV), 2018, pp. 172–188.
[65] L. Song, C. Wang, L. Zhang, B. Du, Q. Zhang, C. Huang, X. Wang, Unsupervised domain adaptive re-identification: Theory and practice, Pattern Recognit. 102 (2020) 107173.
[66] Y. Zhai, Q. Ye, S. Lu, M. Jia, R. Ji, Y. Tian, Multiple expert brainstorming for domain adaptive person re-identification, in: European Conference on Computer Vision (ECCV), 2020, pp. 594–611.
[67] Y. Ge, D. Chen, F. Zhu, R. Zhao, H. Li, Self-paced contrastive learning with hybrid memory for domain adaptive object re-ID, in: NeurIPS, 2020.
[68] Y. Zou, X. Yang, Z. Yu, B.V. Kumar, J. Kautz, Joint disentangling and adaptation for cross-domain person re-identification, in: European Conference on Computer Vision (ECCV), 2020, pp. 87–104.
[69] Z. Zhong, L. Zheng, Z. Zheng, S. Li, Y. Yang, Camstyle: A novel data augmentation method for person re-identification, IEEE Trans. Image Process. 28 (3) (2018) 1176–1190.
[70] Y. Huang, Q. Wu, J. Xu, Y. Zhong, Sbsgan: Suppression of inter-domain background shift for person re-identification, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9527–9536.
[71] F. Yang, Z. Zhong, Z. Luo, S. Lian, S. Li, Leveraging virtual and real person for unsupervised person re-identification, IEEE Trans. Multimed. 22 (9) (2020) 2444–2453.
[72] Y. Yuan, W. Chen, T. Chen, Y. Yang, Z. Ren, Z. Wang, G. Hua, Calibrated domain-invariant learning for highly generalizable large scale re-identification, in: Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 3589–3598.
[73] Z. Zhuang, L. Wei, L. Xie, T. Zhang, H. Zhang, H. Wu, H. Ai, Q. Tian, Rethinking the distribution gap of person re-identification with camera-based batch normalization, in: European Conference on Computer Vision (ECCV), 2020, pp. 140–157.
[74] X. Jin, C. Lan, W. Zeng, Z. Chen, L. Zhang, Style normalization and restitution for generalizable person re-identification, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3143–3152.
[75] D. Wang, S. Zhang, Unsupervised person re-identification via multi-label classification, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10981–10990.
[76] H. Li, J. Pang, D. Tao, Z. Yu, Cross adversarial consistency self-prediction learning for unsupervised domain adaptation person re-identification, Inform. Sci. 559 (2021) 46–50.
[77] H. Li, Y. Chen, D. Tao, Z. Yu, G. Qi, Attribute-aligned domain-invariant feature learning for unsupervised domain adaptation person re-identification, IEEE Trans. Inf. Forensics Secur. 16 (2021) 1480–1495.
[78] P. Isola, J.-Y. Zhu, T. Zhou, A.A. Efros, Image-to-image translation with conditional adversarial networks, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1125–1134.
[79] T. Miyato, M. Koyama, Cgans with projection discriminator, in: International Conference on Learning Representations (ICLR), 2018.
[80] Z. Cao, T. Simon, S.-E. Wei, Y. Sheikh, Realtime multi-person 2d pose estimation using part affinity fields, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 7291–7299.