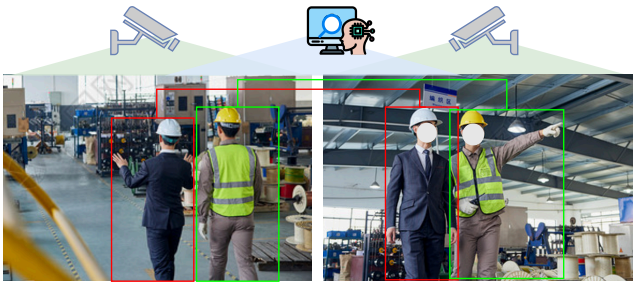
Cross-Camera Discriminative Person Association by Unsupervised Frame Clustering and Selection

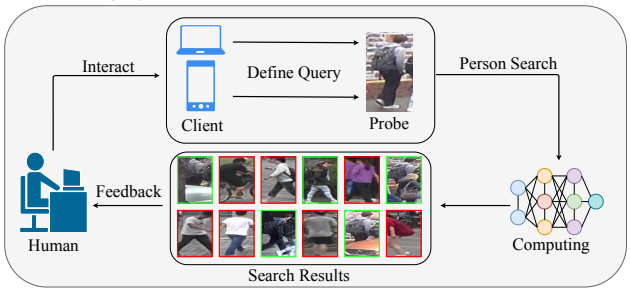
Qilei Li, Mingliang Gao*, Guisheng Zhang, Wenzhe Zhai, and Gwanggil Jeon*

***Abstract*—The objective of cross-camera persson association is to identify individuals captured across disjoint cameras. It is achieved by Re-Identification (ReID) models, which extract unique identity representations from the visual input, dominated by video sequence. Most ReID methods primarily focus on modifying the backbone network architecture to learn more representative features of people. However, these methods often overlook the impact of low-quality frames on the training process. Several studies have confirmed that low-quality data not only hinders the model from learning meaningful content but also diminishes its performance. One possible solution is to manually label the quality of each frames, but this is time-consuming and inefficient. In this paper, we propose a Unsupervised Frame Clustering and Selection framework called UFCS to address this problem by applying the unsupervised clustering to automatically select high-quality frames. Specifically, we applied three unsupervised clustering solutions for high-quality frame selection, namely K-means, Deep K-means, and DBSCAN. These clustering techniques integrate both appearance and, indirectly, temporal consistency by operating within tracklets. These schemes perform clustering in the image or deep feature space to select high-quality frames for network training. This straightforward yet effective approach enables the ReID network to generate a more discriminative representations, thereby improving recognition performance. Experimental results obtained on the challenging video-based person ReID datasets MARS indicate that our proposed scheme can outperform related state-of-the-art methods by a large margin.**

***Index Terms*—Person Association, Re-Identification, Unsupervised Clustering, Representation Learning, Frame Selection.**



(a) An example of the person association use case applied in manufacturing systems.



(b) An example of the person search use case applied in human-computer interaction systems.

Fig. 1: Illustrations of the person association system in practical use cases.

I. Introduction

Person association refers to the task of identifying instances of the same individual across multiple frames or camera views. It is applied to cropped bounding boxes that specifically contain the detected person, which are extracted following a person detection process. The performance of person association heavily relys on the Re-Identification (ReID) model, which is normally designed as a deep learning-based network that extracts unique representations of individuals [1]. ReID model aims to match representations of the same individual within a large gallery that is captured over non-overlapping cameras [2]. This task plays a significant role in various domains, such as ma manufacturing factory or a human-computer interaction system, as shown in Fig. 1.

ReID is often formulated as a representation learning task [3]–[6] and is challenging due to various factors that degrade data quality. For video-based ReID, a tracklet represents a concise sequence capturing the motion of an individual within a delimited temporal scope, which encompasses frames of diverse quality. Within a tracklet, low-quality frames typically caused by motion blur, poor lighting conditions, occlusion, extreme scale variations, or low resolution, all of which can hinder the ability to extract discriminative features for person re-identification. Specifically, These variations in illumination, pose, and camera viewpoint further exacerbate the difficulty of the task, as illustrated in Fig. 2.

In recent years, benefiting from the rapid development of deep learning techniques, many studies have turned to utilize neural networks [7], [8] to extract pedestrian features instead of conventional handcrafted feature-based methods [9], [10]. To enhance the discriminative power of person representations, some methods focus on designing new backbone network architectures, while others develop novel attention modules

Qilei Li, Guisheng Zhang, Wenzhe Zhai, Mingliang Gao and Gwanggil Jeon are with the School of Electrical and Electronic Engineering, Shandong University of Technology, Zibo 255000, China. (e-mail: qilei@ieee.org (Qilei Li), 22504030001@stumail.sdut.edu.cn (Guisheng Zhang), wenzhezhai@outlook.com (Wenzhe Zhai)), mlgao@sdut.edu.cn (Mingliang Gao), ggjeon@gmail (Gwanggil Jeon) Also, Qilei Li is with the School of Electronic Engineering and Computer Science, Queen Mary University of London, London, E1 4NS, United Kingdom.

* Mingliang Gao and Gwanggil Jeon are the corresponding authors.

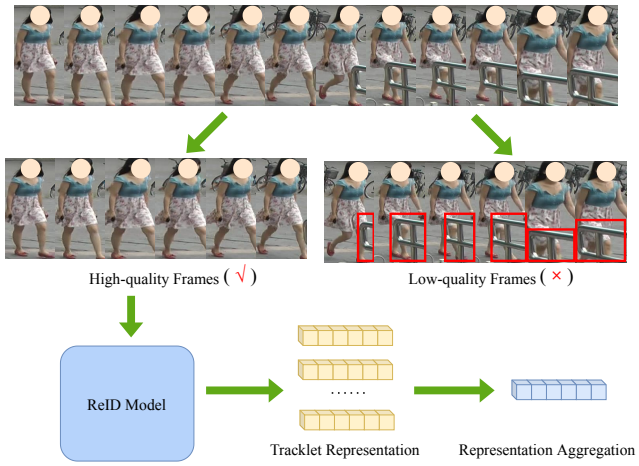


Fig. 2: Demonstration of pedestrian occlusion by an object.

to learn more distinctive feature representations [11]–[13]. For example, Fan *et al.* [14] proposed a spatial channel parallelism network (SCPNet) that utilizes spatial channel corresponding relationships to supervise network learning for overall and local discriminative features. Fu *et al.* [15] introduced a spatial-temporal attention (STA) module that considers both spatial and temporal dimensions of pedestrian discriminative properties, resulting in robust frame-level feature representation. Zhou *et al.* [13] proposed a ReID CNN named Omni-Scale network (OSNet), which realized the function of capturing different spatial scales and encapsulated the synergistic combination of multiple scales. Numerous studies [16], [17] have shown that low-quality data inputs can significantly degrade model performance across various tasks. For instance, in image classification, Samuel *et al.* [17] demonstrated that models trained on blurry images achieved significantly lower accuracy than those trained on high-resolution images. Similarly, Seo *et al.* [18] showed in object detection tasks that noisy or distorted input images lead to decreased precision and recall scores.

Instead of manually labelling the quality of each image, unsupervised clustering has recently emerged as a promising strategy for quality assessment for performance enhancement in several computer vision tasks. Raytchev *et al.* [19] applied unsupervised clustering techniques to partition facial images into distinct quality groups before feeding them into recognition models. Likewise, Sekh *et al.* [20] employed clustering methods to segregate and prioritize high-quality segments in video analysis and further validated the potential of unsupervised clustering as a preprocessing step. For video-based person ReID, the input to the network is the sequence of frames, while the frames can be in various quality and this can impact the performance of ReID model. As depicted in Fig. 3, contemporary video ReID methods typically assemble the frame representations by a pooling operation, which doesn't consider the fact that the representation extracted from the low-quality frame can degrade the aggregate person representation.

In this work, we aim to improve ReID accuracy by learning from high-quality frames selected by unsupervised clustering within a tracklet. To achieve this objective, we propose the Unsupervised Frame Clustering and Selection (UFCS) framework to identify the low-quality frames from the tracklet. The underlying assumption of UFCS is that the dominant object exhibits consistency across all frames within a concise video segment captured over a brief time frame. Specifically, we employ an unsupervised clustering approach to extract frames with high pedestrian-relevant information. These frames contain more meaningful pedestrian information as opposed to low-quality data obscured by various degradations. By doing so, UFCS frame effectively enables group of high-quality frames with more pertinent pedestrian information. By training with such data, ReID model can achieve more accurate identification of people of interest.

In sum, the contributions of the work are three-fold:

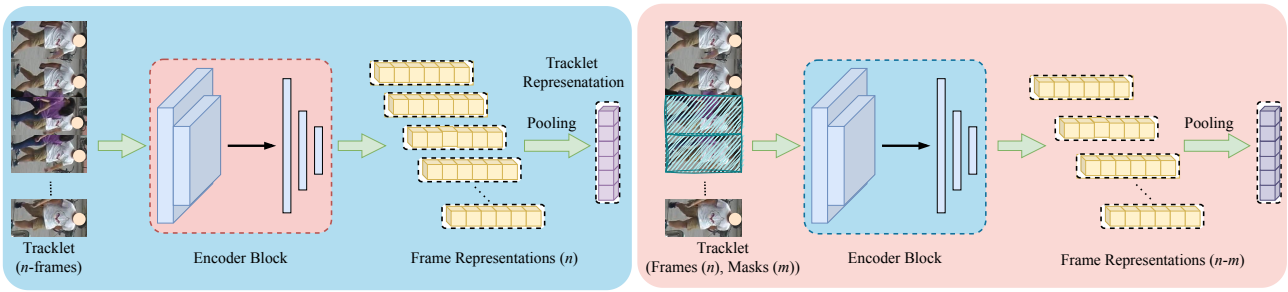
- We improve the video ReID performance by selectively using high-quality frames that are identified by the clustering algorithm. Hence, the ReID network has the capacity to extract more discriminative representation rather than being influenced by low-quality frames caused by occlusion and other interference.
- We design a simple but effective Unsupervised Frame Clustering and Selection (UFCS) scheme that can be easily incorporated into existing ReID networks to improve performance by adaptively selecting high-quality frames. In addition, the proposed UFCS is parameter-free and does not increase network computation overhead.
- For the proposed UFCS, we exclusively study three unsupervised clustering strategies to test its effectiveness, namely K-means [21], Deep K-means and DBSCAN [22]. Experimental results confirm that UFCS performs effectively with any of these clustering strategies.

The rest of the paper is structured as follows: In Section II, we review the recent works related to the proposed method. Section III details the proposed method. Section IV shows the experimental results and the ablation studies, and the paper is concluded in Section V.

II. Related Work

A. Person Re-identification

The objective of Person Re-Identification (ReID) is to accurately identify and match individuals across multiple non-overlapping camera views in a given surveillance network. Person ReID methods can roughly be divided into image-based and video-based approaches. Image-based ReID methods focus on extracting discriminative features from a single frame, often limited by challenges like varying viewpoints and occlusions. On the other hand, video-based ReID utilizes temporal sequences to capture dynamic behaviors and multiple frames. Therefore, it can represent more semantic details over temporal, while the captured frames are subject to various degradations. For video-based ReID, Li *et al.* [23] proposed an Unsupervised Tracklet Association Learning (UTAL) model that learns from extracted person tracklet data from video sequences and optimizes both the per-camera discrimination and cross-camera association losses. Thus, the UTAL model can reduce the cost of manual labeling and improve the real-world scalability of ReID networks. Zhang *et al.* [2] proposed a



(a) Traditional ReID Method (b) The Proposed ReID Method

Fig. 3: Comparisons of frame assemble strategies.

quality-guided metric learning method that enhances domain-adaptive pedestrian re-identification accuracy by assessing sample quality and incorporating adaptive weighted triplet loss. Likewise, Song *et al.* [24] proposed a large-scale dataset named Labeled Pedestrian in the Wild (LPW) and the Region-based Quality Estimation Network (RQEN) to train video-based person re-identification, which can learn part of each image quality and aggregate complementary part information of different frames in an image sequence. To mitigate the potential disruption caused by noisy pseudo-tags during model optimization, Qian *et al.* [25] proposed a successive consensus clustering framework for the iterative optimization of both pseudo-labels and the model. These methods have greatly improved the accuracy of person ReID. Nevertheless, they do not consider the impact of frame quality on ReID models. In this study, we introduce a method named UFCS for acquiring discriminative representations at the tracklet level by identifying high-quality frames. In particular, low-quality frames resulting from occlusion, illumination variation, and weather conditions can diminish the discriminative power of the extracted representations. In this paper, we aim to enhance ReID performance by excluding low-quality frames during training. To achieve this, we employ unsupervised clustering to categorize frames into distinct groups and evaluate their inherent data quality.

B. Unsupervised Data Clustering

Unsupervised data clustering [21], [26]–[28] is a fundamental approach in data mining. It aims to categorize similar data points into discrete sets by identifying common features and patterns within each cluster. Various clustering techniques have been developed from different perspectives and have been successfully applied to several vision tasks, including K-means clustering [29], Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [22], Gaussian Mixture Models (GMM) [30], Self-Organizing Maps (SOM) [31], *etc*. Among various clustering techniques, the K-means algorithm emerges as a classical unsupervised clustering method, renowned for its efficiency and simplicity. It iteratively updates data centroids and cluster assignments, ultimately yielding K distinct clusters based on feature similarity. The GMM is a statistical model. It posits that the data consists of a mixture of multiple Gaussian distributions, and the parameters of each distribution are determined through maximum likelihood estimation. Therefore, the GMM facilitates flexible clustering that is suitable for complex data distributions. Another noteworthy clustering technique is DBSCAN. Unlike K-means, DBSCAN identifies clusters based on the density and connectivity of data points, making it particularly effective in handling clusters of varied shapes and sizes. These unsupervised clustering methods have the advantage of discovering patterns and structures in unlabeled data. In this work, we introduce the UFCS method to evaluate frame quality within video sequences by utilizing efficient unsupervised clustering methods.

III. The Proposed Method

A. Problem Definition

Given a tracklet, denoted as $\mathcal{X}^j = \{x_i^j\}_{i=0}^{N_j}$, where each $\mathcal{X}^j$ represents a brief video clip capturing a person's movement within a limited time frame. These tracklets contain frames with varying quality in terms of representing person. The variation of frame quality brings unexpected challenges like occlusion and ID-switching, which can degrade the person re-identification (ReID) performance. Such occlusion introduces significant random noise to the tracklet, and makes it challenging to distinguish the person's main body.

B. Motivation and Overview

For training a ReID network, A tracklet is fed into the feature extractor to learning a high-dimensional representation corresponding to each input frame that captures key attributes. To achieve this, as depicted in Fig. 3, at the tail of the network, a pooling layer is applied to aggregate the feature map into a representation vector, which serves as the descriptor for the tracklet. However, a drawback of this process is that pooling fails to assess the relative significance of individual frames in terms of frame quality, and further causes information loss. To solve this problem, we proposed a simple yet effective Unsupervised Frame Clustering and Selection (UFCS) scheme to adaptively select high-quality frames from the low-quality ones. The diagram of UFCS is given in Fig. 4. We evaluate the per-frame quality by performing clustering on the input image, which can be applied to the image space or the deep feature space.

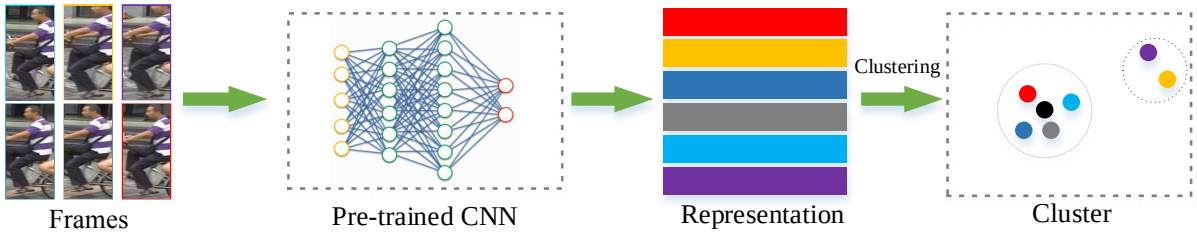


Fig. 4: Overview of the Unsupervised Frame Clustering and Selection (UFCS) framework. When provided with the frames from a single tracklet as input, our approach assesses the quality of each frame concerning the others through unsupervised clustering.

C. Preliminary Work

K-means Clustering is a classical unsupervised machine learning algorithm used for partitioning a dataset into K distinct and non-overlapping clusters. Given a group of data contains N points, $\{x_1, x_2, ..., x_N\}$, where each data point x_i belongs to a d-dimensional feature space ($x_i \in \mathbb{R}^d$), the goal is to partition these data points into K clusters, denoted as $\{C_1, C_2, ..., C_K\}$. Each cluster is characterized by its centroid μ_k, which represents the mean of all data points in the C_k cluster. It is formulated as follows,

$$\mu_k = \frac{1}{|C_k|} \sum_{x_i \in C_k} x_i, \tag{1}$$

where $|C_i|$ is the size of C_i. The K-means algorithm aims to minimize the total within-cluster variance, which is referred to as the "inertia" or "sum of squares" criterion. The process can be mathematically denoted as,

$$J(C) = \sum_{k=1}^{K} \sum_{x_i \in C_k} ||x_i - \mu_k||^2, \tag{2}$$

where $||\cdot||^2$ represents L^2 norm and $J(C)$ is the inertia or sum of squares criterion. Here, we adopt L^2 to ensure the differentiability during model optimization. The K-means algorithm iteratively updates the cluster assignments and centroids until convergence.

DBSCAN Clustering operates based on two key concepts, *i.e.*, Density Reachability and Density Connectivity. First, the DBSCAN defines the concept of "density reachability" to identify clusters. A data point p is considered density-reachable from another data point q if there exists a chain of data points $p_1, p_2, \ldots, p_n$ such that: $p_1 = q$, $p_n = p$, and each i from 1 to $n-1$, p_{i+1} is directly density-reachable from p_i. In simpler terms, p is density-reachable from q if there is a path of nearby data points connecting q to p while adhering to a specified density threshold. In this way, the point p will be identified as a core point. The aforementioned process can be formulated as,

$$|\{p \in P | d(p, q) \leq \epsilon\}| \geq MinPts, \tag{3}$$

where ϵ is the neighborhood size. $d(p, q)$ is the distance between points p and q. $MinPts$ is the minimum number of points that must exist within the neighborhood of a point for it to be classified as a core point.

D. Frame Selection by Clustering

In a tracklet, most frames consistently portray the visual attributes of a person's movements over a short time span. Nevertheless, there are specific frames that diverge from this norm, typically due to occlusions or the presence of multiple individuals. Inspired by this, it's natural to categorize the frames within a tracklet into two clusters: one for the more common frames that convey similar information and another for the outliers that deviate from the characteristics of the primary cluster. Under this assumption, we develop a frame selection machinsim by using two clustering techniques, specifically DBSCAN [22] and K-means [29]. Given a tracklet $\mathcal{X} = \{x_i\}_{i=0}^{N}$, where each $x_i \in \mathbb{R}^{c \times h \times w}$ represents a frame, the frame selection task aims to acquire an indicator function denoted as $\mathbb{1}_{\mathbf{A}}(x)$, with $\mathbf{A}$ representing the cluster containing the majority of elements. By applying this learned $\mathbb{1}_{\mathbf{A}}(x)$ to mask the input frames, the ReID (Re-Identification) backbone can effectively extract meaningful features from the high-quality frames while simultaneously reducing the impact of lower-quality ones. The frame selection process is tightly coupled with clustering: frames are grouped based on feature similarity, and the largest cluster—assumed to represent consistent temporal appearance—is retained for training. This acts as a filtering mechanism before aggregation.

K-Means Clustering for Frame Selection. For the first variation, we apply K-means on the tracklet for selection. In datasets, frames without occlusion are regarded as meaningful frames, whereas those with occlusion are denoted as less meaningful ones. Additionally, meaningful frames" comprise the majority of the dataset. The objective here is to establish two distinct clusters k, one for meaningful frames and another for less meaningful ones. Consequently, we naturally set $k = 2$. This process is formulated as

$$\{\mathbb{1}_{\mathbf{A}}(x_i)\}_{i=0}^{N} = F_{maj}(F_{km}(\mathcal{X})), \tag{4}$$

In the equation above, $F_{km}(\cdot)$ denotes the K-Means function, and F_{maj} is responsible for selecting the cluster with the highest number of data points, assigning its members to the set $\mathbf{A}$ for the indicator function.

DBSCAN Clustering for Frame Selection. As previously mentioned, DBSCAN is a density-based clustering method that eliminates the need to predefine the number of clusters. Consequently, similar to the K-Means clustering approach, we

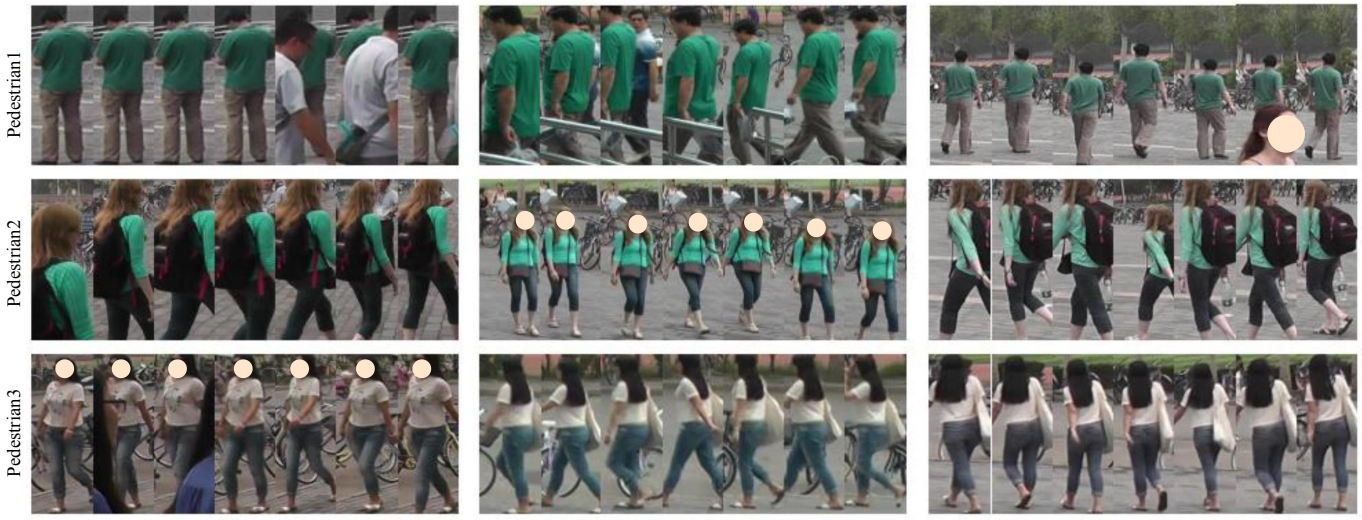


Fig. 5: A few examples sampled from MARS [32] dataset.

input the tracklet data for clustering purposes and select the most dominant cluster to represent the high-quality frames. This process can be expressed as:

$$\{\mathbb{1}_{\mathbf{A}}(x_i)\}_{i=0}^{N} = F_{maj}(F_{db}(\mathcal{X})), \tag{5}$$

where $F_{db}(\cdot)$ denotes the DBSCAN process.

Clustering in Deep Feature Space. To leverage the strong representation learning ability of the CNN network, we extend our approach to perform clustering within the deep feature space extracted by a pretrained CNN network. To accomplish this, we make the assumption that there exists a pretrained network denoted as F_{cnn}. We input the tracklet data into this network to extract deep representations, which are subsequently employed for the clustering process. This procedure can be represented as follows,

$$\begin{aligned} \{v_i\}_{i=1}^{N} &= \{F_{cnn}(x_i)\}_{i=1}^{N}, \\ \{\mathbb{1}_{\mathbf{A}}(x_i)\}_{i=1}^{N} &= F_{\text{cluster}}(F_{cls}(\{v_i\}_{i=1}^{N})), \end{aligned} \tag{6}$$

where $\{v_i\}_{i=1}^{N}$ is the per-frame deep feature, and F_{cluster} is the clustering function that can be either K-Means or DBSCAN.

ReID Model Training. After obtaining the indicator function $\{\mathbb{1}_{\mathbf{A}}(x_i)\}_{i=1}^{N}$, we utilize it as a mask to select frames for subsequent the ReID network training. The final tracklet-level representation is computed by performing average-pooling on the frame-level representations as follows:

$$H(\mathcal{X}) = \text{AvgPooling}(F_{en}(\mathbb{1}_{\mathbf{A}}(x_i) \cdot x_i)_{i=1}^{N}), \tag{7}$$

In this equation, AvgPooling$(\cdot)$ represents the average-pooling operation, $F_{en}(\cdot)$ is the chosen ReID backbone network for feature extraction, and $H(\mathcal{X})$ denotes the track-level representation. During model training, the ReID model is supervised using the cross-entropy loss.

IV. Experiments

A. Evaluation Benchmark and Metrics

Evaluation Benchmark We evaluate UFCS framework on the MARS dataset [32], which is an extensive video-based person ReID dataset, The dataset encompasses data obtained from six closely synchronized cameras and comprises 1,261 distinct pedestrians, each recorded by a minimum of two cameras. Specifically, The dataset comprises 20,478 tracklets and 1,191,003 frames. Moreover, the training and testing subsets of this dataset consist of 625 and 636 tracklets, respectively. Notably, the challenges within MARS stem from the diverse range of pedestrian poses, color variations, fluctuating illumination conditions, and suboptimal image quality. These factors collectively present formidable obstacles to achieving high matching accuracy. The MARS dataset is illustrated with some examples as shown in Fig. 5. It can be observed that there are several low-quality frames exists in each tracklet, which are potentially caused by occlusion, scale variation, illumination changes.

Evaluation Metrics. In order to assess the efficacy of the proposed framework, we employ the Cumulative Matching Characteristic (CMC) score and Mean Average Precision (mAP) as the metrics. The CMC illustrates the cumulative proportion of correct identifications at various ranks, where specific values denote the correct matching rate at particular ranks. Rank-1, Rank-5, and Rank-20 are evaluation metrics derived from the CMC. The higher these ranks, the better the model correctly identifies or ranks the person from the gallery. Meanwhile, the mAP indicates the average identification accuracy of the model across different queries.

B. Implementation Details

We followed a conventional setting to employ the pretrained VGG16 [33] model as the feature extractor. Each mini-batch includes four different identities, with each identity containing eight clips. Each clip consists of ten randomly sampled

frames from the corresponding tracklet. We conducted training for 120 epochs, setting the initial learning rate to 3.5e-5. During the first 30 epochs, we gradually increased the learning rate using a warm-up strategy. The dimension of the feature representation was set to 2048. We used the cosine distance metric to measure the similarity between query and gallery features. We used the Adam optimization for parameter updates. Key hyperparameters, such as the number of clusters (set to 2 for K-means) and DBSCAN's ϵ (set to 70), were empirically chosen after validating on held-out subsets. These parameters significantly impact cluster granularity and, consequently, the frame selection effectiveness. Our experiments were carried out on a system equipped with two P100 GPUs using the PyTorch framework.

C. Comparison with State-Of-The-Art Methods

We compared the proposed methods with the state-of-the-art competitors, including TAUDL [34], EUG [35], Snippet [36], VRSTC [37], GLTP [38], UTAL [23], STMP [39], STAR [40], STA [41], and FGRA [42]. A line of the method [36], [38], [40]–[42] fuses temporal and spatial information to capture long-range dependencies across non-consecutive frames so as to reduce the frame degradation caused by occlusion and noise in video sequences. Another line of methods [37], [39] utilizes spatiotemporal cues to reconstruct the appearance of occluded regions, therefore to improve the quality of all frame and further enhance video re-identification performance.

In Table I, we reported the performance from DeepKmeans and DBSCAN clustering with ϵ set to 70. It is evident from the results that our approach excels in both mAP and CMC score. The evaluation reflects the proposed method to derive highly discriminative representations. Especially, compared with the latest FGRA [42], the proposed method exhibits a remarkable increase of 4.6% in mAP, which underscores the proficiency in capturing fine-grained details. Additionally, our method improves the CMC scores by 2.6%, 1.03%, and 0.5%, which signifies its ability to outperform existing solutions in the task of person re-identification. That indicates our method can extract more discriminative tackle representation for accurate ReID, which is attributed to the fact that the discriminative high-quality frames were selected to learn the frame-level representation.

TABLE I: Compare with SOTA methods on MARS dataset. The best results are highlighted in **bold**.

Methods	mAP	Rank1	Rank5	Rank20
TAUDL [34]	29.1	43.8	59.9	72.8
EUG [35]	67.4	80.8	92.1	96.1
Snippet [36]	76.1	86.3	94.7	98.2
VRSTC [37]	82.3	88.5	96.5	-
GLTP [38]	78.5	87.0	95.8	98.2
UTAL [23]	35.2	49.9	66.4	77.8
STMP [39]	72.7	84.4	93.2	96.3
STAR [40]	76.0	85.4	95.4	97.3
STA [41]	80.8	86.3	95.7	98.1
FGRA [42]	81.2	87.3	96.0	98.1
UFCS (DBSCAN)	84.1	**89.7**	96.8	98.4
UFCS (DeepKmeans)	**85.1**	**89.6**	**97.0**	**98.6**

D. Ablation Study

We conducted a series of ablation studies to investigate the effectiveness of various clustering strategies and the key hyperparameters. Specifically, when utilizing DBSCAN as the clustering method, we conducted experiments with different values for the parameter ϵ by setting it to both 60 and 70. This parameter represents the maximum distance between two data points for them to be considered neighbors within the DBSCAN algorithm. Regarding K-means clustering, we explored two variants that applied clustering on either the image space or the deep feature space. For the deep feature based clustering, we used the pre-trained VGG16 [33] as the feature extractor. The results, shown in Fig. 6, demonstrate a significant performance improvement with the application of both K-means and DBSCAN clustering algorithms. Additionally, clustering applied to deep features further enhances performance. This demonstrated the effectiveness of UFCS framework.

E. Visualization on Selected Frames

We visualized a subset of frames from a pedestrian tracklet in Fig. 7. The frames highlighted as high-quality are depicted in green, while those identified as low-quality are represented in red. In the left subfigure, the pedestrian is walking along a street but becomes briefly occluded by a bicyclist. This occlusion event is short-lived, primarily due to the significantly faster movement of the bicycle compared to walking. The clustering algorithm effectively groups these distracted frames into a distinct category, thereby eliminating them from the model training process. On the right side of the tracklet, the person of interest is riding a bicycle, and some of the captured frames are partially obscured by the leg of another individual. The proposed (UFCS) can identify these frames and categorize them as low-quality. In summary, the aggregated track-level representation achieves the required discriminative power for precise person ReID through the training of the ReID network with these selected discriminative frames.

V. Conclusion

In this work, we propose a simple yet efficient Unsupervised Frame Clustering and Selection (UFCS) framework to adaptively select the most representative frames and discarding low-quality frames within tracklets. The proposed UFCS improves discriminative of assembled feature , thereby enhancing the accuracy of cross-device individual matching. Experimental results indicate the competitiveness of our approach compared to state-of-the-art techniques. Despite the proposed method demonstrating strong performance in this case, the evaluation is conducted under the Independent and Identically Distributed (I.I.D) assumption, where the test data has the same distribution as the training data. In future work, we plan to explore domain adaptation and generalization techniques to improve the model's performance across diverse data distributions.

Acknowledgements

This work was funded by Shandong Province Undergraduate Teaching Reform Project (No.Z2024184).

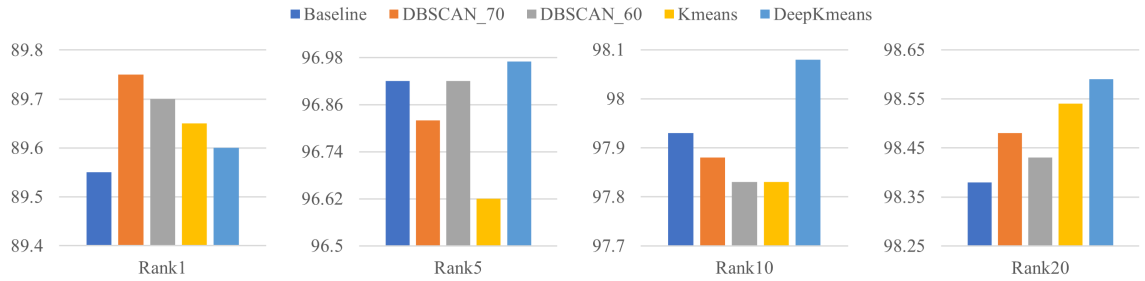


Fig. 6: Impact with different cluster strategies on frame selection.

Fig. 7: Visualization of selected frames (highlighted in green) and discarded frames (highlighted in red) within a tracklet.

Ethical Considerations

All the datasets used in the study are publicly available. The research is intended for academic purposes. We recognize potential biases in recognition models and aim to improve fairness by ensuring diverse representation in training data.

References

[1] E. Ning, C. Wang, H. Zhang, X. Ning, and P. Tiwari, "Occluded person re-identification with deep learning: a survey and perspectives," *Expert systems with applications*, vol. 239, p. 122419, 2024.

[2] L. Zhang, H. Li, R. Liu, X. Wang, and X. Wu, "Quality guided metric learning for domain adaptation person re-identification," *IEEE Transactions on Consumer Electronics*, 2024.

[3] J. Almazan, B. Gajic, N. Murray, and D. Larlus, "Re-id done right: towards good practices for person re-identification," *arXiv preprint arXiv:1801.05339*, 2018.

[4] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *ICCV*, 2015, pp. 1116–1124.

[5] N. Martinel, G. Luca Foresti, and C. Micheloni, "Aggregating deep pyramidal representations for person re-identification," in *CVPR Workshops*, 2019, pp. 0–0.

[6] G. Zou, G. Fu, X. Peng, Y. Liu, M. Gao, and Z. Liu, "Person re-identification based on metric learning: a survey," *multimedia tools and applications*, vol. 80, no. 17, pp. 26 855–26 888, 2021.

[7] Z. Zheng, L. Zheng, and Y. Yang, "A discriminatively learned cnn embedding for person re-identification," *arXiv preprint arXiv:1611.05666*, 2016.

[8] L. Wu, C. Shen, and A. v. d. Hengel, "Personnet: Person re-identification with deep convolutional neural networks," *arXiv preprint arXiv:1601.07255*, 2016.

[9] A. J. Ma and P. Li, "Query based adaptive re-ranking for person re-identification," in *ACCV*, 2014, pp. 397–412.

[10] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM journal on imaging sciences*, vol. 2, no. 1, pp. 183–202, 2009.

[11] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang, "Bag of tricks and a strong baseline for deep person re-identification," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2019, pp. 0–0.

[12] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *CVPR*, 2018.

[13] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang, "Learning generalisable omni-scale representations for person re-identification," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 9, pp. 5056–5069, 2021.

[14] X. Fan, H. Luo, X. Zhang, L. He, C. Zhang, and W. Jiang, "Scpnet: Spatial-channel parallelism network for joint holistic and partial person re-identification," in *ACCV*, 2019, pp. 19–34.

[15] Y. Fu, X. Wang, Y. Wei, and T. Huang, "Sta: Spatial-temporal attention for large-scale video-based person re-identification," in *AAAI*, 2019.

[16] C. Chen, C. Qin, C. Ouyang, Z. Li, S. Wang, H. Qiu, L. Chen, G. Tarroni, W. Bai, and D. Rueckert, "Enhancing mr image segmentation with realistic adversarial data augmentation," *Medical Image Analysis*, vol. 82, p. 102597, 2022.

[17] S. Dodge and L. Karam, "Understanding how image quality affects deep neural networks," in *2016 eighth international conference on quality of multimedia experience (QoMEX)*. IEEE, 2016, pp. 1–6.

[18] J. Seo and H. Park, "Object recognition in very low resolution images using deep collaborative learning," *IEEE Access*, vol. 7, pp. 134 071–134 082, 2019.

[19] B. Raytchev and H. Murase, "Unsupervised face recognition from image sequences based on clustering with attraction and repulsion," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, vol. 2. IEEE, 2001, pp. II–II.

[20] A. A. Sekh, D. P. Dogra, S. Kar, and P. P. Roy, "Video trajectory analysis using unsupervised clustering and multi-criteria ranking," *Soft Computing*, vol. 24, pp. 16 643–16 654, 2020.

[21] K. Krishna and M. N. Murty, "Genetic k-means algorithm," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 29, no. 3, pp. 433–439, 1999.

[22] M. Ester, H.-P. Kriegel, J. Sander, X. Xu *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise." in *Kdd*, no. 34, 1996, pp. 226–231.

[23] M. Li, X. Zhu, and S. Gong, "Unsupervised tracklet person re-identification," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 7, pp. 1770–1782, 2019.
[24] G. Song, B. Leng, Y. Liu, C. Hetang, and S. Cai, "Region-based quality estimation network for large-scale person re-identification," in *Proceedings of the AAAI conference on artificial intelligence*, no. 1, 2018.
[25] J. Qian and X. Xie, "Successive consensus clustering for unsupervised video-based person re-identification," *IEEE Signal Processing Letters*, vol. 29, pp. 822–826, 2022.
[26] D. Parmar, T. Wu, and J. Blackhurst, "Mmr: an algorithm for clustering categorical data using rough set theory," *Data & Knowledge Engineering*, vol. 63, no. 3, pp. 879–893, 2007.
[27] M. Ali, P. Scandurra, F. Moretti, and H. H. R. Sherazi, "Anomaly detection in public street lighting data using unsupervised clustering," *IEEE Transactions on Consumer Electronics*, vol. 70, no. 1, pp. 4524–4535, 2024.
[28] G. Ciocca and R. Schettini, "Supervised and unsupervised classification post-processing for visual video summaries," *IEEE Transactions on Consumer Electronics*, vol. 52, no. 2, pp. 630–638, 2006.
[29] J. MacQueen *et al.*, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, no. 14. Oakland, CA, USA, 1967, pp. 281–297.
[30] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the royal statistical society: series B (methodological)*, vol. 39, no. 1, pp. 1–22, 1977.
[31] T. Kohonen, "Self-organized formation of topologically correct feature maps," *Biological cybernetics*, vol. 43, no. 1, pp. 59–69, 1982.
[32] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian, "Mars: A video benchmark for large-scale person re-identification," in *European Conference on Computer Vision*. Springer, 2016, pp. 868–884.
[33] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CILR*, 2014.
[34] M. Li, X. Zhu, and S. Gong, "Unsupervised person re-identification by deep learning tracklet association," in *ECCV*, 2018.
[35] Y. Wu, Y. Lin, X. Dong, Y. Yan, W. Ouyang, and Y. Yang, "Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning," in *CVPR*, 2018.
[36] D. Chen, H. Li, T. Xiao, S. Yi, and X. Wang, "Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding," in *CVPR*, 2018.
[37] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, and X. Chen, "Vrstc: Occlusion-free video person re-identification," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 7183–7192.
[38] J. Li, J. Wang, Q. Tian, W. Gao, and S. Zhang, "Global-local temporal representations for video person re-identification," in *ICCV*, 2019.
[39] Y. Liu, Z. Yuan, W. Zhou, and H. Li, "Spatial and temporal mutual promotion for video-based person re-identification," in *AAAI*, 2019.
[40] G. Wu, X. Zhu, and S. Gong, "Spatio-temporal associative representation for video person re-identification." in *BMVC*, 2019.
[41] Y. Fu, X. Wang, Y. Wei, and T. Huang, "Sta: Spatial-temporal attention for large-scale video-based person re-identification," in *AAAI*, 2019.
[42] Z. Chen, Z. Zhou, J. Huang, P. Zhang, and B. Li, "Frame-guided region-aligned representation for video person re-identification," in *AAAI*, 2020.