ABSTRACT  Person re-identification (re-ID) is the task of finding the matched person in a non-overlapping and multi-camera system. Because annotating images across multiple cameras is difficult and time-consuming, this paper focuses on fully unsupervised learning person re-ID that can learn person re-ID on unlabeled data. The unsupervised re-ID needs to self-generate pseudo labels to make training possible. Unlike human-annotated true labels, the pseudo labels contain noise labels which substantially hinder the network’s capability on feature learning. In order to refine the predicted pseudo labels, we introduce a novel unsupervised re-ID method named Multiple pseudo Labels Joint Training (MLJT) in this paper. Different from the existing works, the MLJT predicts multiple pseudo labels for each image by mining potential similarities in multiple ways. Based on invariance constraints among multiple pseudo labels, the MLJT is jointly optimized under the supervision of multiple pseudo labels to ease the impact of noises in the single pseudo label. The proposed MLJT predicts three types of pseudo labels for one input image. The first one is the clustering-based pseudo label. The second one is adaptive similarity measurement-based pseudo label. The third one is pseudo sub-labels which are predicted by mining channel-based self-similarities. The proposed MLJT has been extensively evaluated on two mainstream and public person re-ID datasets and outdoor real-world videos. Experiments demonstrate the effectiveness of the proposed multiple pseudo labels joint training strategy and the practicality of the proposed MLJT in real-world unsupervised person re-ID applications. The testing demo can be found at https://drive.google.com/drive/folders/1RvNaEiy6tF18_RcgTNcjiE7jI6eGy8sZL?usp=sharing.

INDEX TERMS  Person re-identification, fully unsupervised learning, pseudo label prediction.

I. INTRODUCTION

The person re-identification (re-ID) system aims to retrieve matched people from different cameras or different occasions. It has become an active research area in recent years due to the growing practicability of computer vision in the industry [1]–[4]. The person re-ID system has wide-ranging applications in security and surveillance systems, including human abnormality detection, the specific person searching, lost person tracking [4], [5]. The idea of person re-ID is to matching features of images, therefore extracting discriminative features is critical and challenging for the person re-ID.

Based on training strategy, current person re-ID methods can be summarized in three categories: supervised person re-ID [6]–[8], Unsupervised Domain Adaptive (UDA) based person re-ID [9]–[13], and the Fully Unsupervised Learning (FUL) based person re-ID [14]–[23]. The illustrations of these methods are shown in Fig. 1.

In past decades, person re-ID works mostly focused on supervised person re-ID. The supervised person re-ID methods [6]–[8] train the network on a labeled dataset. Annotating person identities for every image is required in supervised methods, as shown in Fig. 1(a). It is expensive to manually annotate identity across multiple cameras.

Therefore, some recent works focus on unsupervised person re-ID, which do not require labeled information in a target dataset. Because of the lack of labeled information,
for exploring channel-based self-similarity, multiple pseudo labels prediction, fine-tuning backbone network with computed joint losses.

Unlike predicting a single pseudo label for an input image in MLCReID, the proposed MLJT predicts multiple pseudo labels for the input image by mining potential similarities in different ways. The combination of multiple pseudo labels is more robust than a single pseudo label. According to invariance constraints among these predicted multiple pseudo labels, the MLJT optimizes the network via multiple pseudo labels in a joint manner. In general, there are three types of pseudo labels in MLJT, as shown in Fig. 2(b). The first one is the clustering-based pseudo label $y_i^c$. The second one is adaptive similarity measurement-based pseudo label $y_i^m$. The third one is pseudo sub-labels $y_{ig}$. We will introduce these three pseudo labels one by one in detail in Section III.

As a summary, the contributions of this work is four-fold:

1) Unlike predicting a single pseudo label in the baseline method, this work proposes to predict multiple pseudo labels for an input image by mining multiple potential similarities among samples.

2) Based on the invariance constraints among multiple pseudo labels, this work proposes to train the backbone network jointly to refine pseudo labels effectively.

3) In order to avoid neglecting positive labels of difficult samples, this paper proposes an Adaptive Similarity Measurement-based pseudo label Prediction (ASMP) method to adaptively select positive labels based on the degree of difficulty of samples.

4) This work improves the performance of baseline method [18] by considerable margins on two mainstream and public datasets, Market-1501 [26] (Market) and DukeMTMC-reID [27], [28] (Duke). Testing experiments are also performed in outdoor real-world videos to show the practicality of this work in real-world applications.

The remainder of this paper is organized as follows. Section II summarizes the related work. Section III describes the proposed method. In section IV, the datasets, evaluation metrics, and implementation details are introduced. In section V, extensive ablation studies are reported. The model performance in two public datasets and real-world applications are reported in Section VI. Finally, Section VII concludes this paper.

II. RELATED WORK

A. FUL-BASED RE-ID METHODS

Several FUL-based methods [14]–[18] were proposed, which trained the model without any manually annotated labels. Yu et al. [14] proposed to learn an unsupervised asymmetric metric based on asymmetric clustering on cross-view person images to deal with the view-specific interference. Followed the similar idea of [14], Yu et al. [15] embedded the asymmetric metric into a deep neural network to further improve the performance. Ding et al. [16] proposed an effec-
Q. Tang et al.: Fully Unsupervised Person Re-Identification via Multiple Pseudo Labels Joint Training

![Diagram of the framework of FUL person re-ID method. (a) Baseline method, MLCReID [18]. (b) The proposed MLJT. Given an unlabeled image, d-dimensional feature \( f_i \) is extracted by the backbone network. A feature memory stores feature for all images in a dataset. Before every training iteration, multiple pseudo labels can be predicted. The MLJT fine-tunes the backbone network in each training iteration step by step via the joint loss between multiple classification scores and multiple pseudo labels.]

(1) The baseline method, MLCReID

(2) The proposed MLJT

---

B. PSEUDO LABEL PREDICTION FOR UNSUPERVISED RE-ID

Based on the pseudo label prediction methods, unsupervised person re-ID can be generally divided into two categories: clustering-based label prediction and similarity measurement-based label prediction.

The core idea of clustering-based label prediction [12], [13], [19] is that they perform a clustering algorithm on Convolutional Neural Network (CNN) features to generate pseudo labels for training. Clustering-based methods pull samples close to their class centroids as illustrated in Fig. 3(a). Fan et al. [12] proposed a progressive unsupervised learning (PUL) method, which can be seen as an original clustering-based re-ID work. They iterated clustering and fine-tune CNN step by step until convergence. Because the clustering results may noisy, subsequent researches [13], [19], [20] mainly focus on refining noisy labels.

The core idea of similarity measurement-based label prediction [11], [18], [23] is that they estimate similarities...
among samples to select positive samples for training. Similarity measurement-based methods pull samples close to their near neighbors as illustrated in Fig. 3 (b). The existing methods selected positive samples based on some fixed rules. ECN [11] is the original similarity measurement-based re-ID method. ECN [11] and SSL [23] selected fixed rules. ECN [11] is the original similarity measurement-based method to select positive samples based on some existing methods selected positive samples for training. Subsequently, MLCReID [18] proposed to select positive samples using a pre-defined and fixed similarity threshold \( t \).

The existing researches only focus on one of the pseudo label prediction methods. As mentioned above, clustering-based and similarity measurement-based methods lead the model to learn in different directions, and thus they can be complementary to each other. This paper predicts clustering-based and similarity measurement-based pseudo labels for each image simultaneously to train the model jointly to achieve better performance, as shown in Fig. 3(c).

C. SIMILARITY EXPLORATION FOR UNSUPERVISED RE-ID
Exploring similarity correctly is critical to the re-ID performance. Supervised person re-ID methods can use additional human pose labels [6] or human body part segmentation regions [7] to help model learn human part similarity. However, these pre-annotated auxiliary labels are unknowable in the unsupervised learning task.

In order to boost model learning discriminative features without human part labels, several methods, including supervised method in [8], UDA-based unsupervised method Self-Similarity Grouping (SSG) [22], and FULL-based unsupervised method Softened Similarity Learning (SSL) [23] explored the self-defined part-based self-similarities, as shown in Fig. 4(a). These methods split \( P \times h \times w \) feature maps into \( P \) horizontal stripes to compute \( P \) numbers of part-based similarities. In SSG [22], \( P = 2 \) are used for exploring the potential similarities of upper human body and lower human body of unlabeled samples. They [8], [22], [23] ignore a factor that the human region does not always locate in the center of the image, as shown in Fig. 4(c). The human location variance makes the inconsistency in each partition. Moreover, data augmentation strategies, such as random crop, random erasing, and random rotation, further increases human location variance.

In order to avoid the impact of human location variance and mine potential similarities among images, we propose to explore channel-based self-similarity in this paper, as shown in Fig. 4(b).

III. PROPOSED METHOD
A. FRAMEWORK OVERVIEW

The framework of the proposed method MLJT is shown in Fig. 2(b), which can be mainly divided into five stages: feature extraction by backbone network, feature splitting using channel-wise matching, multiple pseudo labels prediction, classification, network optimization using joint losses.

Given an unlabeled person image \( x_{i[i]=1,2,...,N} \in \mathcal{X} \), \( d \)-dimensional feature \( f_i \) is extracted by the backbone network. \( i \) is the index of the image in an unlabeled person re-ID dataset \( \mathcal{X} \), \( N \) indicates the number of images in \( \mathcal{X} \). A feature memory \( \mathcal{M} \) is used to stores up-to-date features for all images in \( \mathcal{X} \) to make compute similarities among all images possible. The size of \( \mathcal{M} \) is \( N \times d \). \( \mathcal{M} \) is updated by \( f_i \) after every training iteration as,

\[
\mathcal{M}[i] = \alpha \mathcal{M}^t[i] + (1 - \alpha)f_i.
\]

The superscript \( t \) in Eq. (1) indicates \( t \)-th training iteration. \( \alpha \) is the updating rate, \( \alpha \in [0.0, 1.0] \). Updating with the moving averaged weights makes \( \mathcal{M} \) more stable to predict pseudo labels [11], [18]. After updating, the feature of \( x_i \) in \( \mathcal{M} \) are denoted as \( f_i^M \), where \( \mathcal{M}[i] = f_i^M \).

Then, multiple pseudo labels \( y_i = \{y_i^1, y_i^{sm}, y_i^	heta \} \) of \( x_i \) can be predicted. \( y_i^1 \) and \( y_i^{sm} \) are the clustering-based pseudo label and the adaptive similarity measurement-based pseudo label predicted using \( \mathcal{M} \), respectively. \( y_i^\theta \) are sub-labels predicted by mining channel-based self-similarities using sub-feature memories \( \mathcal{M}^\theta \), which are obtained by channel-wise matching, as shown in Fig. 2(b).

The backbone network is optimized progressively by the loss between multiple pseudo labels \( y_i = \{y_i^1, y_i^{sm}, y_i^\theta \} \) and their corresponding classification score \( v_i = \{v_i^1, v_i^{sm}, v_i^\theta \} \) in a joint training manner. The classification scores \( v_i^1 \) and \( v_i^{sm} \) are
computed using the whole feature $f_j$ and the sub-features $f_j^i$, respectively. In the following subsection, we will introduce these three pseudo labels one by one in detail.

**B. MULTIPLE PSEUDO LABELS PREDICTION**

1) **CLUSTERING-BASED PSEUDO LABEL $y_i^{cl}$**

   Following the previous works [19], [25], the unsupervised clustering algorithm DBSCAN [24] is used in this paper. The clustering results is illustrated in Fig. 3 (a). Before each training epoch, DBSCAN assigns pseudo-classes $c_i$ to all samples in $\mathcal{X}$ by computing distances among their features $f_i^j \in \mathcal{M}$. DBSCAN is a density-based clustering algorithm, it treats high-confident samples as clustered inliers ($c_i \geq 0$), and treats low-confident samples as unclustered outliers ($c_i = -1$).

   There is one challenge in the DBSCAN-based clustering method: the numbers of clustered inliers keep changing during the whole training process. Uncertain numbers of clusters make loss function is difficult to design. To address this issue, we encode 1-dimensional pseudo-class $c_i$ to $N$-dimensional clustering-based pseudo label $y_i^{cl}$ as follows,

   \[
   y_i^{cl}[j] = \begin{cases} 1 & c_j = c_i; \\ -1 & c_j \neq c_i; \end{cases} \quad j = 1, \ldots, N. \tag{2}
   \]

   After encoding, the multi-label classification loss function [18] can be easily adopted to compute the loss between $y_i^{cl}$ and $y_i^{cl}$ against the changing of clusters.

2) **ADAPTIVE SIMILARITY MEASUREMENT-BASED PSEUDO LABEL**

   The cosine similarity of $x_i$ and all images $x_{[j]=1,...,N]} \in \mathcal{X}$ is computed using their features in $\mathcal{M}$ as follows,

   \[
   s_i^j = \mathcal{M}[i] \times \mathcal{M}^T. \tag{3}
   \]

   According to the $s_i^j$, the adaptive similarity measurement-based pseudo label $y_{i,m}$ is predicted by our proposed Adaptive Similarity Measurement-based pseudo label Prediction (ASMP).

   As mentioned in Sec II.B, the previous pseudo label prediction methods selected positive labels [11], [18], [23] using fixed rules, i.e., a fixed $k$ number of positive labels in ECN [11] and SSL [23], or fixed label selection threshold $t$ in MLCReID [18]. Using fixed rules [11], [18], [23] to select positive labels is unsuitable, because the similarity distribution of every image is different and keeps changing during the whole training process. More specifically, difficult samples always share relatively low similarity with its $k$-Nearest Neighbors ($k$-NN) because of challenging situations, such as complex or uncommon human poses, occlusion, and complex backgrounds, etc. The examples of similarity distributions of image A (simple case) and image B (difficult case) are illustrated in Fig. 5. The similarities among image B and its $k$-NN are much lower than the similarities among image A and its $k$-NN, as shown in the right graph of Fig. 5. If positive samples are selected using the fixed threshold $t = 0.6$ for all images as [18], the potential positive labels of the difficult sample with similarity less than 0.6 might be neglected continuously.

   To avoid neglect the positive label of the difficult sample, we proposed the ASMP, which first distinguish the difficult sample based on the degree of difficulty $\theta_i$ of the sample, then assigns a lower and adaptive threshold to the difficult sample for selecting more positive labels. Our idea is that, when the model can roughly capture the data distribution and predict similarity, we can therefore utilize the obtained similarity distribution to estimate the degree of difficulty of the image.

   ASMP first computes the similarities among $x_i$ and its $k$-NN as follows,

   \[
   \begin{align*}
   R_i &= \text{argsort}(s_i^j) \\
   K_i &= R_i[1 : k]
   \end{align*}
   \]

   where $R_i$ is the rank list by sorting the similarity $s_i^j$ by descending order. $K_i$ is the collection of $k$-NN of $x_i$. $k$ is a hyper-parameter, controls the numbers of samples in $K_i$. The analysis of $k$ are reported in the Sec. V.

   The degree of difficulty $\theta_i$ is estimated according to statistical characteristics of $K_i$, i.e., the mean $m_i$ and standard deviation $\sigma_i$, as follows,

   \[
   \begin{align*}
   m_i &= \frac{\sum_{j \in k} s_i^j}{k} \\
   \sigma_i &= \sqrt{\frac{\sum_{j \in k} (s_i^j - m_i)^2}{k}} \\
   \theta_i &= m_i + \sigma_i
   \end{align*}
   \]

   If $\theta_i$ is low, $x_i$ is considered a difficult sample. The computed $\theta_i$ will be directly used to select more positive labels for $x_i$ as follows,

   \[
   y_{i,m}[j] = \begin{cases} 1, & s_i^j \geq \min(0.6, \theta_i) \\ -1, & \text{otherwise} \end{cases} \quad j = 1, \ldots, n. \tag{7}
   \]

   In this case, more difficult samples can be assigned a lower $\theta_i$ to mitigate the neglect to potential positive labels. Finally, adaptive similarity measurement-based label $y_{i,m}$ are generated. It is noteworthy that $\theta_i$ for every sample is different according to statistical characteristics of the sample, therefore ASMP is better than fixed rules in [11], [18], [23].
3) PSEUDO SUB-LABELS BY CHANNEL-BASED SELF-SIMILARITY EXPLORATION (CSS)

The third pseudo labels in this paper is pseudo sub-labels $y_i^s$, which is predicted by mining channel-based self-similarities.

The baseline method MLCReID [18] only compared the similarity between two images by global features. To mine more similarity information existing in the unlabeled dataset, several methods [22] split images into horizontal parts to represent human partial regions, as illustrated in Fig. 4(a). Then, unlabeled images are additionally compared using these partial features. The precondition of part-based self-similarity computation is that the human region should always locate in the center of the image. Uncertain human location may cause inconsistency in similarity mining and matching, as we mentioned in Sec II.C.

The critical idea of self-similarity computation is to compute a more robust similarity score by additionally comparing partial features. In order to avoid the impact of human location variance and mine the potential similarity as well, we propose to explore channel-based self-similarity in this paper, as shown in Fig. 4(b).

To formulate the proposed channel-based self-similarity computation, a channel-wise matching module is proposed in this paper. The channel-wise matching module is attached after feature computation, a channel-wise matching module is proposed in this paper, as shown in Fig. 4(b).

Then, unlabeled images are additionally compared using these partial features. The precondition of part-based self-similarity computation is that the human region should always locate in the center of the image. Uncertain human location may cause inconsistency in similarity mining and matching, as we mentioned in Sec II.C.

The classification scores $v_i^g = f_i^g \times (\mathcal{M}^g)^\top$, $g = 1, \ldots, G$. (8)

The $G$ groups of self-similarities of $x_i$ are computed using their corresponding sub-feature memories $\mathcal{M}^g$ as follows,

$$s_i^g = \mathcal{M}^g [i] \times (\mathcal{M}^g)^\top, \quad g = 1, \ldots, G.$$ (9)

Based on $s_i^g$, the corresponding pseudo sub-labels $y_i^g$ can be predicted using the proposed ASMP as Eq. (4)-(7) as follows,

$$y_i^g = \text{ASMP}(s_i^g, k), \quad g = 1, \ldots, G.$$ (10)

C. THE JOINT LOSS FUNCTION

The Memory-based Multi-label Classification Loss (MMCL) [18] $\mathcal{L}^*$ is used to regress classification scores $v_i$ to predicted pseudo label $y_i$ as follows:

$$\mathcal{L}^*(v_i, y_i) = \frac{\delta}{|P_i|} \sum_{v_i[j] \in P_i} (v_i[j] - y_i[j]) + \frac{1}{|N_i|} \sum_{v_i[k] \in N_i} (v_i[k] - y_i[k])$$ (11)

$P_i$ indicates the positive samples of $x_i$, of where $y_i[j] = 1$. $N_i$ indicates the hard negative samples of $x_i$, of where $y_i[k] = -1$. To solve the sparsity of $y_i$, only top 1% negative samples are chosen as hard negative samples to compute $\mathcal{L}^*$. $\delta$ is the balance parameter between positive hard negative sample loss. Following same setting in MLCReID [18], $\delta = 5$ in our paper.

As a summary, for an unlabeled image $x_i$, three types of pseudo labels $y_i = \{y_i^{CL}, y_i^{sm}, y_i^s\}$ can be predicted by Eq. (2), Eq.(7), and Eq.(10). The MLJT is trained using the losses between multiple pseudo labels $y_i = \{y_i^{CL}, y_i^{sm}, y_i^s\}$ and their corresponding classification scores $v_i = \{v_i^0, v_i^1\}$ in an end-to-end and joint manner. The overall joint loss function is formulated as follows,

$$\mathcal{L} = \mathcal{L}^*(v_i^0, y_i^{CL}) + \mathcal{L}^*(v_i^0, y_i^{sm}) + \sum_{g=1}^{G} \mathcal{L}^*(v_i^g, y_i^g).$$ (12)

IV. EXPERIMENT SETTINGS

A. DATASETS AND EVALUATION METRICS

We evaluate the proposed method on the two large-scale and mainstream person re-ID datasets. Every person re-ID datasets consist of three subsets, including training set, query set, and gallery set. The number of identities (IDs) and images in the two datasets are reported in Table 1. The training set is used for training, and both the query and gallery sets are used for testing. The identities of the training set are different from the query and gallery sets.

Market-1501 [26] (Market) has six cameras and 32,668 person images of 1,501 identities in total. 751 identities with 12,936 images are used for training, and 750 identities with 19,732 images are used for testing.

DukeMTMC-reID [27], [28] (Duke) has eight cameras and 36,411 person images of 1,404 identities in total. 750 identities with 16,522 images are used for training, and 702 identities with 19,889 images are used for testing.

Two evaluation metrics are used to measure model performance. The first one is Mean Average Precision (mAP) (%), which is the average value of Average Precision (AP) over all query images. AP is the area under the precision-recall curve. Another evaluation metric is the Cumulative Matching Characteristic (CMC) curve. The CMCs (%) of Rank-1 (R-1), Rank-5 (R-1), and Rank-10 (R-1) are reported, which represents the probability of top-1, top-5, and top-10 ranked gallery samples containing the query identity, respectively.

| Dataset          | Person IDs | Images |
|------------------|------------|--------|
| Market-1501      | T | Q | G | T | Q | G |
| DukeMTMC-reID    | T | Q | G | T | Q | G |

165125
Therefore, the model still cannot distinguish human identities. More accurate pseudo labels help the model learn to extract more discriminative features to re-identify persons.

A. PERFORMANCE OF THE BASELINE METHOD
The No. 1 in Table 2 and Fig. 6 (b) show the results of the baseline method, which does not use clustering-based pseudo labels $y_i^{cl}$, adaptive threshold $\theta_i = 0.6$ in Eq. (7), and sub-labels $y_i^{cl}$ by channel-based self-similarity exploration. In Table 2, the baseline method No. 1 produces unsatisfactory performance on two datasets. For instance, it achieves 43.0% in mAP and 78.9% in R-1 on Market-1501, and 37.4% in mAP and 63.5% in R-1 on DukeMTMC-reID. It shows that using a single pseudo label is not enough to provide robust information. As shown in Fig. 6 (b), the model still cannot produce discriminative features to distinguish difficult samples.

To improve the performance of the baseline method, this paper proposes three types of pseudo labels in Sec III. We further perform experiments to investigate the effectiveness of them in detail by adding them into the baseline model. The results are reported in Table 2 and Fig. 6.

B. THE EFFECTIVENESS OF THE CLUSTERING-BASED PSEUDO LABEL
The effectiveness of the clustering-based pseudo label $y_i^{cl}$ is presented here. From the comparison of No. 1 and No. 2 in Table 2, it is clear that the model performance is improved with the clustering-based pseudo label $y_i^{cl}$. Specifically, the performance improves in mAP from 43.0% to 50.9% and from 37.4% to 40.3% on Market-1501 and DukeMTMC-reID, respectively. The significant improvement is because samples are additionally enforced learning towards their corresponding class centroids under the supervision of clustering labels $y_i^{cl}$, as illustrated in Fig. 3(c). Similar observation also can be found from the comparison of Fig. 6(b) and Fig. 6(e). Features of the same identity are more compact and independent from other clusters by using $y_i^{cl}$. The improvement demonstrates the effectiveness of combining similarity measurement-based and clustering-based label prediction in this paper.

C. THE ANALYSIS OF ADAPTIVE SIMILARITY MEASUREMENT-BASED PSEUDO LABEL PREDICTION (ASMP)

The ASMP is analyzed in three aspects. First, the effectiveness of ASMP is evaluated in Table 2 and Fig. 6. Second, the robustness of hyper-parameter $k$ in Eq.(4) is analyzed in Table 3. Third, we compare our proposed ASMP with other label prediction methods in Table 4.
Q. Tang et al.: Fully Unsupervised Person Re-Identification via Multiple Pseudo Labels Joint Training

1) THE EFFECTIVENESS OF ASMP

ASMP is proposed to adaptively select positive labels for a sample according to the similarity distribution between the sample and its neighbors. If a sample shares low similarities with neighbors, ASMP will consider the sample as the difficult sample and compute a low threshold to choose more positive labels for it. Comparison results between the model trained with or without ASMP are illustrated in Table 2 and Fig. 6. In Table 2, No. 3 surpasses the baseline No. 1 by 6.1% in mAP and 2.7% in R-1 accuracy on Market-1501, and by 3.4% in mAP and 2.3% in R-1 accuracy on DukeMTMC-reID. Similarly, when we compare Fig. 6(b) and Fig. 6(d), we observe that the model with ASMP can gather features of the same identity and enlarges the distances among different identities. It is because that using a fixed and high threshold in the baseline method makes dispersed samples (e.g. blue and orange samples in Fig. 6(b)) lack positive supervisory signals for training, thereby leading them to be neglected. The results demonstrate the necessity of setting a lower threshold to choose more positive labels for difficult samples in unsupervised person re-ID tasks and the effectiveness of our proposed ASMP.

2) COMPARISON OF DIFFERENT k IN ASMP

The hyper-parameter $k$ controls the numbers of near neighbors are chosen to build $K_i$ to estimate the degree of difficulty of the sample in Eq.(4). We validate the influence of different $k$ in Table 3 by vary $k$ from 20 to 160. Compared with baseline results in Table 2, it is clear that using any $k$ enhances model performance consistently, which demonstrates the necessity of adaptive thresholds in ASMP. When $k = 80$, the optimal performance can be obtained. Too small $k$ decreases the performance because the limited number of near neighbors is not enough to represent the statistical characteristics of similarity distribution of the sample for estimating its degree of difficulty $\theta_i$. Also, too large $k$ causes similar statistical characteristics of simple and difficult images, thereby distinguishing them difficultly.

3) COMPARISON WITH DIFFERENT LABEL PREDICTION METHODS

We further compare the proposed ASMP with other label prediction methods, i.e., a fixed $k$ number of positive labels in ECN [11] and SSL [23], and a fixed label selection threshold $t$ in MLReID [18] in Table 4. Table 4 shows that using the fixed threshold $t$ ($t = 0.6$, which achieves the best performance) outperforms using the fixed numbers $k$ ($k = 10$, which achieves the best performance) by large margins. Our proposed adaptive threshold $\theta_i$ in ASMP achieves the best performance. The superior performance demonstrates that, compared with fix rules in existing works, our proposed adaptive threshold is a more reasonable and effective method for pseudo label prediction.

D. THE ANALYSIS OF CHANNEL-BASED SELF-SIMILARITY (CSS)

The CSS is proposed to predict pseudo sub-labels $y^g_i$ by exploring a more robust and precise similarity relation among images. The CSS is analyzed in three aspects here. First, the effectiveness of CSS is evaluated in Table 2 and Fig. 6. Second, different splitting methods and different numbers of splitting groups in CSS on Market-1501 are compared with our proposed CSS. Third, we compare our proposed channel-based self-similarity with part-based self-similarity in existing works [22], [23] in Table 6.

1) THE EFFECTIVENESS OF CSS

As reported in Table 2, No. 4 surpasses the baseline No. 1 by 3.4% in mAP and 1.5% in R-1 accuracy on

### TABLE 2. Ablation study on individual proposed modules in the proposed MLJT.

| No. | Proposed Modules | Market-1501 | DukeMTMC-reID |
|-----|-----------------|-------------|---------------|
|     | Clustering ASMP CSS | mAP R-1 K-5 | mAP R-1 K-5 |
| 1   |                | 43.0 78.9 | 37.4 63.5 |
| 2   | ✔              | 50.9 81.9  | 40.3 64.5 |
| 3   | ✔ ✔            | 49.1 81.6  | 40.8 65.8 |
| 4   | ✔ ✔ ✔          | 46.4 80.4  | 40.2 65.4 |
| 5   | ✔ ✔ ✔ ✔        | 51.4 79.7  | 41.3 64.7 |
| 6   | ✔ ✔ ✔ ✔ ✔      | 52.3 81.2  | 41.7 65.4 |
| 7   | ✔ ✔ ✔ ✔ ✔ ✔    | 50.1 80.8  | 41.4 65.3 |
| 8   | (Full)         | 55.3 81.6  | 42.9 66.3 |

### TABLE 3. Analysis of hyper-parameters $k$ on Market-1501.

| $k$ | 20 | 40 | 60 | 80 | 100 | 120 | 140 | 160 |
|-----|----|----|----|----|-----|-----|-----|-----|
| mAP | 44.8 | 48.1 | 48.3 | 49.1 | 49.3 | 48.3 | 46.9 | 46.8 |
| R-1 | 80.0 | 82.0 | 81.3 | 81.6 | 80.7 | 80.1 | 79.6 | 78.6 |

### TABLE 4. Comparison with different pseudo label prediction methods.

| Method | Market | Duke |
|--------|--------|------|
| Fixed numbers $k$ [11], [23] | 36.7 72.4 | 36.2 62.2 |
| Fixed threshold $t$ [18] | 43.0 78.9 | 37.4 63.5 |
| Adaptive threshold $\theta_i$ (our ASMP) | 49.1 81.6 | 40.8 65.8 |

### TABLE 5. Comparison of different splitting methods and number of splitting groups in CSS on Market-1501.

| Splitting Method | Groups | mAP | Rank-1 | Rank-5 | Rank-10 |
|-----------------|--------|-----|--------|--------|---------|
| In order        | G=2    | 41.3 | 75.3   | 87.4   | 90.8    |
|                 | G=4    | 44.8 | 78.1   | 88.9   | 92.2    |
|                 | G=8    | 46.0 | 80.2   | 89.5   | 92.3    |
| Shuffle         | G=2    | 40.8 | 75.5   | 87.1   | 90.7    |
|                 | G=4    | 45.7 | 78.6   | 89.3   | 92.4    |
|                 | G=8    | 46.4 | 80.4   | 89.6   | 92.7    |
TABLE 6. Comparison with different similarity exploration methods.

| Methods                     | Market-1501 | DukeMTMC-reID |
|-----------------------------|-------------|---------------|
|                            | mAP         | R-1           | mAP         | R-1           |
| w/o self-similarity         | 43.0        | 78.9          | 37.4        | 63.5          |
| w/ Part-based self-similarity | 41.9        | 74.6          | 35.1        | 61.3          |
| w/ Channel-based self-similarity (ours) | 46.4        | 80.4          | 40.2        | 65.4          |

FIGURE 7. The visualization results of feature maps with different similarity exploration methods.

Market-1501, and by 2.8% in mAP and 1.9% in R-1 accuracy on DukeMTMC-reID. As shown in Fig. 6(e), the model can distinguish different identities better than the baseline by mining similarities from global feature to channel-wise partial feature. The improvements and visualization results demonstrate that the proposed CSS is a simple and effective method for mining self-similarity in an unsupervised manner.

2) COMPARISON OF DIFFERENT SPLITTING METHODS AND DIFFERENT G

In channel-wise matching, the feature $f_i$ and the feature memory $M$ can be split into $G$ groups sub-features $f^g_i$ and sub-feature memories $M^g$ in order or shuffle, respectively. Table 5 reports the performance comparisons of these two splitting methods with different splitting groups $G$. The results show that no significant difference in performance between splitting in order and shuffle. It is clear that a large $G$ consistently performs better than a small $G$. When we set $G = 2$, we observe that small $G$ slows down the model convergence speed. Thus, the model is difficult to converge to achieve satisfactory performance even if we set the training epoch very high. Finally, we split the feature memory in a shuffling manner and use $G = 8$.

3) COMPARISON BETWEEN PART-BASED AND CHANNEL-BASED SELF-SIMILARITY

We compare the proposed channel-based self-similarity with part-based self-similarity in [22], [23] in Table 6. For a straightforward comparison, we further visualize the hot maps of feature maps before the last GAP layer of the backbone network in Fig. 7. Fig. 7(a) illustrates six different input images with three different identities. The original height and width of the feature maps are $w_f \times h_f \times d : 8 \times 16 \times 2048$, as mentioned in Fig. 3. We resize them to the same size as input images $W \times H : 128 \times 256$ to more straightforwardly comparisons.

Table 6 reports that using part-based self-similarity drops the model performance from the baseline 43.0% to 41.9% in mAP and 78.9% to 74.6% in R-1 accuracy on Market-1501. Similar, the performance declines are observed on DukeMTMC-reID. It shows that part-based self-similarity method is not robust. The same situations are observed by visualization examples in Fig. 7(c). Mining part-based self-similarity makes the model can not accurately capture features from the human region, especially if the identity does not locate in the center of the image.

Brighter color means the model extracts more features from the regions.

TABLE 7. Performance comparison with state-of-the-art FUL-based methods. The first and second best results are marked in bold and blue, respectively.

| Method     | Market-1501 | DukeMTMC-reID |
|------------|-------------|---------------|
|            | mAP         | R-1           | mAP         | R-1           |
|            | R-5         | R-10          | R-5         | R-10          |
| BoW [12]   | 14.8        | 53.8          | 52.4        | 60.3          |
| UDML [34]  | 12.4        | 34.5          | 52.6        | 59.6          |
| LOMO [33]  | 18.0        | 27.2          | 41.6        | 49.1          |
| CAMEL [14] | 26.3        | 54.5          | -           | -             |
| DECAMEL [15]| 32.4        | 60.2          | 76.0        | 81.1          |
| SSL [23]   | 37.8        | 71.7          | 83.8        | 87.8          |
| BUC [17]   | 38.3        | 66.2          | 79.6        | 84.5          |
| ADTC [20]  | 38.8        | 59.5          | 71.6        | 76.9          |
| DBC [16]   | 41.3        | 69.2          | 83.0        | 87.8          |
| MLCReID [18]| 45.5        | 80.3          | 89.4        | 92.3          |
| NNC1 [21]  | 48.4        | 82.0          | 90.0        | 92.9          |
| DSCE [19]  | 53.9        | 74.8          | -           | -             |
| MLJT (Ours) | 55.3        | 81.6          | 90.4        | 93.5          |
|            | 42.9        | 66.3          | 77.4        | 80.7          |

FIGURE 8. Examples of the top-10 person re-ID results. The first and second rows are generated by the baseline method [18] and the proposed MLJT, respectively. The green boxes denote the true positive results, and the red boxes denote the false positive results.
The proposed channel-based self-similarity can help the model learn to extract more discriminative features, as compared in Fig. 7. The superior performance of the proposed CSS are mainly reflected in four aspects. Firstly, Table 6 reports that using CSS enhance the model performance consistently on Market-1501 and DukeMTMC-reID. Secondly, compared with Fig. 7(b), features of the foreground (human region) are more accurate and brighter in Fig. 7(d). Thirdly, compared with Fig. 7(c), the background area is darker (lower importance) in Fig. 7(d). Fourthly, feature extraction capability of the model is not affected by human location. The visualization results demonstrate that the proposed channel-based self-similarity helps the model extract discriminative features and avoid the impact of human location variance as well.

E. THE ANALYSIS OF JOINT TRAINING STRATEGY

The No. 5-No. 8 in Table 2 shows that combinations of each individual proposed module bring greater improvements. Finally, the full version of the proposed MLJT achieves the best performance 55.3% in mAP and 81.6% in R-1 accuracy on Market-1501, and 42.9% in mAP and 66.3% in R-1 accuracy on DukeMTMC-reID. The t-SNE features generated by the full version of the proposed MLJT are shown in Fig. 6(f). It shows that joint training strategy can overcome demerits and utilize merits of each individual module. Specifically, compared with Fig. 3(c), the clustering errors are eased in Fig. 3(f). Compared with Fig. 3(d) and Fig. 3(e), Fig. 3(f) shows that the model with joint training strategy produces more compact feature clusters. These verify that our proposed multiple pseudo labels joint training strategy is able to fully utilize each individual module for learning better and discriminative features.

VI. PERFORMANCE

A. PERFORMANCE COMPARISON IN PUBLIC DATASETS

We compare our proposed MLJT against state-of-the-art unsupervised person re-ID models in Market-1501 and DukeMTMC-reID in Table 7. We first compare our method MLJT with the hand-crafted feature-based methods, including BoW [32], UDML [34], and LOMO [33], which do not require CNN and any labeled dataset to extract features. The performances of the hand-crafted feature-based methods are not satisfactory enough, because it is difficult to manually design discriminative features with good generalization and robustness, especially on large-scale datasets.

We also compare the MLJT with CNN-based fully unsupervised learning-based methods, including CAMEL [14], DECAMEL [15], SSL [23], BUC [17], ADTC [20], DBC [16], the baseline method MLCReID [18], the best similarity measurement-based method NNCT [21], and the best clustering-based method DSCE [21]. We summarize three observations. 1) The clustering-based methods BUC, DBC, ADTC and DSCE can achieve better clustering accuracy (in mAP) because they assigned the same labels to the samples in the same cluster. During training, intra-class samples are enforced learning towards their class centroids to force these samples to get more compacter. 2) The similarity measurement-based methods SSL, MLCReID, and NNCT can achieve higher ranking accuracy (in R-k) because they assign labels by mining reliable positive neighbors around the sample. The similarity measurement-based methods enforce learning towards reliable neighbors. 3) our method MLJT
outperforms other FUL-based person re-ID methods in mAP and R-k accuracy. The superior performance demonstrate the effectiveness of our proposed multiple pseudo labels joint training strategy for unsupervised person re-ID.

Moreover, we compare the results of the baseline method [18] and the proposed method MLJT by showing their top-10 retrieved images of three query images in Fig. 8. The green and red boundaries denote correct and false re-ID results, respectively. As illustrated, for the same query images, the MLJT can retrieve correct images more accurately than the baseline. For example, the baseline is easily confused by the cloth with white and purple stripes. Overall, the comprehensive comparison results indicate the effectiveness and superiority of the MLJT.

B. PERFORMANCE IN REAL-WORLD APPLICATION

Finally, we test the proposed re-ID method on two outdoor real-world videos to evaluate the model performance in a real-world application. The full testing demo can be found at https://drive.google.com/drive/folders/1RvNaEiy6fIF18_RcgTNCjE7j6eGy8zKL?usp=sharing.

In a real-world application, YOLOv5 [35] is adopted to detect human regions because person regions are a prerequisite for person re-ID. Then, the re-ID network retrieves the same person by matching a pre-defined query image with every detected person region in the frame.

The visualization of re-ID results of the baseline method [18] and our proposed MLJT on two outdoor real-world videos are compared in Fig. 9. The query image (re-ID target) is shown at the left bottom of each frame. From the comparison, we notice that MLJT retrieves correct persons more accurately than the baseline, and MLJT outputs more confident (higher) classification scores than the baseline in two videos.

The runtime of the person detection and re-identification system is shown at the left top of each frame, notated as Frame Per Second (FPS). The system can work in real-time with a processing speed of about 128 FPS with YOLOv5 [35] detector.

VII. CONCLUSION

In this work, we proposed an end-to-end fully unsupervised person re-ID method, which can be trained without using any labeled information. The proposed method achieves superior performance benefit from three aspects. 1) Selecting positive labels adaptively according to similarity distribution of samples. 2) Estimating similarity precisely by the channel-based self-similarities exploration strategies. 3) Optimizing network jointly using multiple pseudo labels to mitigate the impact of noises in a single pseudo label. Extensive experiments and comprehensive analysis demonstrate the effectiveness of the proposed method MLJT. In the future, we will integrate the proposed algorithm in high-level video surveillance tasks like the specific person searching and abnormal human action recognition.

REFERENCES

[1] A. Shahbaz and K.-H. Jo, “Enhanced unsupervised change detector for industrial surveillance systems,” IEEE Trans. Ind. Electron., vol. 68, no. 9, pp. 8973–8981, Sep. 2021.

[2] J. Yang, B. Jiang, B. Li, K. Tian, and Z. Lv, “A fast image retrieval method designed for network big data,” IEEE Trans. Ind. Informat., vol. 13, no. 5, pp. 2350–2359, Oct. 2017.

[3] A. Feizi, “High-level feature extraction for classification and person re-identification,” IEEE Sensors J., vol. 17, no. 21, pp. 7064–7073, Nov. 2017.

[4] D. Li, Z. Zhang, X. Chen, and K. Huang, “A richly annotated pedestrian dataset for person retrieval in real surveillance scenarios,” IEEE Trans. Image Process., vol. 28, no. 4, pp. 1575–1590, Apr. 2019.

[5] 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,” IEEE Trans. Pattern Anal. Mach. Intell., early access, Jan. 26, 2021, doi: 10.1109/TPAMI.2021.3054775.

[6] Q. Zhou, B. Zhong, X. Lan, G. Sun, Y. Zhang, B. Zhang, and R. Ji, “Fine-grained spatial alignment model for person re-identification with focal triplet loss,” IEEE Trans. Image Process., vol. 29, pp. 7578–7589, 2020.

[7] H. Huang, W. Yang, J. Lin, G. Huang, J. Xu, G. Wang, X. Chen, and K. Huang, “Improve person re-identification with part awareness learning,” IEEE Trans. Image Process., vol. 29, pp. 7468–7481, 2020.

[8] Y. Fu, Y. Wei, Y. Zhou, H. Shi, G. Huang, X. Wang, Z. Yao, and T. Huang, “Horizontal pyramid matching for person re-identification,” in Proc. AAAI, 2019, pp. 8295–8302.

[9] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang, “Camera style adaptation for person re-identification,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 5157–5166.

[10] Z. Zhong, L. Zheng, S. Li, and Y. Yang, “Generalizing a person retrieval model hetero- and homogeneously,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2018, pp. 172–188.

[11] Z. Zhong, L. Zheng, Z. Luo, S. Li, and Y. Yang, “Invariance matters: Exemplar memory for domain adaptive person re-identification,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 598–607.

[12] H. Fan, L. Zheng, C. Yan, and Y. Yang, “Unsupervised person re-identification: Clustering and fine-tuning,” ACM Trans. Multimedia Comput., Commun., Appl., vol. 14, no. 4, pp. 1–18, 2018.

[13] Y. Ge, D. Chen, and H. Li, “Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification,” in Proc. Int. Conf. Learn. Represent. (ICLR), 2020. [Online]. Available: https://openreview.net/forum?id=JlnOhVYPS

[14] H.-X. Yu, A. Wu, and W.-S. Zheng, “Cross-view asymmetric metric learning for unsupervised person re-identification,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 994–1002.

[15] H.-X. Yu, A. Wu, and W.-S. Zheng, “Unsupervised person re-identification by deep asymmetric metric embedding,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 42, no. 4, pp. 956–973, Apr. 2020.

[16] G. Ding, S. Khan, Q. Yin, and Z. Tang, “Dispersion based clustering for unsupervised person re-identification,” in Proc. BMVC, 2019, pp. 1–12.

[17] Y. Lin, X. Dong, L. Zheng, Y. Yan, and Y. Yang, “A bottom-up clustering approach to unsupervised person re-identification,” in Proc. AAAI Conf. Artif. Intell., 2019, pp. 8738–8745.

[18] D. Wang and S. Zhang, “Unsupervised person re-identification via multi-label classification,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 10978–10987.

[19] F. Yang, Z. Zhong, Z. Luo, Y. Cai, Y. Lin, S. Li, and N. Sebe, “Joint noise-tolerant learning and meta camera shift adaptation for unsupervised person re-identification,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 4853–4862.

[20] Z. Ji, X. Zou, X. Lin, X. Liu, T. Huang, and S. Wu, “An attention-driven two-stage clustering method for unsupervised person re-identification,” in Proc. ECCV, 2020, pp. 20–36.

[21] Q. Tang and K.-H. Jo, “Unsupervised person re-identification via nearest neighbor collaborative training strategy,” in Proc. IEEE Int. Conf. Image Process. (ICIP), Sep. 2021, pp. 1139–1143.

[22] Y. Fu, Y. Wei, G. Wang, Y. Zhou, H. Shi, U. Uluç, and T. Huang, “Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 6111–6120.

[23] Y. Lin, L. Xie, Y. Wu, C. Yan, and Q. Tian, “Unsupervised person re-identification via softened similarity learning,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 3387–3396.
[24] M. Ester, H. P. Kriegel, J. Sander, and X. Xu, “A density-based algorithm for discovering clusters in large spatial databases with noise,” in Proc. KDD, 1996, pp. 226–231.

[25] Z. Zhong, L. Zheng, D. Cao, and S. Li, “Re-ranking person re-identification with K-reciprocal encoding,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 1318–1327.

[26] L. Zheng, L. Shen, L. Tian, S. Wang, J. Bu, and Q. Tian, “Person re-identification meets image search,” 2015, arXiv:1502.02171.

[27] Z. Zheng, L. Zheng, and Y. Yang, “Unlabeled samples generated by GAN improve the person re-identification baseline in vitro,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Oct. 2017, pp. 3774–3782.

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

[29] 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.

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

[31] L. Van der Maaten and G. Hinton, “Visualizing data using t-SNE,” J. Mach. Learn. Res., vol. 9, pp. 2579–2605, Nov. 2008.

[32] 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.

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

[34] P. Peng, T. Xiang, Y. Wang, M. Pontil, S. Gong, T. Huang, and Y. Tian, “Unsupervised cross-dataset transfer learning for person re-identification,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 1306–1315.

[35] G. Jocher, Ultralytics/YOLOv5: V5.0-YOLOv5-P6 1280 Models, AWS, Supervise.ly and YouTube integrations, Seattle, WA, USA, Apr. 2021.