International Journal of Computer Vision (2020) 128:2936–2961
https://doi.org/10.1007/s11263-020-01349-4

Long-Short Temporal–Spatial Clues Excited Network for Robust Person Re-identification

Shuai Li[1,2] · Wenfeng Song[1] · Zheng Fang[1] · Jiaying Shi[1] · Aimin Hao[1,2] · Qinping Zhao[1,2] · Hong Qin[3]

Received: 1 February 2020 / Accepted: 17 June 2020 / Published online: 15 July 2020

Abstract

Directly benefiting from the rapid advancement of deep learning methods, person re-identification (Re-ID) applications have been widespread with remarkable successes in recent years. Nevertheless, cross-scene Re-ID is still hindered by large view variation, since it is challenging to effectively exploit and leverage the temporal clues due to heavy computational burden and the difficulty in flexibly incorporating discriminative features. To alleviate, we articulate a long-short temporal–spatial clues excited network (LSTS-NET) for robust person Re-ID across different scenes. In essence, our LSTS-NET comprises a motion appearance model and a motion-refinement aggregating scheme. Of which, the former abstracts temporal clues based on multi-range low-rank analysis both in consecutive frames and in cross-camera videos, which can augment the person-related features with details while suppressing the clutter background across different scenes. In addition, to aggregate the temporal clues with spatial features, the latter is proposed to automatically activate the person-specific features by incorporating personalized motion-refinement layers and several motion-excitation CNN blocks into deep networks, which expedites the extraction and learning of discriminative features from different temporal clues. As a result, our LSTS-NET can robustly distinguish persons across different scenes. To verify the improvement of our LSTS-NET, we conduct extensive experiments and make comprehensive evaluations on 8 widely-recognized public benchmarks. All the experiments confirm that, our LSTS-NET can significantly boost the Re-ID performance of existing deep learning methods, and outperforms the state-of-the-art methods in terms of robustness and accuracy.

Keywords Person re-identification · Temporal–spatial clues · Long-short appearance model · Motion-refinement · Low-rank analysis

1 Introduction

Person Re-ID is of fundamental significance to many intelligent surveillance applications pertinent to computational vision. The key is to robustly track persons from one camera view (probe) to another camera view (gallery) subject to stochastically-varying conditions. However, to realize this ambitious goal, there are still several challenges yet to be overcome, including intrinsic temporal–spatial clues/features representation and extraction from cross-camera videos, transferable reuse of the pre-trained model in unknown scenes, etc. Recent technical advances mainly stem from the rapid development of deep learning based models (Chen et al. 2018; Li et al. 2018), though they are critically dependent on a large number of well-labeled datasets (Karanam et al. 2018).

Specially, we should focus more on the unsupervised cross-scenario ReID task, which would require to conduct

Communicated by Patrick Perez.

Shuai Li and Wenfeng Song have contributed equally.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11263-020-01349-4) contains supplementary material, which is available to authorized users.

✉ Wenfeng Song
songwenfenga@163.com

✉ Hong Qin
qin@cs.stonybrook.edu

Shuai Li
lishuai@buaa.edu.cn

[1] State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing, China

[2] Stony Brook University, Stony Brook, USA

[3] Peng Cheng Laboratory, 518055 Shenzhen, China

unsupervised deep transfer learning to make the model (pre-trained on the source dataset) well accommodate certain cross-scenario target dataset. In particular, unsupervised cross-scenario ReID mainly faces with two critical challenges. The first challenge is to extract robust features from long-term cross-camera sequences. To overcome the perception field limitation of the convolution operation, we should decompose and compress the person-related information in advance from multiple ranges of sequences. The second challenge is to preserve the discriminative personalized features at the frame level, which requires to extract the intrinsic appearance features of the pedestrian (e.g., body shapes, pose, motion patterns, etc.). However, these features are hard to be determined in an single frame. Hence, the temporal motion clues should be explored to guide the spatial feature extraction, so that we can flexibly and thoroughly bridge the gap between the temporal and spatial clues. Therefore, the key issue for us to tackle is how to decompose the features in the temporal dimension to obtain a generalized one.

In principle, the person's appearance features in different scenes can be decomposed into two components: 'Background' component and 'Person' component. The 'Background' component is relatively stable in similar scenes but appears diverse across different scenes as shown in Fig. 1 (e.g., being different with clutter background, camera views, poses, accessories, and partial occlusions). As documented in previous works (Zeng et al. 2012; Ye et al. 2017; Liu et al. 2017), the 'Person' component (e.g., person's appearance features) provides temporal clues to preserve/complete person's identity features. However, the cross-scene robustness of these features still needs to be further explored. The shared person features in different scenes are induced from the stable information at multiple temporal ranges. The short-term temporal sequences tend to provide the detailed person appearance features, while the long-term temporal sequences tend to avoid over-fitting with irrelevant distinct clues. In principle, the person covers the moving regions in the videos, therefore, we uniformly name the 'Person-related component' as 'Motion' in the following texts.

Essentially, the 'Background' component would have influences on the feature integrity of the persons, and perturb the person's discriminative features. One feasible way is to eliminate the influences of the background (Tian et al. 2018; Kalayeh et al. 2018). Thus, it should focus on the extraction of the person's distinct clues by suppressing the background. For example, some research works (Li et al. 2017) resort to learning the intrinsic dictionaries for feature representation in non-overlapping camera views. Other pose-sensitive methods (Sarfraz et al. 2018; Liu et al. 2018; Huang et al. 2018) try to utilize local/global person descriptions by incorporating the body pose/part information, which could significantly improve the person Re-ID performance. Besides, some methods couple primary spatial–temporal clues to combine the

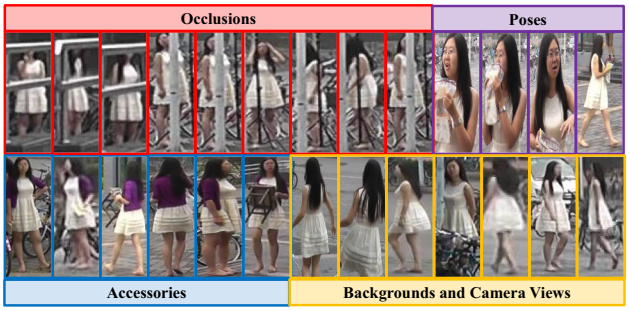

Fig. 1 Illustration of the different backgrounds, accessories and poses, of which, the intrinsic features can only be captured by considering long-term sequence dependencies

sequential and spatial CNN features by aggregating them in RNN (Zhou et al. 2017) or conducting pooling operation on the sequences (Xu et al. 2017). These methods (McLaughlin et al. 2016; Zhou et al. 2017; Xu et al. 2017) tend to directly extract the temporal and spatial features in an alternative way or simply exploit the spatial information without considering temporal priors. Consequently, they fall short in learning local-to-global coherency from multi-range video sequences, and they may also be overwhelmed by clutter background, which would further weaken the discriminative ability of the person's intrinsic features.

To alleviate, our key insight includes two aspects: (1) The long-short temporal sequences can provide exhaustive person-related clues to complete the occluded regions and distinct regions, and we shall employ the temporal feature to promote the Re-ID improvement; (2) In realistic scenes, it is hard to guarantee the scalability across scenes, we should employ an unsupervised multi-range low-rank decomposition (most of the'multi-range' expressions relate to the temporal clues underlying multiple lengths of sequences) to conduct transferrable motion extraction from temporal sequences. Since the sequences provide temporal coherency prior for certain frame, we can retain more discriminative person-related features while eliminating the influences from the environmental factors. Therefore, our motion feature is built upon the temporal clues with corresponding feature response in CNNs, which is more robust to the interference of irrelevant objects and drastically-changing poses. Moreover, different from attention based methods (Kalayeh et al. 2018; Tian et al. 2018; Peng et al. 2016), we seek to represent the person in certain frame by aggregating all the relevant frames' information, which can better preserve the temporal context information, while being less sensitive to noise and occlusions. On the other hand, human beings are more sensitive to moving objects than stationary ones (Burr and Santoro 2001). To emulate, we propose to extract temporal features to excite the highly-relevant spatial features, which have been rarely studied in the previous exploration works on spatial–temporal clues.

Specifically, our new method can extract features in both temporal and spatial domains, which are characterized in three modules (all serving as enablers): temporal motion clue capture, robust spatial feature representation, and the aggregation mechanism. The salient contributions of this paper can be summarized as follows.

- We propose an unsupervised long short-term motion decomposition scheme to decouple the underlying intrinsic motion information via low-rank analysis over multiple ranges of cross-camera video sequences, which can adaptively capture and highlight the temporally-invariant person motion priors.
- We propose an interactive motion-refinement scheme to select the intrinsic cross-scenario motion clues via a self-attention mechanism (enhanced by message-passing), which can automatically activate the motion regions, augment the coherence clues, and reduce the redundancy of different temporal motions.
- We design an end-to-end network [equipped with a motion excitation block (MEB)] to exploit the distinct cross-scenario person features, which can flexibly aggregate and refine the critical temporal clues to excite person's discriminative features for robust cross-scenario person ReID.

2 Related Works

Deep learning based methods (Zhao et al. 2017a; Liu et al. 2018) have been dominating the person Re-ID research community. These methods can be roughly classified into two categories. The first category mainly focuses on extracting invariant features (Zhong et al. 2017a) from images to improve the discriminative power. The second one extends the feature exploration from images to video sequences by integrating the temporal clues into the spatial features (McLaughlin et al. 2016; Zhou et al. 2017; Xu et al. 2017).

2.1 Image based Person Re-ID Methods

Image based methods usually emphasize to make the feature representation and measurement be insensitive to illumination, pose, and camera views. Many works (Zhao et al. 2017, a; Zheng et al. 2017; Li et al. 2017) have analyzed the pose-invariant features used to align the pedestrians, which gives rise to better improvement w.r.t the methods without considering the pose alignment. To suppress the irrelevant information (Song et al. 2018; Xu et al. 2018; Si et al. 2018) tried to remove the background and paid more attention on the pedestrian-relevant regions. Zhao et al. (2017b) resorted to evaluating the saliency of the pedestrian regions. These works achieved remarkable improvements compared to conventional whole image based methods. Some other works (Yu et al. 2017; Zhong et al. 2017a) respectively proposed new feature distance metrics to better measure the similarity relationships of different samples. It is obviously that, single image based methods cannot avoid ignoring the intrinsic relations, as a result, these methods lack generalization ability when they are applied to un-known scenarios.

To handle the large varieties caused by different views, existing works mostly try to align the pedestrians via intrinsic feature learning. For example, Wang et al. (2018) developed to explicitly designing a feature embedding space for supervised re-id task. Li et al. (2017) proposed cross-view dictionary learning method to improve the representation ability. Yu et al. (2017) proposed a cross-view asymmetric metric learning for unsupervised person Re-ID. Li et al. (2019) proposed to incrementally discover and exploit the underlying discriminative information from automatically-generated person tracklet data in end-to-end way. Peng et al. (2018) proposed to decompose the image into semantic, latent discriminative and latent background attributes, respectively. Among the previous works, unsupervised transfer learning (Peng et al. 2016; Lv et al. 2018; Wang et al. 2018; Lin et al. 2017) shown advantages over others in the cross-scenario Re-ID task. On this basis, we observe that, different components can be shared more or less in different degrees. For example, the pedestrians' intrinsic attributes, like hair, coat, walking action and body shape, are stable across different natural scenarios, while the views, pose and illumination conditions are easily changed. The separated transfer learning should be more efficient for the pedestrian identification. Therefore, we would decompose the feature into scenario-specific component and scenario-intrinsic component, and only the latter one will be focused on when conducting feature transfer.

2.2 Video based Person Re-ID Methods

For video based re-identification task, a person is rarely fully-visible in all frames. However, different frames often contain temporal clues, thus, the temporal context aggregation facilitates to accommodating extreme variation in person appearance. Recently, an increasing number of works have been proposed for video-based person Re-ID by combining spatial–temporal features (McLaughlin et al. 2016; Simonyan and Zisserman 2014). These works could be roughly classified into two categories: the first one uses the *local* temporal features to augment the spatial features, and the second one exploits temporal context towards the *global* consistency in videos. Existing works (Simonyan and Zisserman 2014; Li et al. 2018; Liu et al. 2017; Simonyan and Zisserman 2014; Xu et al. 2017) tried to extract temporal features from consecutive frames using optical flow or other registration mechanism. For example, Wang et al. (2016)

and Zhou et al. (2017) selected the most discriminative video fragments from video sequences, so that more reliable spatial–temporal features could be extracted. However, the short-term sequence based learning methods are still hard to represent the across-scenario long-term coherency, though the long-term coherency is important to distinguish the stable background and the dynamic foreground. To this end, Li et al. (2019) proposed to learn the discriminative Global-Local Temporal Representations (GLTR) from a video sequence by embedding both short and long-term temporal cues. In fact, GLTR is an innovative and powerful framework, which solves the critical challenges: aggregating the features in temporal dimension to obtain a discriminative representation, This works is relevant to our research works. At the macro-level, the overall pipeline of our method is in spirit similar to that of GLTR. Yet, our LSTS-NET mainly focuses on the unsupervised cross-scenario ReID task, which would require to conduct unsupervised deep transfer learning to make the model (pre-trained on the source dataset) well accommodate certain cross-scenario target dataset. Towards this specific objective, we concentrate our research efforts much more on the generic and robust person-related feature extraction across different scenarios.

Recently, to meet the requirement of unsupervised transfer ReID tasks, some global video sequence based learning methods were proposed. For example, Wu et al. (2018) proposed to exploit unlabeled tracklets by gradually improving the discriminative capability of the CNN feature representation via step-wise learning. Liu et al. (2017) accumulated the motion context clues to exploit the long-term motion context to identify the same person under challenging conditions. Although global video sequence based learning methods are able to extract the stable coherent features in temporal space, it might fail when interferences or corruptions occur. Besides, these methods tend to extract more sharing features, which would overwhelm the discriminative features for different identities. Specially, neither local nor global sequences based method alone is able to fully leverage intrinsic appearance cues for motion prediction. We find that, the local temporal context sequences can provide details to recover the neighboring frames contaminated by the occlusions, while the global sequences can provide common features across long-term changes. Thus, we will fuse the multi-range temporal contexts with spatial features seamlessly.

2.3 Spatial–Temporal Clues Based Works

Some recent works proposed to leverage spatial–temporal models and attention models. For example, Li et al. (2020) proposed a compact 3D convolution layer to capture the temporal cues. Subramaniam et al. (2019) proposed a co-segmentation mechanism to combine the temporal and spatial clues. Fu et al. (2019) exploited the persons' discriminative parts in both spatial and temporal dimensions via the attention model with an inter-frame regularizer. In summary, most of these works improve the discriminative ability in supervised video-person ReID tasks, and they demonstrate that, the temporal and spatial clues could work collaboratively to achieve better performance. Inspired by the aforementioned research works, our LSTS-NET aims to extend the temporal clues for unsupervised cross-scenario person ReID.

2.4 Attention Modeling for Person Re-ID

Attention modeling usually involves temporal and spatial clues, which aims to augment the cross-scenario common features. For example, Tian et al. (2018), Chen et al. (2018), Li et al. (2019b), and Kalayeh et al. (2018) respectively proposed human parsing maps to learn more discriminative person part features to solve the background-biasing problem. These works only model the common features in a limited ranges of sequences, and lack an effective mechanism to fuse temporal–spatial clues in CNNs, thus, they are hard to handle the cases with large differences. Alternatively, Chen et al. (2018) divided long sequences into multiple short video snippets, and aggregated the top-ranked snippet similarities estimated by a temporal co-attention. Given different parts of the pedestrians, Li et al. (2018) demonstrated that, the diversity of temporal–spatial attention could benefit the video based Re-ID. Si et al. (2018) proposed a dual attention matching network to align the temporal context-aware features among sequences. Li et al. (2019a) extended the scalability to large scale re-id deployment scenarios by attention selection. Inspired by these works, we will leverage hierarchical temporal attention mechanism to excite the spatial features maps by incorporating low-rank analysis (Zhou and Tao 2011; Candès et al. 2011) over multi-range video sequences, which would capture the motion in local range and extract the stable background-related features in global range.

Recently, more sophisticated attention-mechanism have been proposed to precisely model the local and global relationships for the ReID task. For example, Tay et al. (2019) integrated person attribute attention maps into a classification framework for ReID task. Xia et al. (2019) proposed an attention mechanism to directly model long-range relationships via second-order feature statistics. Meanwhile, Chen et al. (2019) proposed the High-Order Attention (HOA) module to utilize the complex and high-order statistics information. Zhou et al. (2019) proposed a consistent attention regularizer and an improved triplet loss to learn foreground attentive features for person ReID. Similar to our work in spirit, these works attempt to learn the foreground regions. However, most of the existing works are designed to learn the spatial clues, which are hard to be applied in video datasets. Zhang

et al. (2019) adopted video pairs and output their matching scores via non-parametric attention mechanism. We further exploit the temporal clues and embed them in the attention mechanism.

3 Method Overview

The overall architecture of our LSTS-NET is shown in Fig. 2. The main components of the network include a long-short motion appearance model, motion-refinement module, and a motion excitation network. All the cross-camera videos are divided into snippets, which aims to partially extract the long-term background from video sequences in rolling way. The frame snippets serve as the inputs of our LSTS-NET. Following the global low-rank decomposition over across-camera frame snippets, we conduct local low-rank decompositions over the same-camera snippets to get multi-range motion priors (denote the component related to pedestrian's moving pattern decoupled from multiple ranges of video sequences), which benefits to preserve the person-related features. Afterwards, we further refine the motion features via motion-refinement module, which naturally intergrades semantic distinct feature maps into the temporal motion features. Subsequently, four newly-designed motion excitation blocks (MEB) and an identity mapping of the parallel motion flow are embedded in the CNN, which can provide cross-frame spatial prior to complete the missing person-related features. The above network structure is shared in both training and testing phases.

3.1 Long Short-Term Appearance Model

We propose to utilize the multi-range temporal contexts in unsupervised way, and it is inherently superior over supervised methods in transferring the knowledge of source dataset to the unknown datasets. More concretely, to model the motion features closely related to the person Re-ID task, we decompose the multi-range frame snippets into background components and the person components. we uniformly name the person-related component as 'Motion' in the following texts. It is detailed in Sect. 4.

3.2 Motion-Refinement Module

Given the temporal features resulted from the long-short temporal appearance model, since it tends to focus more on the common person-related features across scenarios, it loses the individual details. Therefore, we propose to further introduce motion-refinement layers to promote the discriminative ability. It is detailed in Sect. 5.1.

3.3 Motion Excitation Network

The person component and the learned importance of each pixels in this component jointly constitute the motion prior, respectively corresponding to global temporal clues and local discriminative features, which will be fed into motion excitation blocks. The details is described in Sect. 5.

4 Cross-Scenario Long Short-Term Appearance Model

An unsupervised method tends to extract more generic features for various scenarios. Hence, we propose to utilize the multi-range video sequences in an unsupervised way, which is robust and stable across scenarios. Meanwhile, the supervised method will be limited by the training datasets. Namely, it will outperform the supervised methods in extracting the common knowledge underlying the labeled source dataset and the un-labeled target dataset. Furthermore, to extract the person-related motion features, we decompose the multi-range frame snippets into background components and the person's motion components. The long sequence involves the global background information and more camera views, while the short sequence contains much specific person-related features. We propose the cross-scenario long short-term appearance model to sufficiently aggregate these clues to provide both generic and critical details for cross-scenarios ReID tasks. For clear description, we summarize the involved symbols in Table 1.

4.1 Cross-Scenario Motion Clues Decoupling

To capture the specific motion cues, we decompose the video snippets into moving and static components. Of which, the cross-scenario moving persons have different visual appearances, however, their motion patterns are relatively stable and invariant. Such characteristics of the motion features are suitable to be transferred among different scenarios. Intuitively, the temporal sequences provide underlying information to discriminate the moving objects, even with severe occlusion, clutter background and large view differences. The videos are decomposed into two components: 'motion' and 'background', wherein, the 'motion' features are temporally complementary among the discrete frames.

The key challenge of cross-scenario Re-ID is to distill the person-intrinsic features in temporal space, which should simultaneously be robust and distinct for each person. The pre-fixed ranges commonly used for temporal clue extraction are insufficient for crowded scenarios with various environment distractions. Therefore, we propose a hierarchical integration mechanism to capture the multi-range temporal clues from long-short video sequences. Specifically, we

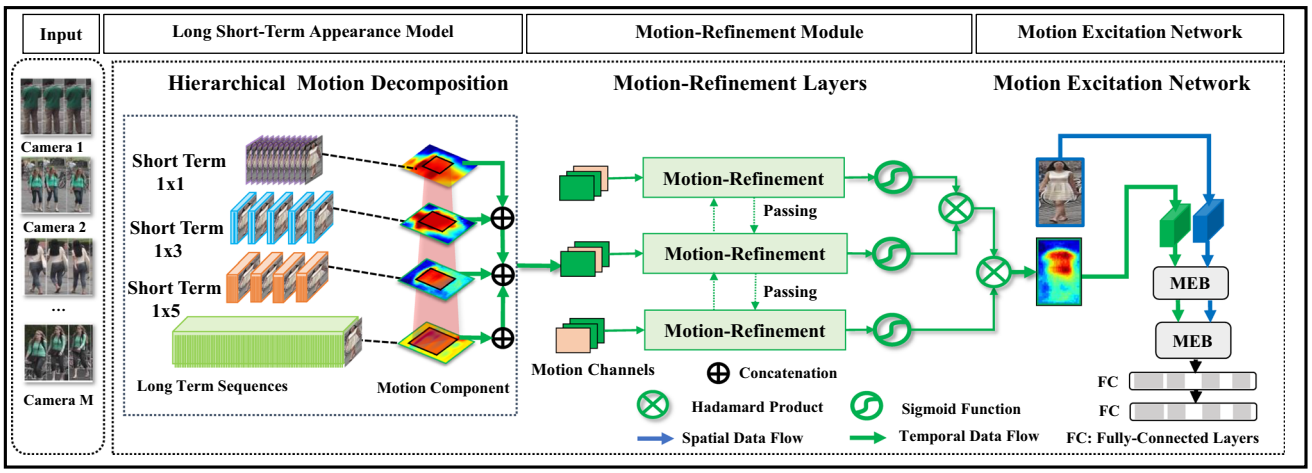

Fig. 2 Architecture overview of our LSTS-NET. To model the motion prior, we divide the same-camera and across-camera video sequences into snippets, which are then used for low-rank analysis in multiple branches: long-term low-rank decomposition over across-camera snippets, multiple ranges of short-term low-rank decomposition over same-camera snippets. Of which, the short-term low-rank decomposition results and the long-short snippets are separately fed into the motion-refinement layers with motion-passing operations, of which they are aggregated by self-attention mechanism. Meanwhile, we employ the motion features to activate the relevant pedestrian's features via Motion Excitation Blocks (MEB)

Table 1 Notations list

Symbols	Meanings
w	Temporal window size for low-rank analysis
m, n	Height and width of the input video frame
f_r	the rth frame
C	Channels of the network structure
$Y(), F()$	The feature maps output by MEB embedded CNN, baseline CNN
$\mathcal{R}_A, \mathcal{R}_B$	Long-term sequences, short-term sequences
$\mathbf{M}, \mathbf{M}(w)$	'Motion' and its extraction function (sparse matrix) over the frame set with the temporal window w
$\mathbf{D}$	Video frames spanned matrix
$\mathbf{B}$	Background component (low-rank matrix) over the frame set
$\mathbf{B}(w)$	Background extraction over the frame set with the temporal window w
$\mathbf{B}_r$	Background decomposed from the frame set: frame #1 to frame #r.
$\mathcal{R}$	Relation set w.r.t the current image

define the sliding windows as $W = \{w_1, w_2, \ldots, w_N\}$ to cover the temporal motion features within different ranges. Supposing we have a video frame spanned matrix, denoted as $\mathbf{D} = \{f_1, \ldots, f_r\} \in \mathbb{R}^{m \times n \times r}$, it can be decomposed into a low-rank matrix $\mathbf{B}$ (representing the background) and a sparse matrix $\mathbf{M}$ (consisting of the foreground objects). For convenience, we use $\mathbf{M}(f_r | f_r \in w_r)$ to represent $\{\mathbf{M}_{r-w}, \ldots, \mathbf{M}_r, \ldots, \mathbf{M}_{r+w}\}$. Then, the frame-wise decomposition can be formulated as follows.

$$\begin{aligned} \mathbf{D}(r) = \sum_{w_r} (&\mathbf{B}(f_r | r \in w_r) + \mathbf{M}(f_r | r \in w_r) \\ &+ \mathbf{G}(f_r | r \in w_r)), \\ s.t. \, &rank(\mathbf{B}(w_r)) \leq T_r, card(\mathbf{M}(w_r)) \leq T_c, \end{aligned} \tag{1}$$

where $\mathbf{G}$ is the noise component decomposed from $\mathbf{D}$, f_r is the rth frame image, T_r is the rank of matrix $\mathbf{B}(w_r)$, T_c is the cardinality range of $\mathbf{M}(w_r)$. In practice, we handle the motion regions from the low-rank and sparse components in separate cases. Specifically, the rank constraint of the matrix indicates the number of unrelated entries, namely, the higher rank of the matrix will preserve more orthometric entries. Namely, the objects with different moving patterns will be preserved. The cardinality controls the elements number of the decomposed sparse matrix. The smaller the cardinality is, the less meaningful information it will contain. Consequently, the person-related information will be preserved and the background component will be removed. However, the environment is complex in person ReID tasks, as foreground

Fig. 3 Illustrations about the low-rank decomposition results under different cardinalities ($T_c = 100, 5000, 10{,}000$). As 'CARD' increasing, the decomposed sparse component (moving pedestrians) are preserved with more details. Blue boxes mark the differences among various sparse components

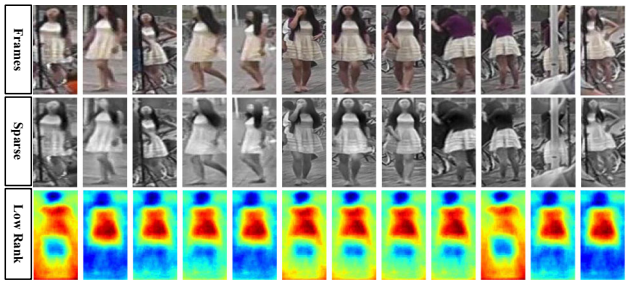

Fig. 4 Illustrations about the low-rank decomposition results under different rank constraints ($T_r = 2, 20, 50$). As 'rank' increasing, the decomposed low-rank component (stable background) is preserved with more information, and the background is also influenced by the unrelated information. Yellow boxes mark the differences among various low-rank components

moving objects are located in clutter background with occlusions and noise. Hence, we should simultaneously constrain the rank of matrix to obtain the stable component, and simultaneously constrain the cardinality to preserve the parts most relevant to the pedestrians.

In theory, the low-rank decomposition over video frames should result in time-varying component (sparse component) and stable component (low-rank component) across video sequences.

1. From the holistic perspective (holistic frames): the motion should relate to the sparse component, and the background should relate to the low-rank component, because the stable image content corresponds to the background component in the frame-spanned matrix $\mathbf{D}$, while the varying content corresponds to the foreground objects, such as pedestrians, cars, etc. We show some examples in Fig. 3.
2. From the cropped perspective (cropped frames with pedestrians): the motion should relate to the low-rank component, while the background should relate to the sparse component. In the cropped frame sequences, the pedestrian locates in the center of each cropped image. The stable content across frames corresponds to the body regions of the pedestrians. The varying content corresponds to the clutter background, the moving legs, hands or the noisy occlusions. We show some examples in Fig. 4.

Since we aim to decompose the image into two components, based on our experiments, T_r is empirically set as 2, and T_c is set as 0.05 in soft GoDesc. Equation 1 can be approximated by alternatively solving the following two subproblems until convergence,

$$\begin{cases} \mathbf{B}_r^t = \underset{rank(\mathbf{B_r}) \le T_r}{\arg\min} \parallel \mathbf{D}_r - \mathbf{B}_r - \mathbf{M}_r^{t-1} \parallel_F^2, \\ \mathbf{M}_r^t = \underset{card(\mathbf{M_r}) \le T_c}{\arg\min} \parallel \mathbf{D}_r - \mathbf{B}_r^t - \mathbf{M}_r \parallel_F^2 . \end{cases} \tag{2}$$

Here, $\mathbf{D}_r = (\mathbf{D}(f_1), \mathbf{D}(f_2), \ldots, \mathbf{D}(f_r))$, $||\cdot||_F$ represents the Frobenius norm, t denotes the iteration times. Although Eq. 2 has non-convex constraints, their global solutions $\mathbf{B}_r$ and $\mathbf{M}_r$ can be well approximated using GoDec (Zhou and Tao 2011) method. In fact, this problem can be solved by updating $\mathbf{M}_r^{t-1}$ via singular value hard thresholding of $\mathbf{D}_r - \mathbf{B}_r$ and updating $\mathbf{M}_r^t$ via entry-wise hard thresholding of $\mathbf{D}_r - \mathbf{M}_r^t$, respectively. The multi-range low-rank analysis provides adequate temporal motion clues for the current frame. Nonetheless, how to flexibly select the most relevant surrounding sequences and how to aggregate the multi-range motion features still need to be solved.

4.2 Appearance Model Construction

We observe that the consecutive frames have a large overlapping motion field, which is redundant. To efficiently aggregate the hierarchical clues, we propose a long short-term appearance model to extract temporally-coherent features.

Long-term batch $\mathbf{D}_A$ involves a fixed length of inter-camera video sequences $\mathbf{D}_d$, which cover the diversities of view and illumination, and a fixed length of intra-camera sequences $\mathbf{D}_s$, which focuses on the pedestrians' discriminative features. We re-organize the two kinds of sequences ($\mathcal{R}_d$, $\mathcal{R}_s$) into fixed size of batches $\mathcal{R}_A = \mathcal{R}_d \cup \mathcal{R}_s$. Short-term batch $\mathbf{D}_B$ involves a fixed length of intra-camera sequences, which convey the consistent pedestrian's appearance. We denote the set of short-term batches as $\mathcal{R}_B$.

4.2.1 Long-Term Clues

The long-term appearance model aims to provide the common clues in global temporal ranges while caring less about the person's individual features. $\mathcal{R}_s$ is decomposed into the 'background' and 'motion' components appearing in the same cameras. $\mathcal{R}_d$ is decomposed into camera-invariant features and variant components. Specifically, we divide all

the sequences into batches $\mathbf{D}_A$, which involve all the different cameras views. Meanwhile, we dynamically update the batches to go through all the sequences. Based on Eq. 1, the long-term 'motion' clues $\mathbf{M}_A$ and camera-invariant background $\mathbf{B}_A$ are formulated as,

$$\begin{aligned} \mathbf{D}_A &= \mathbf{B_A}(f_r|r \in \mathcal{R}_A) + \mathbf{M_A}(f_r|r \in \mathcal{R}_A) \\ &\quad + \mathbf{G}(f_r|r \in \mathcal{R}_A), \\ rank(\mathbf{B}_A) &\leq T_r, card(\mathbf{M}_A) \leq T_c. \end{aligned} \tag{3}$$

Here $\mathbf{D}_A$ is decomposed into motion clues and the scenario-specific components, $\mathbf{B}_A$ represents camera-invariant components, as shown in Fig. 5. When the number $||\mathcal{R}_A||$ of set $\mathcal{R}$ is larger, the motion $\mathbf{M}_A$ will tend to involve more global clues, which are averaged over larger ranges of frames.

4.2.2 Short-Term Clues

Based on the long-term decomposition model, we get $\mathbf{B}_A$ as the prior clues, which should be considered from the short-term sequences to get coarse motion regions, and $\mathbf{M}(r \in \mathcal{R}_B)$ is decomposed from $\mathbf{D_B}$ based on Eq. 1. The short-term model aims at exhaustively aggregating the individual details to represent different persons. To this end, we introduce a pyramid of short-term sequences to perceive video contents in multiple-range neighboring frames. The pyramid, which refers to mean-pooling video frame representations over different temporal ranges, is gradually enlarged along the temporal dimension to get continuously-increasing perception scope. The pyramid aggregation model is formulated as,

$$\begin{aligned} \mathbf{M_B}(f_r) = \frac{1}{||W|| + 1}(&\mathbf{M}(f_r|r \in w_1) + \mathbf{M}(f_r|r \in w_2) \\ &+ \mathbf{M}(f_r|r \in w_3) + \ldots + \mathbf{M}(f_r|r \in w_k)), \end{aligned} \tag{4}$$

where $k \in [0, N]$, and $1 * (2 * k * s + 1)$ means the size of the sliding window along the temporal dimension, s is set to 10, covering 10 frames. When k is larger, $\mathbf{M_B}$ tends to be the person-related mean motion clues in larger-range sequences, $\mathbf{M}_A$ provides the motion field coarsely in global range, $||W||$ is the number of the elements in set W. The main advantage of the short-term low-rank decomposition is that, it recovers the person region from occlusions, and extracts the intrinsic features of the same person among different frames. It is observed that, the decomposed 'motion' $\mathbf{M_B}(f_r)$ ranging from 0 to $2 * k * s + 1$ captures appearance features from specific to general, which is non-trivial in distinguishing different persons. The principle advantage of the pyramid short-term model in W is that, it produces a multi-range component aggregation mechanism to extract feature from global to local, while creating a feature pyramid strongly related to the person at all temporal ranges $\{f_{r-k}, \ldots, f_r, \ldots, f_{r+k}\}$.

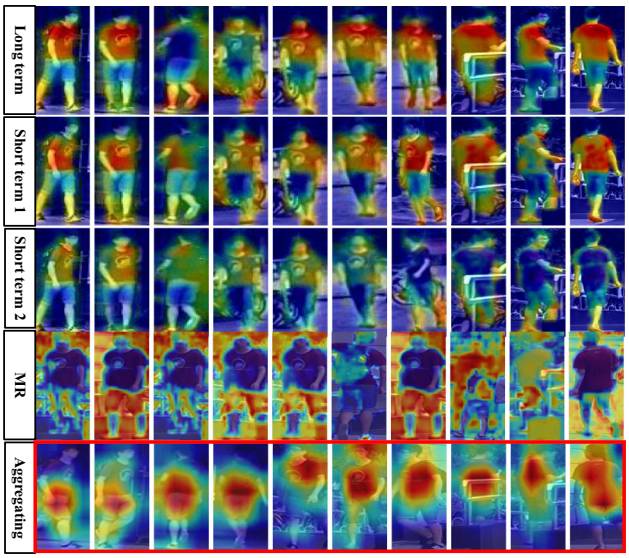

Fig. 5 Visualization of LSTS-NET, including long-term motion clues, two examples of short-term motion clues, the feature maps with 'motion-refinement' clues (simply denoted as 'MR'), and aggregating clues. The long-term model facilitates to recovering the critically-occluded regions. The short-term model can preserve more details relevant with person Re-ID. The motion-refinement module extracts the shallow feature maps in CNNs, while the aggregating scheme produces high response to the discriminative regions

The long-term temporal sequences $\mathbf{D}_k$ in large w_k provide context clues for the obscured person regions, and capture the walking pattern periodically. The short-term temporal sequences $\mathbf{D}_k$ in small w_k preserve the detailed common appearance features robustly, with small variance for the individual features. The intermediate level changes (e.g., pose) are captured in the gradually-changing temporal scales. The representative examples are shown in Fig. 5. Besides, the long short-term appearance models motion feature maps are also shown in Fig. 5. Specially, the model construction is summarized in Algorithm 1.

Algorithm 1 Construction of Long Short-Term Appearance Model.

Input: The set of gallery/query images $\mathbf{D}$;
Output: Motion feature map $\mathbf{M}$.
1: Divide long-term batches $\mathbf{D}(f_r \in \mathcal{R}_A)$;
2: Decompose $\mathbf{D}_A = \mathbf{M}_A + \mathbf{B}_A$ via GoDec;
3: Divide short-term batches into multi-ranges $\mathbf{D}(f_r \in \mathcal{R}_B)$;
4: **for** $<r \in \mathcal{R}_B>$ **do**
5: Decompose $\mathbf{D}_B(f_r) = \mathbf{M}_B(f_r) + \mathbf{B}_B(f_r)$;
6: Update $\mathbf{M}_r$ based on Eq. 4;
7: **end for**

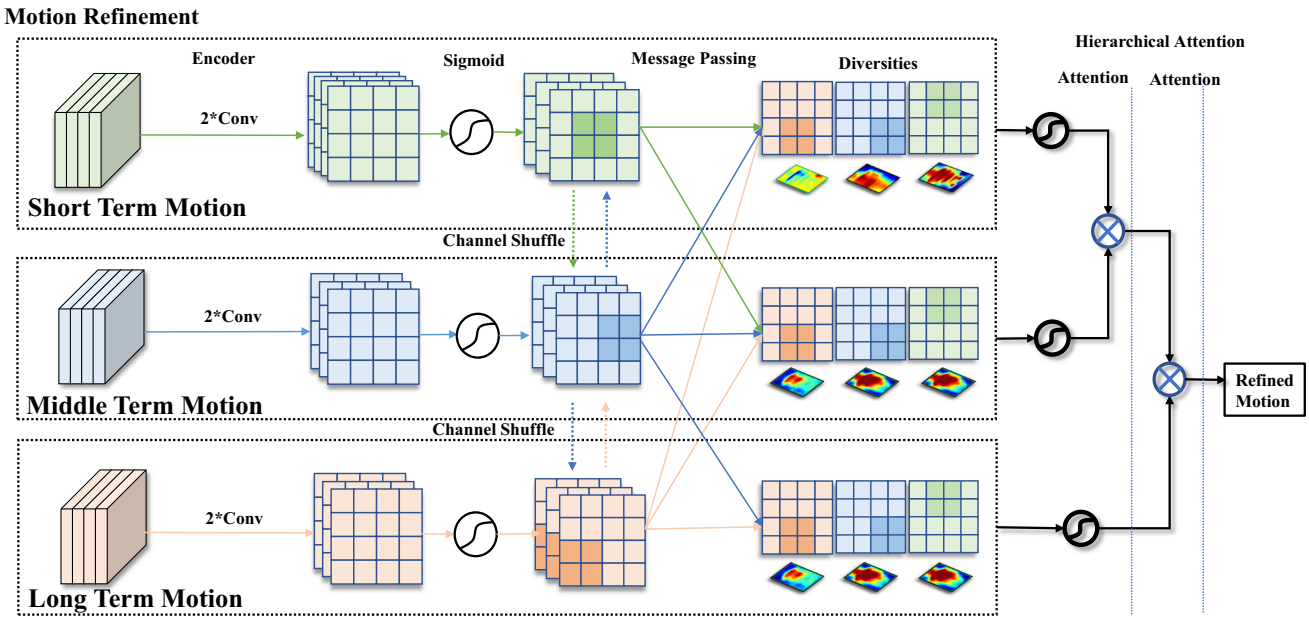

Fig. 6 Illustrations about motion refinement module: message passing and hierarchical attention mechanism. To capture the temporal clues from short-term to long-term video sequences, we first feed the sequences into the motion refinement module to encode temporal clues via a sequential encoder structure: a two-layer convolution (2*conv) and a Sigmoid function. We then employ the message passing operation to enable the channels to communicate and exchange information, which increases the diversities of the motion clues. Finally, we hierarchically employ the attention layer from local short-term to middle-term, and continue to long-term features. This module outputs the refined motion maps to the next layer

5 Temporal–Spatial Motion Clues Excited Deep Network

In this section, we introduce how to efficiently embed the temporal–spatial motion prior into the CNN to excite the person-related features.

5.1 Motion-Refinement Aggregation

We aim to endow the multi-range temporal features $\mathbf{M}(f_r)$ with person-related discriminative ability, and simultaneously avoid losing personally-distinct features during low-rank decomposition. Hence, we propose to tackle person Re-ID by refining the motion features, which can implicitly learn the discriminative compact encoding of the person's appearances and motions via specially-designed CNN layers.

As shown in Fig. 6, this network structure consists of two convolution layers with 3×3 kernel size, a sigmoid function to normalize motion values, and a network interaction mechanism, which is the 'motion shuffle' operation. There is a critical problem when we extract features via the long short-term appearance model, only a small fraction of interaction information among multiple-range video sequences can be derived. This phenomenon hinders the information flow between different motion maps resulted from multiple-range temporal clues and weakens the representation. To alleviate, motion passing aims to break the barriers between different motion feature channels. 'Motion passing' induces a channel-wise interactive network structure. This operation is extended from Zhang et al. (2018), which is then used to enable information communication between the two branches. Different from Zhang et al. (2018), we shuffle the motion features from the cross-scenario long short-term appearance model, instead of the group channels. In this manner, our motion clues could benefit from the multiple-range temporal clues. We first separately encode the pyramid motions, and then we use a sigmoid function for the feature maps, afterwards, the channel-wise transformer recursively moves the feature maps and shuffles the channels randomly, which bridges the gap between different motions. As a result, this operation increases the diversity of the motion maps, and make the temporal feature much richer.

To fuse all ranges of temporal clues, we carefully-design a hierarchical attention mechanism. The attention starts from the short-term feature map branch, afterwards, the feature map activates the middle-term feature map via element-wise product operation. The activated feature map is further combined with the long-term feature map to obtain more global clues from the cross-scenario frames. During the aggregating phase, we utilize the long-term temporal clues to activate the middle level clues. After cooperating the two levels, we

further utilize the two levels into the short-term clues. All the activation operation is implemented by self-attention network to adaptively learn from the critical regions. Finally, we obtain an importance map, which describes the person-relative degree in pixel-wise way.

In fact, the refine net incorporates the semantic features into motion features along two directions (as shown in Fig. 7): bottom-up and top-down directions. Here, the 'Semantic' features represent the person-related features. First, from the bottom-up direction, given the coarse motion regions, the message-passing layers increase the diversity and semantic representation ability of these regions, such that the refine net could let these feature maps convey much richer semantic clues. Furthermore, the self-attention mechanism activates the distinct regions related to the persons, which tends to refine the motion features in the discriminative semantic regions. At the end of the refine net, the motion feature maps are concatenated with the RGB channels, which are further fed into the 'Motion Excitation Blocks' layers for feature extraction. Second, from the top-down direction, these CNN layers are constrained via the cross entropy Softmax loss, which back propagates the semantic messages from the deep layers to the shallow layers, specifically, the back propagation loss of the refinement layer comes from the 'Motion Excitation Blocks'. Under the constraint of the loss function, the refinement module updates the self-attention layer, message passing layer, and motion encoder layer towards semantic features extraction.

5.2 Motion Excitation Block Design

We integrate the temporal motion feature map resulted from motion-refinement layers into CNN structure. We transit the motion feature map $\mathbf{M}$ in two orthogonal directions: embedding the motion feature map by making the network deeper and wider, which can concurrently activate the person-related regions in the CNN feature maps. Meanwhile, we employ the skip-connection via identity map to avoid the gradient vanishing problem. CNNs are ideal in hierarchically perceiving the spatial information. Therefore, we stack the temporal motion feature maps with the CNNs. In detail, we encode the global spatial information with a channel descriptor, so that the temporal motion feature maps can be respectively stacked as an additional channel, together with the original RGB image. This is achieved by using globally average pooling to generate channel-wise statistics. A feature map $\mathbf{M}_c$ can be generated by shrinking $x \in \mathbb{R}^c = (R, G, B, \mathbf{M_c})$ over the spatial feature maps with c channels, which can be calculated as

$$\mathbf{M}_c = \frac{1}{c}\sum_{c=1}^{N} F_c(x), \tag{5}$$

where F denotes the feature map function (e.g., convolution, pooling the filtered feature maps), and the transformation output $\mathbf{M}_c$ can be interpreted as a collection of local descriptors, whose statistics are expressive for the person-related regions.

The motion information is the most sensitive excitation for the human vision system among many kinds of features. To mimic the efficient mechanism, we model the 'motion' from the sequences. We simplify this operation as *motion excitation*, since it indicates how visual appearance and structure relates to dynamic motions in person identification.

The excitation motion map is crucial for learning the scene-independent, invariant, person-related features with the 'motion' component. In order to make the temporal motion activate the person motion related regions with equal-size perception, we add the identity map for every few stacked layers. The feature map in channel c can be formulated as,

$$Y(x, \mathbf{M}_c) = F(x \otimes \mathbf{M}_c) \oplus x \oplus \mathbf{M}_c, \tag{6}$$

where x is the input feature map from the stacking layers $x_0, \ldots, x_{l-1}$; $\otimes$ is the element-wise product operation, which is used to activate the feature map x with the motion-related component $\mathbf{M}_c$; and $\oplus$ is the channel-wise concatenation operation. A fundamental built-in block is shown in Fig. 8. The shortcut connections in Eq. 6 introduce neither extra parameter nor computation complexity, as demonstrated in He et al. (2016). The adaptive weights indicate the importance of motion-embedded frames at each spatial location and temporal channels. To make use of the information integrated in the temporal motion feature, we periodically embed the excitation block, which aims to fully capture the temporal relations in a skip-connection manner. More specifically, in the motion excitation block, all layers are directly connected to all subsequent layers. Namely, in order to make the temporal motion activate the regions related to the person motion, we employ the dense connection to build the 'motion excitation blocks' to pass the motion maps to all the subsequent layers. We first add the extra motion channel at the beginning of the network, then all the channels of each layer are passed to all the succeeding layers with concatenation connection, which strengthens the motion propagation ability. Each layer receives the feature maps of all the preceding layers, including the motion maps, such that the motion map could be fully passed through all the layers. This densely-connected network maximizes the influence of the motion clues, passing information from temporal to spatial. Afterwards, we add an extra skip connection on the motion maps between 'motion excitation blocks'. The skip connections employ the identity map to fully pass the motion to the middle and higher layers. These blocks are connected with the transition layers: a convolution layer, a max pooling layer, and a newly-added motion channel concatenation. It may be noted that, the con-

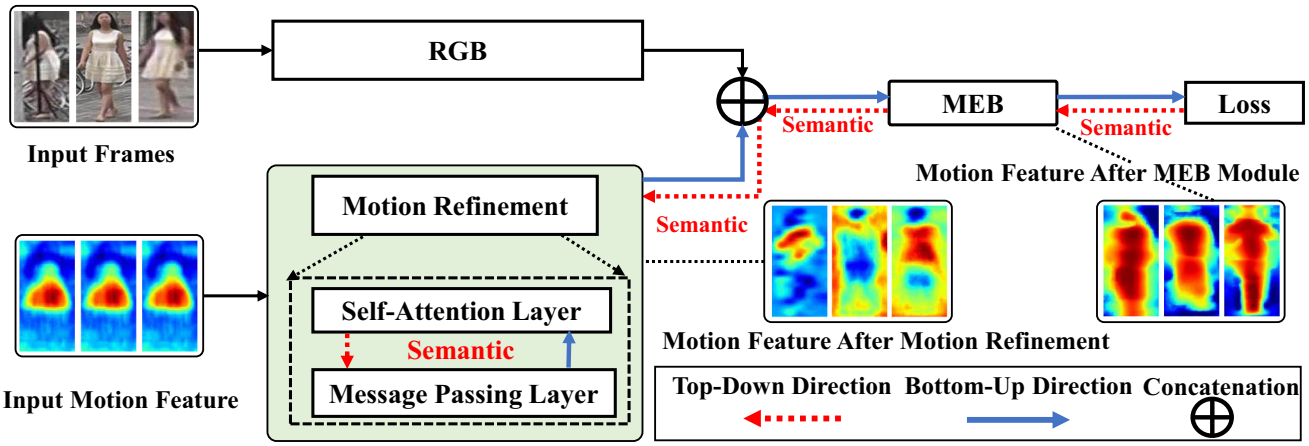

Fig. 7 Illustration about incorporating the semantic features into motion features. (1) From the bottom-up direction, the message-passing layer increases the diversity of the semantic regions. Furthermore, the self-attention mechanism activates the distinct regions related to the persons. The motion feature maps are concatenated with the RGB channels, which are further fed into the 'Motion Excitation Network'. (2) From the top-down direction, these CNN layers are constrained via the cross entropy Softmax loss, which back propagates the semantic label messages from the loss function, where the back propagation loss of the refinement layer comes from the succeeding layers

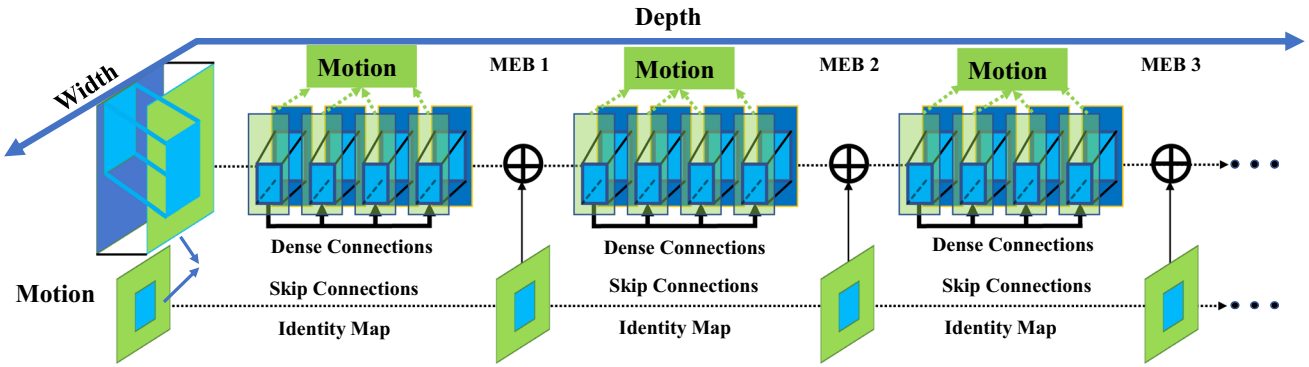

Fig. 8 Illustrations of the densely-connected motion excitation networks (3 blocks) in our LSTS-NET. Specifically, in the motion excitation block, all layers are directly connected to all subsequent layers, which strengthens the motion propagation ability. The green boxes show the motion maps, while the blue boxes denotes the spatial feature maps. The layers between the two adjacent blocks are concatenated with the 'motion' layer. The skip connections employ the identity map to fully pass the motion to middle and higher layers. These blocks are connected with the transition layer: a newly-added motion channel concatenation, a convolution layer, and a max pooling layer

catenation operations are employed to combine the spatial and temporal information flows based on the feature maps and motion maps.

Figure 8 illustrates the connections between the layers. Consequently, the l-th layer could receive the motion excited feature maps from all the preceding layers, and $\{x_0, \ldots, x_{l-1}\}$, $\{m_0, \ldots, m_{l-1}\}$ respectively represent the input sequential layers and the motion feature maps. It is formulated as,

$$Y_l = Y([x_0, \ldots, x_{l-1}], [m_0, \ldots, m_{l-1}]), \tag{7}$$

where [,...,] refers to the concatenation result of the feature maps produced in the layers $0, 1, \ldots, l$. In this densely-connected manner, the motion feature maps can excite all the inner block layers without losing pixels. Therefore, the inner block layers can extract the person-related features under the guidance of the 'motion'. The loss function of MEB networks is defined as:

$$L = -\sum_{k=1}^{N} (p_k * \log q_k), \tag{8}$$

where q is the ground-truth identity, q is the predicted label, N denotes the number of the input images.

In order to exhaustively analyze the optimal network structure for the fusion of 'motion' and spatial features, we investigate four structures: dense connection by adding auxiliary channels, short connection, couple connection by additional branches with shared weights, and separate con-

nection without shared weights, as shown in Fig. 9. Besides the first one, the short connection is similar to the 'bottleneck' in residual network. The third one combines the two pathes of the same network, one path trains the RGB image, the other one trains the motion feature. Finally, before the 'fc' layer, they are concatenated into a long feature vector. The fourth one has the similar structure with the third one, the difference is that the two paths do not share weights. MEB help recover the missing person-related features in two aspects: recovering the integrity of the person, and exciting the most critical regions with high attention values. There are two critical problems when the persons miss appearance features. (1) The integrity of the person will make the network bias towards noise (e.g., the distinct accessories, occlusions). Moreover, the network will be over-fitted to the person-unrelated distractors. (2) The person's motion prior has complex relations with person ReID tasks, but it is hard to determine the relevance in advance. Namely, the person-related temporal clues will play different roles on the specific background and person feature's discriminative ability. Confronting with the two main challenges, our MEB has two main advantages. (1) Our MEB embeds the recovered body regions (obtained from the motion appearance model) into the disturbed feature maps, which provides pedestrian's torso regions, including both generic and characteristic appearance details (long short-term sequences). (2) Our MEB is to learn the attention maps to describe the relevance to ReID tasks in the context of specific background. The enhanced attention mechanism could determine the importance of the motions at the pixel level, corresponding to the specific person's appearance, which guides the network to focus on the critical parts while ignoring the un-related regions.

Algorithm 2 The $t-th$ iteration of LSTS-NET's transfer process.

Input: The set of the gallery/query images $\mathbf{D}$;
Output: Person-related feature;
1: Extract feature map $F(\mathbf{D})$ from the last 'fc' layer of the *baseline networks (resnet50)*;
2: Construct the sequences set $\mathcal{R}_A, \mathcal{R}_B$ from $\mathbf{D}$;
3: Decompose the sequence $\mathbf{D}$ into 'motion' components $\mathbf{M}_c$ and background components $\mathbf{B}_c$ based on $\mathcal{R}_B, \mathcal{R}_A$;
4: Compute the final 'motion' guided by the pre-extracted semantic mask;
5: Output the final 'motion' to MEB network $Y(x, \mathbf{M}_c)$ for person-related feature.

6 Experiments and Evaluations

We verify our LSTS-NET's effectiveness on eight benchmarks, and compare it with various state-of-the-art architectures. In order to demonstrate the significant improvement benefiting from our LSTS-NET, we firstly conduct comparisons with baselines. To better understand the mechanism of our LSTS-NET, we then conduct comprehensive ablation and parameter analysis, meanwhile, we compare it with supervised state-of-the-art methods. Afterwards, we conduct extensive experiments to demonstrate the scene-independent ability of our method, and compare it with unsupervised state-of-the-art methods. Finally, we test the convergence of our LSTS-NET.

To be noted that, we also adapt our method to image-based datasets and compare it with other state-of-the-art methods. The Algorithm 2 describes the testing process without continuous video sequences as input. When only single image is available, the temporal sequences resulted from the images should be constructed with a newly-designed method. When we have camera information, we can simply cluster the images from the same camera as a group. The sequences with the same background will be easily decomposed by our long-short low-rank analysis. Otherwise, if we do not have camera ID and only single image is available, we should firstly encode the images as the features with a pre-trained resnet50 network, and predict the person's feature in each image. The similar ones will be clustered into the same group, and preserve their corresponding ranking list. When extracting the motion based on low-rank decomposition, they will be viewed as a continuous sequence.

6.1 Experiment Settings

6.1.1 Dataset Setting and Protocol

Our experiments are conducted on eight publicly available datasets, including PRW (Person Re-identification in the Wild) (Zheng et al. 2018), CUHK03-NP (Zhong et al. 2017a), Market (Market1501) (Zheng et al. 2015), MARS (Zheng et al. 2016), PRID (Hirzer et al. 2011), iLIDS Video re-Identification (iLIDS-VID) (Li et al. 2018), Duke-Video (Wang et al. 2014), and Duke (DukeMTMC-Re-ID) (Ristani et al. 2016). **Video-Based Datasets.** For PRID (Hirzer et al. 2011), we follow the evaluation protocol from Wang et al. (2014). Datasets are randomly split into probe/gallery identities. This procedure is repeated 10 times for computing averaged accuracy. MARS dataset totally has 1,191,993 person images, with a training/testing split of 509,914/681,089 images corresponding to 625 and 636 persons, respectively. iLIDS-VID dataset consists of 300 different pedestrians, of which, we use the training/testing split settings provided in the original paper. Duke-Video (Wang et al. 2014) dataset contains 702 gallery identities and 408 distractors. **Image-Based Datasets.** Market consists of 12,936 images for training, and 19,732 images are used for testing. The PRW dataset consists of 11,816 video frames, of which, 34,304 bounding boxes are assigned for 932 person IDs. The train-

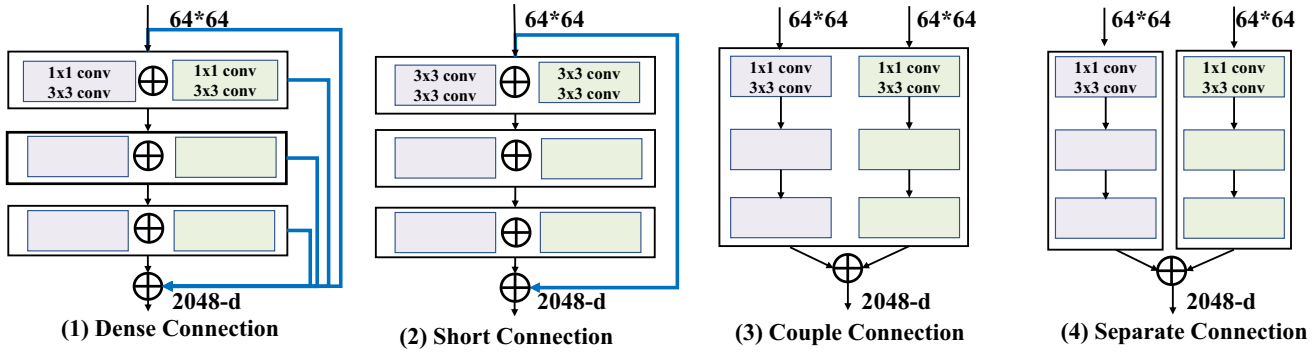

Fig. 9 Four kinds of connections for MEB. ⊕ denotes the concatenation of different channels. The layers between two adjacent blocks are concatenated with the motion layer

Table 2 Dataset statistics

Properties	PRID	Market	PRW	MARS	Duke (duke-video)	CUHK03-NP	iLIDS-VID
Camera per ID	2	6	6	6	8	2	3
Sequence length per ID	Long	Normal	Normal	Long	Long	Short	Long
Background	Simple	Norm	Normal	Complex	Complex	Normal	Complex
Occlusion	Little	Normal	Normal	Normal	Normal	Little	Heavy

The datasets are divided into three levels based on sequence length per ID, background, occlusion, respectively

ing/testing datasets include 482/450 different persons. The new protocol (Zhong et al. 2017a) splits the CUHK03-NP dataset into training set and testing set, similar to Market, which consists of 767 and 700 identities respectively. More detailed properties are summarized in Table 2.

6.1.2 Evaluation

In all of our experiments, we employ two protocols for different datasets. First, the Cumulative Matching Characteristics (CMC) curve is widely used in ReID task, which is a precision curve, documenting the detection or recognition precision for each rank. The horizontal line is the recognition rank, and the vertical line is the precision percentage. Here, the ReID task is considered as a ranking problem. Usually there is only one matched ground-truth result for a given query. We report the Rank-1, Rank-5, and Rank-10 scores in the CMC curve. The CMC metric is effective when each query corresponds to only one ground-truth clip in the gallery.

Second, the mean Average Precision (mAP) is a comprehensive metric, which can measure both single-matching and multiple-matching results. We use the single query mode and the general evaluation metrics (the same as previous works Zhong et al. 2017b): rank (r) 1, 5, 10 and mAP. For each query, its average precision (AP) is computed from its precision-recall curve. The mAP is calculated as the mean value of the average precisions across all queries. For ease of comparison, we only report the cumulated re-identification accuracy at selected ranks. In testing, we make sure that each query identity is selected from two cameras, so that the cross-camera search can be performed. In evaluation, the truly-matched images from the camera containing the query image are regarded as 'junk', which means that these images have no influence on Re-ID accuracy (CMC/mAP).

6.1.3 Implementation Details

On the Market, MARS, PRW, CUHK03-NP, Duke-Video, and Duke datasets, we first use the batch size of 300 for the long-term sequences ($\mathcal{R}_s$) from the same camera to decompose the background and the 'motion' $\mathbf{M}$, and use the batch size of 180 to extract the 'motion' component in different cameras' sequences ($\mathcal{R}_d$), with batch size of 30 in each camera. Based on the experiments, we observe that, the sequence longer than 300 frames can extract more motion features existing in more than one frames, and the sequence shorter than 300 frames can not extract the background sufficiently. We analyze the statistics, which will be detailed in 'parameter analysis' section. We further extract the short-term motion feature in the local pyramid with 1, 3, 5 continuous frames in sequences. On the PRID, Duke-Video, and iLIDS-VID dataset, the long-term sequences with 300 frames from the same camera and 100 (50 in each camera) frames from different cameras are used to extract the background. When adopting our LSTS-NET model, the batch size is 32.

All the networks are trained using stochastic gradient descent (SGD). On PRW, Market, PRID, iLIDS-VID, MARS, CUHK03-NP, Duke-Video and Duke datasets, we

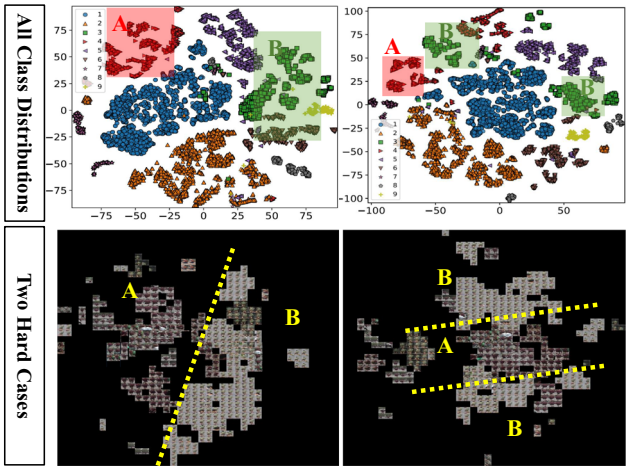

Fig. 10 Illustrations about the occluded persons's distributions via 't-distributed stochastic neighbor embedding' (van der Maaten 2014). The 1st column: LSTS-NET (with MEB), the 2nd column: Baseline Network (Densenet121 without MEB). 'A', 'B' respectively denotes two easily mixed-up cases due to the missing person-related features. It is obvious that, the MEB improves the discriminative ability of the occluded videos

use batch size of 32 for 40, 60, 40, 40, 60, 60, 60, and 60 epoches, respectively. Following Huang et al. (2017), the initial learning rate is set to be 0.1, and is respectively reduced 90% when the total training epochs reach to 50% and 75%, a weight decay of 10^{-4} and a Nesterov momentum of 0.9 without dampening are utilized. The weight initialization is set according to He et al. (2016). For all of the eight datasets, we don't use any data augmentation, we add a dropout layer (Srivastava et al. 2014) after each convolutional layer (except the first convolutional layer), and set the dropout rate to be 0.5. All the experiments are conducted on 4 P100 and 4 K80 Tesla GPUs.

6.2 Comparisons with Baselines

We have studied most of the recent works, of which, two high-performance baselines (He et al. 2016; Huang et al. 2017) are used as our backbone networks, both of which have achieved impressive results in classification tasks. We compare our method with the two baseline networks based on the experiments over all of the eight datasets. The results are documented in Table 3 (including the results on iLIDS-VID, MARS, and Duke-Video, while the results on PRW, CUHK03-NP, PRID, and Duke are provided in supplementary material), and it demonstrates that our LSTS-NET outperforms all the baselines on the eight datasets. The improvement of rank-1 accuracy ranges from 12.57 (96.81–84.24) to 21.62 (60.92–39.3)% over Duke-Video and iLIDS-VID datasets, as shown in Table 3. Specifically, on the MARS dataset, our LSTS-NET has an improvement of 8.32 (89.22–80.9) %. It is mainly because the motion is better modeled on this video sequences, which has higher-quality videos (1920×1080 resolution) and more continuous sequences than other datasets (average 200 frames). Besides, our LSTS-NET respectively gains large improvement w.r.t the baselines, because the two datasets both have more clutter backgrounds, accessories and poses variance, it states that these conditions can be well accommodated by our LSTS-NET.

6.3 Ablation Studies and Parameter Analysis

6.3.1 Technical Elements Analysis

We design ablation experiments to demonstrate the effectiveness of our novel technical elements, including long short-term appearance model, the motion-refinement (motion-passing, aggregating), and the MEB (Motion Excitation Block) network. We observe that the aforementioned technical elements will have consistent effects on different datasets, which is documented in Table 3. For example, on ILIDS-VID dataset, the LSTS-NET performs best, the improvement of the MEB (4) ranks the second, and the Low-rank component ranks the third. While on the dataset ILIDS-VID, the MEB gives rise to better performs than the other elements. With the increasing number of the MEB, the performance will be improved, however, after 4 times, the network will be too large, which may have the risk of causing over-fitting problem. Further, both motion-passing and aggregating mechanism plays critical roles in motion-refinement, because the low-rank decomposition may provide irrelevant clues for Re-ID task.

To demonstrate the contribution of our MEB in completing the pedestrians' missing features, we show the detailed results of some highly-occluded videos of 9 persons (e.g., occluded by clutter background, other pedestrians, buildings, etc.) in Fig. 10. The results show that our MEB could make the occluded pedestrians more discriminative.

Besides, we analyze the hard cases in Densenet121, which can be better accommodated by our LSTS-NET (images are provided in supplementary material). The hard cases mostly involve large view variances or different poses w.r.t the query, however, benefiting from the motion region excitation, the results are improved, because single image is insufficient to distinguish different persons with volatile non-trivial factors. The performance gap between the variants (Table 3) and our method confirms that, the motion clues facilitate to capture more stable features for identification.

According to the more results documented in Table 3, we can draw two main conclusions about the person-related and scenario-specific components. (1) 'Motion' can benefit the spatial features. Only taking 'motion' with one block MEB as input can gain a large margin compared with the baselines, showing that the motion region contains useful information

Table 3 Performance comparisons among baselines and our LSTS-NET with different-number MEBs on the MARS, ILIDS-VID, and Duke-Video datasets

Dataset model	ILIDS-VID				MARS				Duke-Video			
	r1	r5	r10	mAP	r1	r5	r10	mAP	r1	r5	r10	mAP
Original ResNet50 (He et al. 2016)	38.30	50.51	55.97	–	80.90	89.1	93.9	71.1	84.24	89.19	92.21	76.11
Original Densenet121 (Huang et al. 2017)	39.28	53.81	56.21	–	82.32	90.01	93.42	74.21	88.30	90.51	95.97	77.59
Low Rank +MEB (4)	48.09	59.05	66.71	–	87.14	92.25	96.24	81.39	92.15	95.92	98.22	83.01
Motion-Refinement+MEB (4)	39.70	58.79	63.16	–	85.24	90.98	94.12	76.12	91.70	96.03	96.34	80.92
Motion+MEB (1)	47.70	68.46	72.64	–	86.70	93.21	95.21	80.21	92.92	96.52	97.23	82.45
Motion+MEB (2)	52.47	73.12	79.92	–	87.05	93.81	96.20	81.33	93.82	97.26	98.27	89.82
Motion+MEB (3)	58.74	78.95	83.13	–	87.89	94.23	96.54	82.50	96.13	98.35	99.17	90.37
Motion+MEB (4) (LSTS-NET)	**60.92**	**82.81**	**88.63**	–	**89.22**	**96.81**	**97.96**	**83.12**	**96.81**	**99.32**	**99.80**	**93.91**

'MEB (n)' means to embed MEB n times into CNN. The bold indicates the best performance, while the italics indicates the second best performance

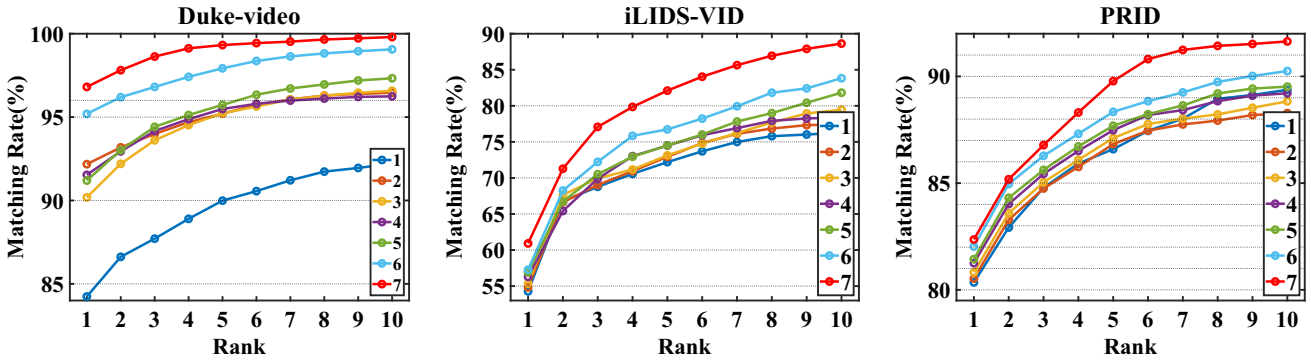

Fig. 11 Ablation Studies about long short-term appearance model

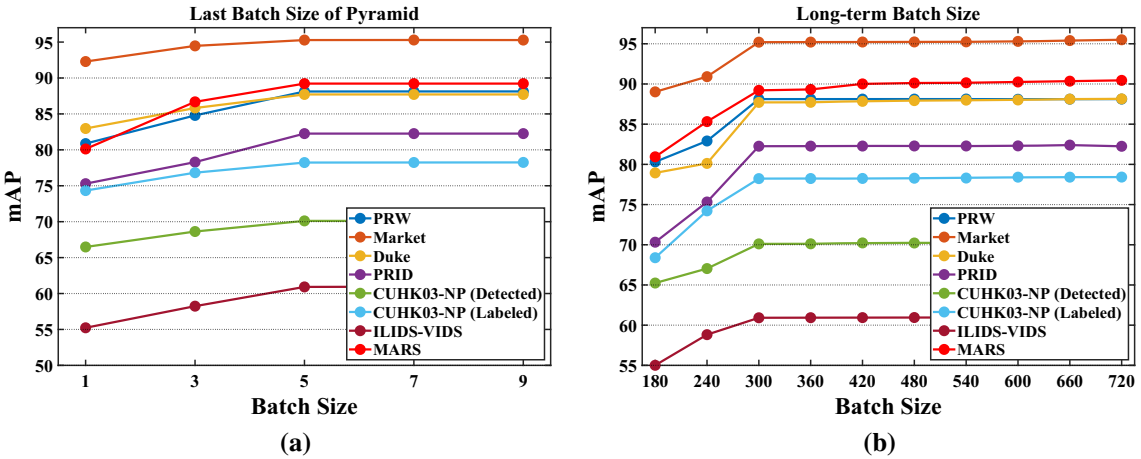

Fig. 12 Analysis about influences of different batch sizes in long short-term appearance model on eight datasets. When the long-term batch size is set to 300 and the short-term batch size is set to 1, 3, 5 in a pyramid structure, it achieves the best performance

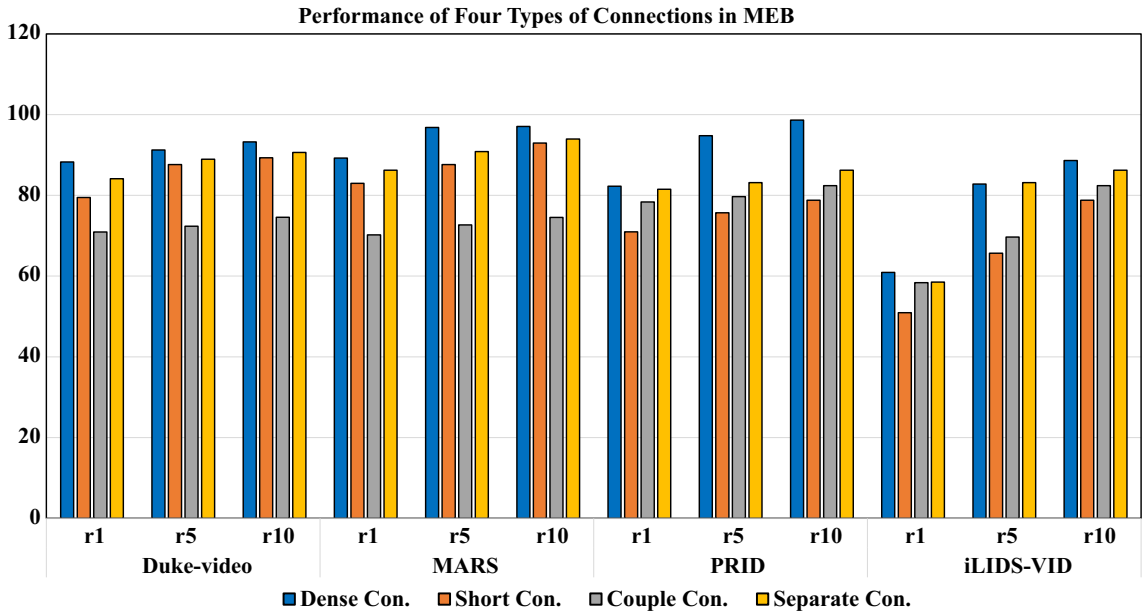

Fig. 13 Four types of connections in MEB. 'con.' denotes the connection

associated with the identity; (2) MEB improves the motion performance. Adding motion-auxiliary MEBs can achieve better accuracy than original baseline networks.

6.3.2 Long Short-Term Appearance Model Analysis

To further analyze how our long short-term appearance model works, we design another ablation study to evaluate the contribution of the long-term appearance model, short-term appearance model, and low-rank prior guided mechanism. One of the results on MARS dataset is documented in Table 4. Based on the performance of long short-term appearance model, on most of the datasets, long-term model performs the best, short-term model performs the second best. Based on the performance, we can classified the datasets into two categories. (1) Long-term clues preferred datasets, including PRID, iLIDS-VID, and MARS. These datasets mostly have long sequences, especially when the backgrounds are highly variable among different scenarios, the improvement gains are significantly great. (2) Motion-refinement preferred datasets, including Duke-video, this dataset mostly involves various camera views, especially when the backgrounds are more complex, the improvement gains are significantly great. For example, on the Duke dataset, the Motion-refinement gives rise to an improvement of 5.42% in rank-1 accuracy in Fig. 11. To clearly show the improvements, we document the ablation study results in Fig. 11.

6.3.3 Parameter Analysis

We analyze the critical parameters that determine the performance, wherein the batch size (in Eq. 4) is the most important parameter in our long short-term appearance model. We analyze its affects by fixing other parameters in our LSTS-NET. The long-term batch size for the intra-camera sequences ($\mathcal{R}_s$) ranges from 180 to 720 with a step of 30. The changes of batch size for the inter-camera sequences ($\mathcal{R}_d$) has few influences on the decomposition results. The last-layer batch size of the short-term sequence pyramid ranges from size 1–9, with a step of 2. We observe that, when the long-term batch size is set to 300 and the short-term batch size is set to 1, 3, 5 in a pyramid structure, it gives rise to the best performance. As the batch size increases, the performance keeps stable, however, the computation cost will increase. Therefore, we set global batch size to 300 and set pyramid batch size to 1, 3, 5. Experimental results are shown in Fig. 12a, b. This setting of temporal pyramid indicates the observations as follows. When the range of the long-term model is too large, the specific clues describing the person' feature are smoothed. Since the moving region in person-walking video is periodic, the clues in this period is too diverse to distinguish from other persons. Specifically, if the range exceeds this period, the low-rank component will preserve less unique information of the frame.

6.3.4 Network Structures Analysis

In order to sufficiently pass the temporal motion feature into the CNN for further spatial feature extraction, we compare the performance of three network structures on five representative datasets. The results are provided in Fig. 13, which show the superiority of dense connections for our LSTS-NET with MEB. The connection performs consistently on different datasets. When adding auxiliary person related motion feature, directly using the short connections will also increase the high-level semantic cues in discriminating different persons. While the two path connections lack the information transmitting between the two networks, the accuracies are lower than other two connections. The results demonstrate that more connections and more information transmitting from the temporal motion clues can increase the features' discrimination ability. Although the 'separate couple path connection' performs close to the dense connection induced MEB, its corresponding time consuming and memory cost are nearly 4 times higher than that of the dense connection induced MEB. Besides, it tends to cause over-fitting on MARS dataset. In summary, to better trade off the cost and the accuracy, we chose the dense connection induced MEB for our LSTS-NET framework.

6.4 Comparisons with State-of-the-Art Supervised Methods

We further compare our method with state-of-the-art supervised methods based on the experiments on the all of the eight datasets, of which, 4 are video-based datasets, another 4 are image-based datasets. Sixteen state-of-the-art Re-ID methods are utilized for comparison. They are classified into three categories. (1) Feature representation and metrics based methods: MGCAM (Song et al. 2018), DuATM (Si et al. 2018), AACN (Xu et al. 2018), Scalable (Bai et al. 2017), NPSM (Liu et al. 2017), OIM (Xiao et al. 2017), SVDNet (Sun et al. 2017), ACS (Huang et al. 2018), Pose Basel (D,Tri) (Liu et al. 2018), PSE (Sarfraz et al. 2018), TriNet+REDA (Zhong et al. 2017b); (2) Multi-range temporal context aggregation methods: SpindleNet (Zhao et al. 2017), Re-rank (Zhong et al. 2017a), ASTPN (Xu et al. 2017), SFT (Zhou et al. 2017); (3) Attention models: SPRe-IDcombined-ft* (Kalayeh et al. 2018), Consistent-Aware (Lin et al. 2017). Note that SPRe-IDcombined-ft* (Kalayeh et al. 2018) requires multiple datasets for training, thus, they apply data augmentation to generate more training samples on Market and Duke dataset, and it also uses multi-shot setting. Besides, AACN uses additional poses. For fair

Table 4 Component analysis for long short-term appearance model on MARS dataset

ID	Long term	Short term	Motion-Refinement	Motion-Passing	Aggregating	mAP	r1	r5	r10
1	✓	✓	✓	✓	✓	59.23	69.28	75.81	76.21
2	✓	✗	✗	✗	✗	65.11	78.59	81.90	81.69
3	✗	✓	✗	✗	✗	62.45	72.34	82.84	85.93
4	✗	✗	✗	✓	✗	77.68	86.94	88.72	93.13
5	✗	✗	✓	✓	✗	74.75	86.32	87.43	91.88
6	✗	✗	✗	✗	✓	78.43	87.16	89.15	94.32
7	✗	✗	✗	✗	✗	**83.12**	**89.22**	**96.81**	**97.96**

Table 5 Performance comparisons on MARS dataset

Methods	Conf.	r1	mAP
MGCAM (Song et al. 2018)	CVPR18	77.17	71.17
MGCAM-Siamese (Song et al. 2018)	CVPR18	76.01	70.13
MSCAN-bodySiamese (Li et al. 2017)	CVPR17	68.23	51.82
SFT (Zhou et al. 2017)	CVPR17	70.6	50.7
ASTPN (Xu et al. 2017)	ICCV17	44	–
IDE+XQDA (Zhong et al. 2017a)	CVPR17	70.51	55.12
MSCAN-Fusion (Li et al. 2017)	CVPR17	71.77	56.05
IDE+XQDA+Rerank (Zhong et al. 2017a)	CVPR17	73.94	68.45
DuATM (Si et al. 2018)	CVPR18	78.74	62.26
DRSA (Li et al. 2018)	CVPR18	82.3	–
PSE (Sarfraz et al. 2018)	CVPR18	76.7	71.8
ADFD (Zhao et al. 2019)	CVPR19	87	78.2
STCnet (Hou et al. 2019)	CVPR19	*88.5*	*82.3*
Ours (LSTS-NET)	–	**89.22**	**83.12**

Table 6 Performance comparisons on PRID and iLIDS-VID dataset (rank-1 accuracy %)

Method	PRID			ILIDS-VID		
	r1	r5	r10	r1	r5	r10
RFA (Yan et al. 2016)	58.2	85.5	93.4	49.3	76.8	85.3
RNN+OF (Wu et al. 2016)	70	90	95	58	84	91
RCN+KISSME (Zhang et al. 2018)	69	88.4	96.4	46.1	76.8	89.7
CNN+SRM+TAM (Zhou et al. 2017)	79.4	94.4	–	55.2	86.5	–
CAR (Zhang et al. 2017)	83.3	–	60.2	85.1	–	
QAN (Liu et al. 2017)	90.3	98.2	99.32	68	86.8	95.4
ST2N+TRL (Dai et al. 2019)	87.8	97.4	–	57.7	81.7	–
Salience (Zhao et al. 2013)	25.8	43.6	52.6	10.2	24.8	35.5
LOMO (Liao et al. 2015)	40.6	66.7	79.4	9.2	20	27.9
STFV3D (Liu et al. 2015)	42.1	71.9	84.4	37	64.3	77
DTW (Ma et al. 2017)	41.7	67.1	79.4	31.5	62.1	72.8
UnKISS (Khan and Bremond 2016)	58.1	81.9	89.6	35.9	66.3	74.9
SMP (Liu et al. 2017)	80.9	**95.6**	**98.8**	41.7	66.3	74.1
DGM+MLAPG (Ye et al. 2017)	73.1	92.5	96.7	37.1	61.3	72.2
TKP (Gu et al. 2019)	–	–	–	54.6	79.4	86.9
Ours	**82.26**	*94.78*	*98.64*	**60.92**	**82.81**	**88.63**

Table 7 Performance comparisons on Duke-Video and Duke datasets

Method	Conf.	r1	r5	r10	mAP
Basel LSRO (Zheng et al. 2017)	ICCV17	67.7	–	–	47.1
Basel OIM (Xiao et al. 2017)	CVPR17	68.1	–	–	–
SVDNet (Sun et al. 2017)	ICCV17	76.7	86.4	89.9	56.8
DuATM (Si et al. 2018)	CVPR18	81.16	92.47	–	67.73
AACN (Xu et al. 2018)	CVPR18	76.84	–	–	59.25
ACS (Huang et al. 2018)	CVPR18	84.11	–	–	78.19
Pose Basel (D,Tri) (Liu et al. 2018)	CVPR18	77.03	–	–	55.34
CAM (Yang et al. 2019)	CVPR19	85.8	–	–	72.9
HOM (Chen et al. 2019)	ICCV19	87.5	–	–	75.2
CAR (Zhou et al. 2019)	ICCV19	86.3	–	–	73.1
PGFA (Miao et al. 2019)	ICCV19	82.6	–	–	65.5
SSG (Fu et al. 2019)	ICCV19	76	85.8	89.3	60.3
PSE (Sarfraz et al. 2018)	*CVPR18*	*85.2*	–	–	*79.8*
Ours (LSTS-NET)	–	**87.71**	**94.31**	**96.23**	**80.86**
Ours (LSTS-NET) Video	–	**96.81**	**99.3**	**99.80**	**93.91**
STCnet (Hou et al. 2019) video	CVPR19	95	99.1	99.4	93.5
GLTR (Li et al. 2019) video	ICCV19	*96.29*	*99.3*	*99.71*	*93.74*
SPRe-IDcombined-ft* (Kalayeh et al. 2018) video	CVPR18	85.95	92.95	94.52	73.34
+re-ranking (Kalayeh et al. 2018) video	CVPR18	88.96	93.27	94.75	84.99

comparison, we use the same training-testing splits for the compared methods whenever possible.

6.4.1 Comparisons on Video Datasets

The results are shown in Table 5 (MARS), Table 6 (PRID and iLIDS-VID), and Table 7 (Duke-Video). The results show that, our LSTS-NET outperforms most of the state-of-the-art methods, except for the ones that cannot be fairly compared (extra training datasets or different protocols), and we list them in a separate sub-table for reference. Specifically, when comparing with the second best approach on each dataset, our LSTS-NET achieves 0.72%, and 0.52% improvement in rank-1 accuracy on MARS and Duke-video, respectively. On the small-scale dataset PRID and iLIDS-VID with relatively-simple backgrounds and camera views, our LSTS-NET also achieves competitive performance compared with other state-of-the-arts methods. The results indicate that, our LSTS-NET can well learn the temporal features to benefit the person Re-ID task. Since some works utilized extra datasets, such as DRSA (Li et al. 2018), and some works utilized multiple query, such as ST2N+TRL (Dai et al. 2019), thus, they cannot be compared with our LSTS-NET directly.

6.4.2 Comparisons on Image Datasets

Our LSTS-NET also could be applied to image based datasets, which is the salient advantages over other conventional methods. We conduct experiments on Market1501, PRW, and CUHK03-NP datasets, and the results are shown in Table 8 (PRW), Table 9 (Market), Table 7 (Duke-Video and Duke), Table 10 (CUHK03-NP). Specifically, when comparing with the second best approach on each dataset, our LSTS-NET gains 1.03%, 2.51%, and 2.8% improvement in rank-1 accuracy on CUHK03-NP (Labeled), Duke, and Market respectively. The results indicates that our LSTS-NET could extract clues for Re-ID tasks in image-based datasets. For example, on Market dataset, compared with SpindleNet, which uses person landmarks to pool features from person regions, our method is able to achieve higher re-identification performance, because the motion feature maps are able to provide more accurate human layout than the rough human joints. More importantly, SpindleNet and other existing methods do not consider how to handle the background-biasing problem. Benefitting from our LSTS-NET, though the rank-1 accuracy improvements on CUHK03-NP (Detected) dataset are slight, the mAP improvement is respectively 2.5%, which is still significant.

On PRW dataset, it is a retrieval problem, we focus on the identification ability. To fairly compare with the other state-of-the-art methods, we employ the widely-used faster R-CNN (Ren et al. 2015) as the detector with mAP of 80.14%, our LSTS-NET achieves significant improvements compared with person search methods. Our method outperforms the state-of-the-art Context-Graph (Yan et al. 2019) by 28.56%, 14.52% in terms of mAP, rank-1 accuracy. These results also demonstrate the effectiveness of our method.

Table 8 Performance comparisons on PRW dataset

Rank	Conf.	mAP	r1
CNN_v + IDNetOIM (Chen et al. 2018)	ECCV18	28.2	66.7
OIM (Xiao et al. 2017)	CVPR17	21.3	49.9
NPSM (Liu et al. 2017)	ICCV17	24.2	53.1
Context-Graph (Yan et al. 2019)	CVPR19	33.4	73.6
Resnet50	–	41.92	75.82
Densenet121	–	53.78	80.75
Ours (LSTS-NET)	–	**61.96**	**88.12**

'–' means no implementation code or no reported result is available

Table 9 Performance comparisons on market dataset

Method	Conf.	r1	r5	r10	mAP
MGCAM (Song et al. 2018)	CVPR18	83.55	–	–	74.25
MGCAM-Siamese (Song et al. 2018)	CVPR18	83.79	–	–	74.33
Scalable (Bai et al. 2017)	CVPR17	82.21	–	–	68.8
Consistent-Aware (Lin et al. 2017)	CVPR17	73.84	–	–	47.11
Spindle (Zhao et al. 2017)	CVPR17	76.9	91.5	94.6	–
Re-ranking (Zhong et al. 2017a)	CVPR17	77.11	–	–	63.63
GAN (Zheng et al. 2017)	ICCV17	78.06	–	–	56.23
MSCAN (Li et al. 2017)	CVPR17	80.31	–	–	57.53
DLPAR (Zhao et al. 2017a)	ICCV17	81	–	–	63.4
DaF (Yu et al. 2017)	BMVC17	82.3	–	–	72.42
SVDNet (Sun et al. 2017)	ICCV17	82.3	–	–	62.1
Basel. LSRO (Zheng et al. 2017)	ICCV17	84	–	–	66.1
Background (Tian et al. 2018)	CVPR18	81.2	94.6	97	–
DuATM (Si et al. 2018)	CVPR18	91.42	–	–	76.62
AACN (R.E) (Xu et al. 2018)	CVPR18	88.69	–	–	82.96
ACS (Huang et al. 2018)	CVPR18	88.66	–		83.3
Pose Basel (D,Tri) (Liu et al. 2018)	CVPR18	86.73	–	–	67.78
PSE (Sarfraz et al. 2018)	CVPR18	90.3	–	–	84
VPM (Sun et al. 2019)	CVPR19	*93*	*97.8*	*98.8*	*80.8*
Ours (LSTS-NET)	–	**95.8**	**97.9**	**98.85**	**93.0**
SPRe-IDcombined (Kalayeh et al. 2018)	CVPR18	94.63	96.82	97.65	90.96
UED (Bai et al. 2019)	CVPR19	95.9	–	–	92.75

Table 10 Performance comparisons on CUKU03-NP dataset

Method	Conf.	Labeled		Detected	
		r1	mAP	r1	mAP
MGCAM (Song et al. 2018)	CVPR18	49.29	49.89	46.29	46.74
MGCAM-Siamese (Song et al. 2018)	CVPR18	50.14	50.21	46.71	46.87
Re-rank (Zhong et al. 2017a)	CVPR17	38.1	40.3	34.7	37.4
ACS (Huang et al. 2018)	CVPR18	–	–	56.09	54.56
TriNet+REDA (Zhong et al. 2017b)	*ArXiv17*	58.1	53.8	55.5	50.7
SVDNet (Sun et al. 2017)	ICCV17	40.93	37.83	41.5	37.3
Pose Basel (D,Tri) (Liu et al. 2018)	CVPR18	45.1	42	41.6	37.26
IA (Hou et al. 2019)	CVPR19	77.2	72.4	**71.7**	*65.4*
Ours (LSTS-NET)	–	**78.23**	**72.3**	*70.11*	**67.9**

Table 11 Comparisons about unsupervised transfer learning ability (video-based and image and video involved datasets)

ID	Method	Source	Target	r1	r5	r10	mAP
1	**LSTS-NET**	**Duke**	**MARS**	**53.20**	**66.80**	**78.30**	**29.62**
2	LSTS-NET	CUHK03 (labeled)	MARS	11.80	16.20	18.10	8.97
3	RACE (Ye et al. 2018)	–	MARS	41.00	55.60	61.90	22.30
4	SMP (Liu et al. 2017)	–	MARS	23.59	35.81	44.90	–
5	DGM (Ye et al. 2017)	–	MARS	36.80	54.00	61.60	21.30
6	**LSTS-NET**	**Duke**	**CUHK03-NP (Labeled)**	**65.71**	**78.57**	**80.00**	**46.55**
7	LSTS-NET	Market	CUHK03-NP (Labeled)	57.14	61.43	67.14	45.02
8	LSTS-NET	MARS	CUHK03-NP (Labeled)	11.80	16.20	18.10	8.97
9	LSTS-NET	Market	Duke	45.43	49.43	51.98	32.40
10	LSTS-NET	CUHK03-NP (Labeled)	Duke	32.37	41.95	45.23	27.69
11	**LSTS-NET**	**MARS**	**Duke**	**68.25**	**75.22**	**81.98**	**46.80**
12	TAUDL (Li et al. 2018)	–	Duke	61.7	–	–	43.5
13	HHL (Zhong et al. 2018)	Market	Duke	46.90	–	–	27.20
14	**LSTS-NET**	**Duke**	**Market**	**58.51**	**64.71**	**66.74**	**32.57**
15	LSTS-NET	CUHK03-NP (Labeled)	Market	36.70	45.50	50.90	20.97
16	Disentangled (Ma et al. 2018)	–	Market	30.70	–	–	10.00

Table 11 continued

ID	Method	Source	Target	r1	r5	r10	mAP
17	PTG (Wei et al. 2018)	CUHK03	Market	27.80	–	54.60	–
18	PTG (Wei et al. 2018)	Transformed CUHK03	Market	31.50	–	60.20	–
19	HHL (Zhong et al. 2018)*	CUHK03-NP (Labeled)	Market	56.80	74.70	81.40	29.80
20	HHL (Zhong et al. 2018)*	Duke	Market	62.20	–	–	31.40
21	kLFDA-N (Xiong et al. 2014)	–	PRID	9.10	–	–	–
22	SADA+kLFDA (Xiong et al. 2014)	–	PRID	8.70	–	–	–
23	AdaRSVM (Ma et al. 2015)	–	PRID	4.90	–	–	–
24	GL (Kodirov et al. 2016)	–	PRID	25.00	–	–	–
25	UDML (Peng et al. 2016)	–	PRID	24.20	–	–	–
26	SSDAL (Su et al. 2016)	–	PRID	20.10	–	–	–
27	TJ-AIDLDuke (Wang et al. 2018)	Duke	PRID	34.80	–	–	–
28	TJ-AIDLMarket (Wang et al. 2018)	Market	PRID	26.80	–	–	–
29	**LSTS-NET**	**MARS**	**PRID**	**45.84**	**56.83**	**62.96**	**29.04**
30	LSTS-NET	Duke	PRID	38.91	46.79	58.92	22.21
31	LSTS-NET	CUHK03-NP (Labeled)	PRID	19.24	28.94	35.65	9.38

All the models are trained only on source dataset, and then are used to conduct unsupervised feature extraction on the target dataset

6.5 Comparisons with State-of-the-Art Unsupervised Methods

We conduct three kinds of experiments on six representative datasets (1) video datasets (2) image datasets (3) video and image mixed datasets, to demonstrate our LSTS-NET's unsupervised transfer learning ability. All the configurations of LSTS-NET are trained only on the source dataset. All the pre-trained models are then used to conduct unsupervised feature extraction on the target dataset.

6.5.1 Comparisons on Video Datasets

We firstly compare our LSTS-NET with the unsupervised baseline methods under the same dataset configuration. The results are shown in Table 11. With the pre-trained model on the MARS, PRID, ILIDS-VID and Duke-Video datasets separately, we directly use the second-to-last layer's output of our LSTS-NET to represent the feature of the target dataset. In most of the transfer learning cases, our LSTS-NET outperforms the baseline networks by a large margin. Specially,

in the case when the target dataset is 'MARS' and the source dataset is 'Duke-Video', as shown in the row (id = 11) marked with bold in Table 11, the transferred LSTS-NET network performs better than the supervised training networks, including the Resnet50 and the Densent121 networks on the Duke-Video dataset, compared with the rows 'Original ResNet50', 'Original Densenet121', 'Motion + MEB (4)' in Table 3. It may be attributed to the long range of video sequences in MARS dataset and Duke-Video dataset. Besides, the distinct motion features in the 'MARS' dataset are extracted from the original videos instead of the cropped human regions, which could benefit the Re-ID. However, most datasets only leave bounding box regions, which lose some of the context background. The performance comparisons demonstrate that, our motion feature excitation network can significantly improve the performance of cross-scenario person Re-ID.

6.5.2 Comparisons on Image Datasets

We compare our LSTS-NET with the unsupervised configuration on image-based datasets. The results are shown in Table 12. With the pre-trained model on the CUHK03-NP(Detected, Labeled), Duke, PRW, and Market datasets separately, in most of the transfer learning cases, our LSTS-NET outperforms the baseline networks by a large margin. Specially, in the case when the target dataset is 'Duke' and the source dataset is 'CUHK03-NP(Labeled)', as shown in the row (id = 6) marked with bold in Table 12, the transferred LSTS-NET network performs better than the supervised training networks. It may be attributed to the multiple cameras dataset, from which it can extract more general motion clues across scenarios.

6.5.3 Comparisons on Video and Image Involved Datasets

Our LSTS-NET can be flexibly applied on both video and image involved datasets. We compare our LSTS-NET on video and image involved datasets by pre-training on image/video datasets and testing on other image/video datasets. The results in Table 12 show that: (1) training on video based datasets could improve the testing performance on image-based datasets (id = 8, 11, 29) by a large margin, since our person-related component are stable across scenarios; (2) training on image based datasets (id = 1, 2, 30, 31) could improve the testing performance (ranging from 4.11 to 12.2% in rank-1 accuracy) on video-based datasets, benefitting from the prior of clutter background component obtained from image datasets.

We further conduct another experiment to compare our method with two categories of existing unsupervised Re-ID methods: (1) Person-related feature extraction methods without transfer learning, including DGM (Ye et al. 2017), RACE (Ye et al. 2018), and graph learning based model (GL) (Kodirov et al. 2016), those features are designed to be view invariant; (2) Source identity and attribute knowledge based transfer methods, including HHL (Zhong et al. 2018), kLFDA-N (Xiong et al. 2014), SADA+kLFDA (Xiong et al. 2014), AdaRSVM (Ma et al. 2015), UDML (Peng et al. 2016), SSDAL (Su et al. 2016), TJ-AIDLDuke (Wang et al. 2018), and SMP (Liu et al. 2017), these methods tend to encode the attributes as intrinsic pedestrian features.

From Table 11, we can draw the following conclusions. (1) Our method outperforms all existing state-of-the-art models, improving the rank-1 accuracy by 12.2 (53.2–41.00)%, 6.93 (45.84–38.91)%, 21.81 (58.51–36.7)%, 6.55 (68.25–61.7)% w.r.t the previous best performance method on MARS/PRID/Market/Duke datasets, respectively. On CUHK03-NP (Labeled) dataset, we achieve rank-1 accuracy of 78.23%. On Market dataset, the performance our LSTS-NET is close to the second best one in rank-1 accuracy. Besides, our method cannot be directly compared with HHL (Zhong et al. 2018), since it assumes the target domain is known. This proves the overall advantages of our LSTS-NET in capturing the condition-independent pedestrian component for cross-domain unsupervised Re-ID. (2) When learning from the longer sequences, the performance will be better. For example, when using MARS as the source dataset, the target dataset outperforms the Duke dataset. This indicates the importance of learning temporal motion clues in cross-domain Re-ID tasks. It also means that, using more supervision in cross-domain transfer learning is non-trivial, particularly when the target dataset has clutter backgrounds, and the long sequences will introduce more irrelevant components. It proves the advantages of our LSTS-NET in exploiting the diverse knowledge from different types of labeled data.

Finally, it is worth noted that, the performance gains of our LSTS-NET are achieved with much less supervision data, which lacks diversity (only from one source dataset: 16,522 images of 702 identities/classes on Duke, or 12,936 images of 751 identities on Market) w.r.t the competitors. For example, the methods in the second category use 7 different person Re-ID datasets with high varieties, including a total of 44,685 images and 3791 identities. The UDML (Peng et al. 2016) leverages the datasets from three different source domains. The SSDAL (Su et al. 2016) benefits from 10 diverse datasets, consisting of 19,000 images with 8705 person identities and another 20,000 images with 1221 person tracklets. They all use much more supervised datasets, however, the performance is worse than ours.

6.6 Efficiency and Convergence Analysis

The time cost of our LSTS-NET involves two main parts: the long-short appearance modeling and the spatial feature

Table 12 Comparisons about unsupervised transfer learning ability on image-based datasets

Method	Source	Target	Performance			
			r1	r5	r10	mAP
Resnet50	PRW	Market	52.70	65.56	70.96	35.93
	Duke	Market	42.46	53.47	58.58	21.55
	Market	PRW	89.22	94.63	96.86	85.12
	Duke	PRW	24.12	28.41	31.41	14.12
	PRW	Duke	22.13	34.29	38.56	18.42
	Market	Duke	30.21	40.75	46.72	20.72
Densenet121	PRW	Market	85.54	91.33	93.44	79.66
	Duke	Market	45.23	55.56	60.96	25.93
	Market	PRW	92.64	96.14	97.15	89.06
	Duke	PRW	28.24	32.85	41.54	18.23
	PRW	Duke	29.35	38.91	44.57	20.46
	Market	Duke	42.21	45.25	50.29	26.83
LSTS-NET	PRW	Market	86.25	96.64	95.09	82.91
	Duke	Market	58.51	64.71	66.74	32.57
	Market	PRW	78.51	84.71	86.74	70.57
	Duke	PRW	41.45	52.41	53.12	28.25
	PRW	Duke	53.50	65.83	71.38	36.97
	Market	Duke	45.43	49.43	51.98	32.40
TJ-AIDJ (Wang et al. 2018)	Duke	Market	58.2	–	–	26.5
	Market	Duke	44.3	–	–	23

All the models are trained only on the source dataset, and then are used to conduct unsupervised feature extraction on the target dataset

extraction. The temporal motion prior modeling includes short-term and long-term frame-spanned matrix decomposition. The spatial feature extraction part is processed offline on K80 GPU. Low-rank decomposition and motion integration respectively cost 10 ± 4 ms, 1000 ± 360ms, and 200 ± 10 ms. The result shows that, our efficiency bottleneck is the semantic label prediction, therefore, there should be a trade-off between the time cost and accuracy. The simplified LSTS-NET (only with the low-rank analysis component, without semantic labels) has the fastest speed and can also give rise to significant performance improvement compared with the baseline networks (see Table 3). In the training stage, the convergence speeds of our LSTS-NET and the baselines (including the ResNet50 and Densenet121) are shown in Fig. 14. The result shows that, our LSTS-NET needs a similar number of epoches as the baselines to reach convergence, since the training process of our LSTS-NET does not bring extra computation cost in average. (*Please refer to our supplementary materials for more experiment results and evaluations.*)

7 Discussion and Conclusion

7.1 Discussion and Limitation

For extremely complicated scenes, our LSTS-NET has shown superiority in accommodating cross-scene datasets with clutter backgrounds and large camera view variations. For scenes with simple background and without any camera variance, our LSTS-NET only exhibits slight improvement, which mainly results from the temporal clue aggregation. For example, on PRID dataset, the involved two cameras have similar views, whose background is also relatively simple, but the difficulty stems from complex illumination, our LSTS-NET gains slight improvement in rank-1 accuracy. The temporal sequences can provide clues for both person integrity and motion field, but can hardly handle other challenges (e.g., illumination, low resolution) so far. In terms of 'motion' and background component decomposition, the components corresponding to motion and background might swap their roles with each other on different kinds of datasets. For example, consider an image in its entirety, the motion should belong to the sparse component and the background should belong to the low-rank component in a generic low-rank analysis framework. However, in images with cropped pedestrians, the motion should belong to the low-rank com-

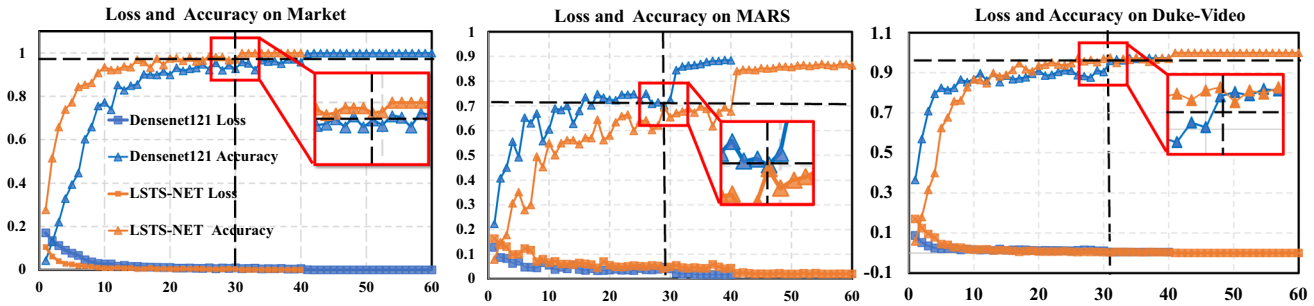

Fig. 14 Loss and validation accuracy comparisons. On both Market and Duke-Video datasets, our LSTS-NET converges faster than the Densenet121. On MARS dataset, our LSTS-NET converges slightly slower than Densenet121

ponent while the background should belong to the sparse component. In terms of single image/video sequence, our LSTS-NET relies on the video sequences with similar backgrounds to extract the person-relevant features. In the case where only single image is available, our method will become degenerated to a single-level pyramid based long-short appearance model.

7.2 Conclusion and Future Work

In this paper, we have detailed a novel LSTS-NET architecture for robust and high-performance scene-independent person Re-ID. In particular, our LSTS-NET could transfer both 'motion' and background information across different scenes by integrating the temporal–spatial motion priors learned from unsupervised low-rank analysis. Furthermore, we proposed a motion excitation scheme to enhance the person-related spatial feature extraction. Experiments showed that, our LSTS-NET outperforms the state-of-the-art person Re-ID methods on the scene-independent datasets by a large margin. In the near future, we shall generalize our LSTS-NET to handle other critical tasks with more flexible and relaxed scene conditions, such as raining day and night time. Since our method is efficient in the intrinsic representation of cross-scene video contexts, our method can also contribute to other video-related tasks that are sensitive to cross-scene appearance features. In particular, when the objective of certain video processing tasks become independent of the subjects and varieties of different scenes, our method has potential to gain significant improvement. For example, both the action recognition and pose estimation tasks' performance could deteriorate when encountering occlusions in short-term sequences, but our new LSTS-NET could provide valuable long-short temporal clues to benefit the intrinsic feature extraction in a more flexible temporal range, which promises performance improvement. Besides, in our future work, we should consider converting the low-rank decomposition into a neural network based model and training everything all together, so that the matrix decomposition could be uniformly converted to the convolution operations. However, neural net based decomposition will become a pixel-level task, which may require higher computation cost. Hence, to sufficiently exploit the advantage of our LSTS-NET, we should explore an efficient on-line decomposition approach to easily adapt for more computer vision tasks.

Acknowledgements National Key R & D Program of China (No. 2018YFB1700603), National Natural Science Foundation of China (NO. 61672077 and 61532002), Beijing Natural Science Foundation Haidian Primitive Innovation Joint Fund (L182016), National Science Foundation of USA under Grant IIS-1715985 and IIS-1812606.

References

Bai, S., Bai, X., & Tian, Q. (2017). Scalable person re-identification on supervised smoothed manifold. In *CVPR* (pp. 2530–2539).

Bai, S., Tang, P., Torr, P. H., & Latecki, L. J. (2019). Re-ranking via metric fusion for object retrieval and person re-identification. In *CVPR* (pp. 740–749).

Burr, D. C., & Santoro, L. (2001). Temporal integration of optic flow, measured by contrast and coherence thresholds. *Vision Research*, *41*(15), 1891–1899.

Candès, E. J., Li, X., Ma, Y., & Wright, J. (2011). Robust principal component analysis? *JACM*, *58*(3), 11.

Chen, B., Deng, W., & Hu, J. (2019). Mixed high-order attention network for person re-identification. In *ICCV* (pp. 371–381).

Chen, D., Li, H., Xiao, T., Yi, S., & Wang, X. (2018). Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding. In *CVPR* (pp. 1169–1178).

Chen, D., Zhang, S., Ouyang, W., Yang, J., & Tai, Y. (2018). Person search via a mask-guided two-stream CNN model. In *ECCV* (pp. 734–750).

Dai, J., Zhang, P., Wang, D., Lu, H., & Wang, H. (2019). Video person re-identification by temporal residual learning. *TIP*, *28*(3), 1366–1377.

Fu, Y., Wang, X., Wei, Y., & Huang, T. S. (2019). Sta: Spatial–temporal attention for large-scale video-based person re-identification. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 33(01), pp. 8287–8294).

Fu, Y., Wei, Y., Wang, G., Zhou, Y., Shi, H., & Huang, T. S. (2019). Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification. In *ICCV* (pp. 6112–6121).

Gu, X., Ma, B., Chang, H., Shan, S., & Chen, X. (2019). Temporal knowledge propagation for image-to-video person re-identification. In *ICCV* (pp. 9647–9656).
He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *CVPR* (pp. 770–778).
Hirzer, M., Beleznai, C., Roth, P. M., & Bischof, H. (2011). Person re-identification by descriptive and discriminative classification. In *SCIA* (pp. 91–102).
Hou, R., Ma, B., Chang, H., Gu, X., Shan, S., & Chen, X. (2019). Interaction-and-aggregation network for person re-identification. In *CVPR* (pp. 9317–9326).
Huang, G., Liu, Z., Weinberger, K. Q., & van der Maaten, L. (2017). Densely connected convolutional networks. In *CVPR* (pp. 4700–4708).
Huang, H., Li, D., Zhang, Z., Chen, X., & Huang, K. (2018). Adversarially occluded samples for person re-identification. In *CVPR* (pp. 5098–5107).
Kalayeh, M. M., Basaran, E., Gokmen, M., Kamasak, M. E., & Shah, M. (2018). Human semantic parsing for person re-identification. In *CVPR*.
Karanam, S., Gou, M., Wu, Z., Rates-Borras, A., Camps, O., & Radke, R. J. (2018). A systematic evaluation and benchmark for person re-identification: Features, metrics, and datasets. In *TPAMI* (p. 1).
Khan, F. M., & Bremond, F. (2016). Unsupervised data association for metric learning in the context of multi-shot person re-identification. In *AVSS* (pp. 256–262).
Kodirov, E., Xiang, T., Fu, Z., & Gong, S. (2016). Person re-identification by unsupervised L1 graph learning. In *ECCV* (pp. 178–195).
Li, D., Chen, X., Zhang, Z., & Huang, K. (2017). Learning deep context-aware features over body and latent parts for person re-identification. In *CVPR* (pp. 384–393).
Li, J., Wang, J., Tian, Q., Gao, W., & Zhang, S. (2019). Global-local temporal representations for video person re-identification. In *ICCV* (pp. 3958–3967).
Li, J., Zhang, S., & Huang, T. (2020). Multi-scale temporal cues learning for video person re-identification. *IEEE Transactions on Image Processing*, *29*, 4461–4473.
Li, M., Zhu, X., & Gong, S. (2018). Unsupervised person re-identification by deep learning tracklet association. In *ECCV* (pp. 737–753).
Li, M., Zhu, X., & Gong, S. (2019). Unsupervised tracklet person re-identification. In *TPAMI*.
Li, S., Bak, S., Carr, P., & Wang, X. (2018). Diversity regularized spatiotemporal attention for video-based person re-identification. In *CVPR* (pp. 369–378).
Li, S., Shao, M., & Fu, Y. (2017). Person re-identification by cross-view multi-level dictionary learning. In *TPAMI* (p. 1).
Li, W., Zhu, X., & Gong, S. (2019a). Scalable person re-identification by harmonious attention. In *IJCV* (pp. 1–19).
Li, W., Zhu, X., & Gong, S. (2019b). Harmonious attention network for person re-identification. *IEEE Access*, *7*, 22457–22470.
Liao, S., Hu, Y., Zhu, X., & Li, S. Z. (2015). Person re-identification by local maximal occurrence representation and metric learning. In *CVPR* (pp. 2197–2206).
Lin, J., Ren, L., Lu, J., Feng, J., & Zhou, J. (2017). Consistent-aware deep learning for person re-identification in a camera network. In *CVPR* (pp. 5771–5780).
Liu, H., Feng, J., Jie, Z., Karlekar, J., Zhao, B., Qi, M., et al. (2017). Neural person search machines. In *ICCV* (pp. 493–501).
Liu, H., Jie, Z., Jayashree, K., Qi, M., Jiang, J., Yan, S., & Feng, J. (2017). Video-based person re-identification with accumulative motion context. *TCSVT*.
Liu, J., Ni, B., Yan, Y., Zhou, P., Cheng, S., & Hu, J. (2018). Pose transferrable person re-identification. In *CVPR* (pp. 4099–4108).
Liu, K., Ma, B., Zhang, W., & Huang, R. (2015). A spatio-temporal appearance representation for viceo-based pedestrian re-identification. In *ICCV* (pp. 3810–3818).
Liu, Y., Yan, J., & Ouyang, W. (2017). Quality aware network for set to set recognition. In *CVPR* (pp. 4694–4703).
Liu, Z., Wang, D., & Lu, H. (2017). Stepwise metric promotion for unsupervised video person re-identification. In *ICCV* (pp. 2448–2457).
Lv, J., Chen, W., Li, Q., & Yang, C. (2018). Unsupervised cross-dataset person re-identification by transfer learning of spatial–temporal patterns. In *CVPR* (pp. 7948–7956).
Ma, A. J., Li, J., Yuen, P. C., & Li, P. (2015). Cross-domain person reidentification using domain adaptation ranking SVMS. *TIP*, *24*(5), 1599–1613.
Ma, L., Sun, Q., Georgoulis, S., Van Gool, L., Schiele, B., & Fritz, M. (2018). Disentangled person image generation. In *CVPR* (pp. 99–108).
Ma, X., Zhu, X., Gong, S., Xie, X., Hu, J., Lam, K.-M., et al. (2017). Person re-identification by unsupervised video matching. *PR*, *65*, 197–210.
McLaughlin, N., del Rincon, J. M., & Miller, P. (2016). Recurrent convolutional network for video-based person re-identification. In *CVPR* (pp. 1325–1334).
Miao, J., Wu, Y., Liu, P., Ding, Y., & Yang, Y. (2019). Pose-guided feature alignment for occluded person re-identification. In *ICCV* (pp. 542–551).
Peng, P., Tian, Y., Xiang, T., Wang, Y., Pontil, M., & Huang, T. (2018). Joint semantic and latent attribute modelling for cross-class transfer learning. *TPAMI*, *40*(7), 1625–1638.
Peng, P., Xiang, T., Wang, Y., Pontil, M., Gong, S., Huang, T., & Tian, Y. (2016). Unsupervised cross-dataset transfer learning for person re-identification. In *CVPR* (pp. 1306–1315).
Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. In *NIPS* (pp. 91–99).
Ristani, E., Solera, F., Zou, R., Cucchiara, R., & Tomasi, C. (2016). Performance measures and a data set for multi-target, multi-camera tracking. In *ECCVW*.
Sarfraz, M. S., Schumann, A., Eberle, A., & Stiefelhagen, R. (2018). A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. In *CVPR* (pp. 420–429).
Si, J., Zhang, H., Li, C.-G., Kuen, J., Kong, X., Kot, A. C., & Wang, G. (2018). Dual attention matching network for context-aware feature sequence based person re-identification. In *CVPR* (pp. 5363–5372).
Simonyan, K., & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. In *NIPS* (pp. 568–576).
Song, C., Huang, Y., Ouyang, W., & Wang, L. (2018). Mask-guided contrastive attention model for person re-identification. In *CVPR* (pp. 1179–1188).
Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *JMLR*, *15*(1), 1929–1958.
Su, C., Zhang, S., Xing, J., Gao, W., & Tian, Q. (2016). Deep attributes driven multi-camera person re-identification. In *ECCV* (pp. 475–491).
Subramaniam, A., Nambiar, A., & Mittal, A. (2019). Co-segmentation inspired attention networks for video-based person re-identification. In *ICCV* (pp. 562–572).
Sun, Y., Xu, Q., Li, Y., Zhang, C., Li, Y., Wang, S., et al. (2019). Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification. *CVPR* (pp. 393–402).
Sun, Y., Zheng, L., Deng, W., & Wang, S. (2017). Svdnet for pedestrian retrieval. In *ICCV* (pp. 3800–3808).
Tay, C.-P., Roy, S., & Yap, K.-H. (2019). Aanet: Attribute attention network for person re-identifications. In *CVPR* (pp. 7134–7143).

Tian, M., Yi, S., Li, H., Li, S., Zhang, X., Shi, J., Yan, J., & Wang, X. (2018). Eliminating background-bias for robust person re-identification. In *CVPR* (pp. 5794–5803).
van der Maaten, L. (2014). Accelerating t-sne using tree-based algorithms. *JMLR*, *15*(93), 3221–3245.
Wang, H., Zhu, X., Gong, S., & Xiang, T. (2018). Person re-identification in identity regression space. *IJCV*, *126*(12), 1288–1310.
Wang, J., Zhu, X., Gong, S., & Li, W. (2018). Transferable joint attribute-identity deep learning for unsupervised person re-identification. In *CVPR* (pp. 2275–2284).
Wang, T., Gong, S., Zhu, X., & Wang, S. (2014). Person re-identification by video ranking. In *ECCV* (pp. 688–703).
Wang, T., Gong, S., Zhu, X., & Wang, S. (2016). Person re-identification by discriminative selection in video ranking. *TPAMI*, *38*(12), 2501–2514.
Wei, L., Zhang, S., Gao, W., & Tian, Q. (2018). Person transfer Gan to bridge domain gap for person re-identification. In *CVPR* (pp. 79–88).
Wu, L., Shen, C., & Hengel, A. V. D. (2016). *Deep recurrent convolutional networks for video-based person re-identification: An end-to-end approach*. CoRR. arXiv:1606.01609.
Wu, Y., Lin, Y., Dong, X., Yan, Y., Ouyang, W., & Yang, Y. (2018). Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning. In *CVPR* (pp. 5177–5186).
Xia, B. N., Gong, Y., Zhang, Y., & Poellabauer, C. (2019). Second-order non-local attention networks for person re-identification. In *ICCV* (pp. 3760–3769).
Xiao, T., Li, S., Wang, B., Lin, L., & Wang, X. (2017). Joint detection and identification feature learning for person search. In *CVPR* (pp. 3376–3385).
Xiong, F., Gou, M., Camps, O., & Sznaier, M. (2014). Person re-identification using kernel-based metric learning methods. In *ECCV* (pp. 1–16).
Xu, J., Zhao, R., Zhu, F., Wang, H., & Ouyang, W. (2018). Attention-aware compositional network for person re-identification. In *CVPR* (pp. 2119–2128).
Xu, S., Cheng, Y., Gu, K., Yang, Y., Chang, S., & Zhou, P. (2017). Jointly attentive spatial–temporal pooling networks for video-based person re-identification. In *ICCV* (pp. 4733–4742).
Yan, Y., Ni, B., Song, Z., Ma, C., Yan, Y., & Yang, X. (2016). Person re-identification via recurrent feature aggregation. In *ECCV* (pp. 701–716). Berlin: Springer.
Yan, Y., Zhang, Q., Ni, B., Zhang, W., Xu, M., & Yang, X. (2019). Learning context graph for person search. In *CVPR* (pp. 2158–2167).
Yang, W., Huang, H., Zhang, Z., Chen, X., Huang, K., & Zhang, S. (2019). Towards rich feature discovery with class activation maps augmentation for person re-identification. In *CVPR* (pp. 1389–1398).
Ye, M., Lan, X., & Yuen, P. C. (2018). Robust anchor embedding for unsupervised video person re-identification in the wild. In *ECCV* (pp. 170–186).
Ye, M., Ma, A. J., Zheng, L., Li, J., & Yuen, P. C. (2017). Dynamic label graph matching for unsupervised video re-identification. In *ICCV* (pp. 5152–5160).
Yu, H.-X., Wu, A., & Zheng, W.-S. (2017). Cross-view asymmetric metric learning for unsupervised person re-identification. In *ICCV* (pp. 994–1002).
Yu, R., Zhou, Z., Bai, S., & Bai, X. (2017). Divide and fuse: A re-ranking approach for person re-identification. In *BMVC*.
Zeng, Z., Chan, T.-H., Jia, K., & Xu, D. (2012). Finding correspondence from multiple images via sparse and low-rank decomposition. In *ECCV* (pp. 325–339). Berlin: Springer.
Zhang, W., Hu, S., & Liu, K. (2017). Learning compact appearance representation for video-based person re-identification. arXiv:1702.06294.
Zhang, R., Li, J., Sun, H., Ge, Y., Luo, P., Wang, X., et al. (2019). Scan: Self-and-collaborative attention network for video person re-identification. *IEEE Transactions on Image Processing*, *28*(10), 4870–4882.
Zhang, W., Yu, X., & He, X. (2018). Learning bidirectional temporal cues for video-based person re-identification. *CSVT*, *28*(10), 2768–2776.
Zhang, X., Zhou, X., Lin, M., & Sun, J. (2018). Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *CVPR* (pp. 6848–6856).
Zhao, H., Tian, M., Sun, S., Shao, J., Yan, J., Yi, S., Wang, X., & Tang, X. (2017). Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In *CVPR* (pp. 1077–1085).
Zhao, L., Li, X., Zhuang, Y., & Wang, J. (2017a). Deeply-learned part-aligned representations for person re-identification. In *CVPR* (pp. 3219–3228).
Zhao, R., Ouyang, W., & Wang, X. (2013). Unsupervised salience learning for person re-identification. In *CVPR* (pp. 3586–3593).
Zhao, R., Ouyang, W., & Wang, X. (2017b). Person re-identification by saliency learning. *TPAMI*, *39*(2), 356–370.
Zhao, Y., Shen, X., Jin, Z., Lu, H., & Hua, X.-s. (2019). Attribute-driven feature disentangling and temporal aggregation for video person re-identification. In *CVPR* (pp. 4913–4922).
Zheng, L., Bie, Z., Sun, Y., Wang, J., Su, C., Wang, S., & Tian, Q. (2016). Mars: A video benchmark for large-scale person re-identification. In *ECCV* (pp. 868–884).
Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., & Tian, Q. (2015). Scalable person re-identification: A benchmark. In *ICCV* (pp. 1116–1124).
Zheng, L., Zhang, H., Sun, S., Chandraker, M., Yang, Y., & Tian, Q. (2018). Person re-identification in the wild. In *ECCV* (pp. 176–193).
Zheng, Z., Zheng, L., & Yang, Y. (2017). Unlabeled samples generated by Gan improve the person re-identification baseline in vitro. In *ICCV* (pp. 3774–3782).
Zhong, Z., Zheng, L., Cao, D., & Li, S. (2017a). Re-ranking person re-identification with k-reciprocal encoding. In *CVPR* (pp. 3652–3661).
Zhong, Z., Zheng, L., Kang, G., Li, S., & Yang, Y. (2017b). Random erasing data augmentation. In *CoRR*. arXiv:1708.04896.
Zhong, Z., Zheng, L., Li, S., & Yang, Y. (2018). Generalizing a person retrieval model hetero-and homogeneously. In *ECCV* (pp. 172–188).
Zhou, S., Wang, F., Huang, Z., & Wang, J. (2019). Discriminative feature learning with consistent attention regularization for person re-identification. In *ICCV* (pp. 3760–3769).
Zhou, T., & Tao, D. (2011). Godec: Randomized low-rank and sparse matrix decomposition in noisy case. In *ICML* (pp. 33–40).
Zhou, Z., Huang, Y., Wang, W., Wang, L., & Tan, T. (2017). See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification. In *CVPR* (pp. 6776–6785).