Multi-Scale Temporal Cues Learning for Video Person Re-Identification

Jianing Li, Shiliang Zhang, Member, IEEE, and Tiejun Huang, Senior Member, IEEE

Abstract—Temporal cues embedded in videos provide important clues for person Re-Identification (ReID). To efficiently exploit temporal cues with a compact neural network, this work proposes a novel 3D convolution layer called Multi-scale 3D (M3D) convolution layer. The M3D layer is easy to implement and could be inserted into traditional 2D convolution networks to learn multi-scale temporal cues by end-to-end training. According to its inserted location, the M3D layer has two variants, i.e., local M3D layer and global M3D layer, respectively. The local M3D layer is inserted between 2D convolution layers to learn spatial-temporal cues among adjacent 2D feature maps. The global M3D layer is computed on adjacent frame feature vectors to learn their global temporal relations. The local and global M3D layers hence learn complementary temporal cues. Their combination introduces a fraction of parameters to traditional 2D CNN, but leads to the strong multi-scale temporal feature learning capability. The learned temporal feature is fused with a spatial feature to compose the final spatial-temporal representation for video person ReID. Evaluations on four widely used video person ReID datasets, i.e., MARS, DukeMTMC-VideoReID, PRID2011, and iLIDS-VID demonstrate the substantial advantages of our method over the state-of-the-art. For example, it achieves rank1 accuracy of 88.63% on MARS without re-ranking. Our method also achieves a reasonable trade-off between ReID accuracy and model size, e.g., it saves about 40% parameters of 3D CNN.

Index Terms—Video Person ReID, Convolutional Neural Networks, Spatial Temporal Feature Learning

I. INTRODUCTION

PERSON Re-Identification (ReID) aims to identify a specific person by matching his/her images or video sequences from other cameras. Because of its importance in security related applications like smart surveillance and criminal investigation, person ReID is currently a highly active area of research.

Manuscript received April 25, 2019; revised October 13, 2019 and December 30, 2019; accepted January 28, 2020. Date of publication February 14, 2020; date of current version February 21, 2020. This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFE0118400, in part by the Beijing Natural Science Foundation under Grant JQ18012, and in part by the Natural Science Foundation under Grant 61936011, Grant 61425025, Grant 61620106009, Grant 61572050, and Grant 91538111. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Yun Fu. (Corresponding author: Shiliang Zhang.)

Jianing Li and Shiliang Zhang are with the Department of Computer Science, School of Electronic Engineering and Computer Science, Peking University, Beijing 100871, China (e-mail: ljm-vmc@pku.edu.cn; slzhang.jdl@pku.edu.cn).

Tiejun Huang is with the Department of Computer Science, School of Electronic Engineering and Computer Science, Peking University, Beijing 100871, China, and also with the Peng Cheng Laboratory, Shenzhen 518066, China (e-mail: tjhuang@pku.edu.cn).

Current research mainly focuses on two lines of tasks depending on individual images and video sequences, respectively. Recent years have witnessed the impressive progresses in image person ReID, e.g., deep visual representations have significantly boosted the ReID performance on image ReID datasets [1]–[7]. Since videos contain plenty of spatial-temporal cues, video person ReID has potential to better handle some challenges in image person ReID. Fig. 1 shows several sampled frames from video sequences in MARS dataset [8]. As shown in Fig. 1(a), those two pedestrians show similar appearance, thus are hard to be distinguished solely relying on visual cues. However, those two pedestrians can be easily distinguished by their temporal cues, e.g., the gait of pedestrian. The two pedestrians in Fig. 1(b) show similar gait cues, but can be easily distinguished by their spatial and appearance cues. It is easier to infer that, both spatial and temporal cues are important when identifying a specific person. Therefore jointly extracting and leveraging spatial and temporal cues will benefit the video person ReID.

Fig. 1. Illustration of video frames sampled from person tracklets. (a) shows two persons with similar appearance but different gait; (b) shows two persons with similar gait but totally different appearance.
temporal relations on high-level features, hence is not capable to capture temporal cues on image local details. The optical flow effectively represents temporal cues among adjacent frames, but could be sensitive to spatial misalignment and detection errors. Therefore, more effective and robust way of acquiring spatial-temporal feature still should be investigated.

As shown in Fig. 2 (d), 3D CNN is recently introduced to jointly learn the spatial-temporal representation in many video vision tasks e.g., video person ReID [18], [19] and action recognition [20], [21]. Through sliding convolution kernels on both spatial and temporal dimensions, the 3D CNN model jointly encodes visual appearance and temporal cues. Promising performances have been reported in many studies [20], [22], [23]. Because a single 3D convolution kernel can only cover a short temporal range, researchers usually stack several 3D kernels together to gain stronger temporal cue learning ability. Although showing better performance, stacked 3D convolutions result in substantial growth of parameters, e.g., the widely used C3D [22] network reaches the model size of 321MB with only 8 3D convolution layers, about 3 times to the 95.7MB parameters of ResNet50 [24]. Too many parameters not only make 3D CNNs computationally expensive, but also lead to the difficulty in model training and optimization. This makes 3D CNN not readily applicable on video person ReID, where the training set is commonly small and person ID annotation is expensive.

This work aims to explore the abundant spatial and temporal cues in videos for person ReID through applying 3D convolution, while mitigating the shortcomings in existing 3D CNN models. A novel Multi-scale 3D (M3D) convolution layer is proposed as a more efficient and compact alternative to traditional 3D CNN layers. The proposed M3D layer is implemented using several parallel temporal convolution kernels with different temporal ranges. The M3D layer has two types of variants, i.e., the local M3D and global M3D layers to capture temporal cues from different scales. The local M3D layer is inserted between 2D convolution layers to learn spatial-temporal cues among adjacent 2D feature maps. The global M3D layer is computed on adjacent frame feature vectors to learn their global temporal relations. The local and global M3D layers hence learn complementary temporal cues. Two types of M3D layers are inserted into 2D CNN to compose the M3D CNN with residual connections, which allow us to effectively initialize the M3D CNN with existing 2D CNN models. The resulting M3D CNN introduces marginal parameters to 2D CNN, but gains the multi-scale temporal cues modeling ability.

Compared with the conference version, this work makes the following differences. 1) We design two variants of M3D layers, e.g., global M3D and local M3D, respectively. The local M3D layer captures the local spatial-temporal details but could be sensitive to misalignment between adjacent frames. The global M3D layer is hence applied on adjacent frame features to learn global temporal cues, as well as to boost the feature robustness. Although the M3D layer is a general structure for temporal cues learning, its local and global variants work on different feature scales and learn complementary features. 2) The SAL attention layers in the conference version improve the robustness of the learned feature, but are expensive to compute and train. Since the global M3D layer has the ability to learn global temporal cues, we discard the SAL attention layers to chase better efficiency. 3) This efficient new design brings substantial performance gains, e.g., boosting the mAP on MARS dataset from 74.06% to 79.46%. More ablation studies and experimental comparisons are provided. A new dataset, i.e., the recently released DukeMTMC-VideoReID is also tested. As shown in our experiments, the combination of local and global M3D layers introduces a fraction of parameters, but leads to strong multi-scale temporal feature learning capability.

Based on M3D CNN, we further propose a two-stream feature fusion architecture, that consists of M3D CNN and 2D CNN to learn temporal and spatial features respectively. Extensive experiments demonstrate that our method outperforms a wide range of state-of-art ones on four widely used video person ReID datasets, i.e., MARS [8], DukeMTMC-VideoReID [25], [26], PRID2011 [27] and iLIDS-VID [28], respectively. For example, our fused feature achieves rank1 accuracy of 88.63% without re-ranking on MARS, outperforming the recent STMP [29] by 4.2%.
Moreover, we achieve a reasonable trade-off between ReID accuracy and model size. Introducing only about 15MB parameter overhead to the 2D CNN, M3D CNN boosts the mAP of 2D CNN from 62.54% to 78.64% on MARS. The 3D CNN model I3D [20] achieves mAP of 62.84% with 186MB parameters. Compared with I3D, M3D performs better and saves about 76MB of parameters, thus could be a better temporal feature learning model for video person ReID.

The contribution of this work can be summarized into following aspects. 1) We propose a novel M3D convolution layer to jointly learn multi-scale temporal cues for video person ReID. 2) A two-stream architecture consisting of M3D CNN and 2D CNN is hence proposed to learn spatial-temporal feature, which achieves competitive performance compared with the state-of-the-art in extensive experiments. 3) To the best of our best knowledge, this is also an early attempt of studying light-weight and efficient 3D CNN for video person ReID.

II. RELATED WORK

Existing studies on person ReID can be divided into two categories, e.g., image and video based person ReID, respectively. Most of image based person ReID works focus on two aspects: 1) learning discriminative image features [2], [30]–[35] and 2) learning distance metrics for feature matching [36]–[38], respectively. Li et al. [39] propose a novel separation loss to separate positive and negative support neighbors, and a squeeze loss to penalize the variance among positive samples. These two losses are complemented with each other and achieve promising performance. Li et al. [40] proposes to jointly learn dictionaries and mapping to bridge the gap across low and high resolution person images. Multi-scale feature learning is widely explored in image ReID task. MDL [41] learns representation at 3 different spatial scales, e.g., patch-level, part-level, and image-level, and fuses the multi-level spatial features for ReID. MDLA [42] exploits multi-scale spatial cues and proposes a saliency-based learning strategy. ASTPN [43] uses a spatial pyramid pooling structure to learn multi-scale spatial feature. These methods show strong capability to learn robust spatial features. Different from these methods, our method aims to exploit multi-scale cues in temporal dimension for video ReID.

Early works regard video person ReID as an extension of image person ReID and design hand-crafted spatial-temporal features. For instance, 3D-SIFT [44] and HOG3D [45] are adopted as spatial-temporal features for video person ReID. Hand-crafted features present limited robustness and discriminative power compared with deep features [8]. Recent deep feature learning works can be summarized into four categories, as shown in Fig. 2. The following parts briefly review related works in each category.

Temporal pooling directly aggregates frame features across all temporal stamps. Many works apply image ReID methods for frame feature extraction and then utilize temporal pooling or weight learning strategy for feature fusion [8], [10], [46]. Temporal pooling could effectively utilize algorithms designed for image person ReID and shows promising efficiency. However, it is not effective in capturing the temporal cues embedded in adjacent frames.

Optical flow effectively encodes motion cues across adjacent frames. Many works apply two-stream network to learn the spatial and temporal features from still images and stacked optical flows, respectively [47], [48]. McLaughlin et al. [14] combine optical flow and RNN to model short and long term temporal cues. However, optical flow is sensitive to the spatial misalignment and detection errors across adjacent bounding boxes, which commonly exist in video ReID datasets. This leads to inaccurate temporal feature. And the computation of optical flow is also time-consuming.

Recurrent Neural Network (RNN) is also introduced for temporal feature learning in video ReID tasks. McLaughlin et al. [14] and Liu et al. [29] apply recurrent network to model temporal cues cross frame features and further utilize temporal pooling to generate the final feature. RNN only builds temporal connections on high-level features and ignores the low-level temporal cues.

3D CNNs treat the video as a 3D signal and jointly model the spatial-temporal cues through sliding 3D convolution kernels on both spatial and temporal dimensions. Carreira and Zisserman [20] build a deep Inflated 3D (I3D) network through inflating 2D convolution kernels to corresponding 3D version. Qiu et al. [21] factorize the 3D convolutional filters into spatial and temporal components, which reduces a large number of parameters without degrading the performance. Fig. 3 illustrates the detailed architecture of 2D CNN, I3D [20] and P3D [21] convolution kernels. Although these 3D CNNs have exhibited promising performances, existing works need to stack a certain number of 3D convolution kernels to capture long-range temporal cues, resulting in large parameter overheads and increased difficulty for CNN optimization.
Moreover, existing 3D CNN layers are also sensitive to the spatial misalignment.

Multi-scale 3D CNN is investigated by several works in medical image processing. Our M3D differs with them in the way to achieve the multi-scale feature learning. Existing methods generally feed multi-scale inputs into different network streams to achieve multi-scale feature learning. Kamnitsas et al. [49] propose a two-stream 3D CNN with different convolution kernel sizes and different input sizes. Li et al. [50] introduce a multi-branch network with different input scales. Ghafoorian et al. [51] also propose a multi-branch network and crop image patches with different sizes to learn the multi-scale cues. We name our method as M3D because it uses dilated temporal convolution to achieve multi-scale feature learning in a compact single-stream 3D CNN, hence could be more efficient than previous designs.

Compared with existing 3D CNN layers, M3D layer presents better temporal feature learning ability with a more compact architecture. Different from previous works using stacked optical flow as input, our method directly extracts temporal feature from original video sequences, hence would be more efficient to compute and more robust to misalignment errors. Extensive experimental results on four datasets demonstrate the advantage of the proposed method over existing ones.

III. Two-Stream M3D Convolution Network

A. Problem Formulation

Person ReID aims to identify a specific person from a large scale database, which can be implemented as a retrieval task. Given a query video sequence \( S_q = (s^1, s^2, \ldots s^T) \), where \( T \) is the sequence length and \( s^t \) is the \( t \)-th frame at time \( t \). Video person ReID can be tackled by ranking gallery sequences \( G \) based on the video representation \( f \), and a distance metric \( D \) computed between \( S_q \) and each gallery sequence \( S_g \). In the returned rank list, sequences containing the identical person with query \( S_q \) are expected to appear on top of the list. Therefore, learning discriminative video representation \( f \) and designing the distance metric \( D \) are two critical steps for video person ReID.

This work focuses on designing a discriminative video representation \( f \). As illustrated in Fig. 1, both the spatial and temporal cues embedded in video sequences are important for identifying a specific person. Because the spatial and temporal cues are complementary with each other, we extract them with two modules. The video representation \( f \) can be formulated as

\[
f = [f_s, f_t],
\]

where \( f_s \) and \( f_t \) denote the spatial and temporal representations respectively, and \([,]\) denotes feature fusion, which could be implemented with different strategies including weighted fusion and adaptive fusion [52]. This paper utilizes vector concatenation for its simplicity and high efficiency.

Existing image based person ReID works have proposed many successful 2D CNNs for spatial feature extraction. We refer to existing works and utilize 2D CNN to extract the sequence spatial feature \( f_s \). Specifically, this is finished by first extracting spatial representation from each individual video frame, then aggregating frame features through temporal average pooling, i.e.,

\[
f_s = \frac{1}{T} \sum_{t=1}^{T} F_{2d}(s^t),
\]

where \( F_{2d} \) refers to 2D CNN, which is used to extract the frame feature. As a commonly used pooling strategy, average pooling has no parameter to tune and works for arbitrary length of sequences.

As discussed in the above sections, more effective ways of acquiring temporal feature should be investigated. For the temporal representation \( f_t \), we propose a Multi-scale 3D (M3D) CNN to extract the temporal feature from the video sequence, i.e.,

\[
f_t = F_{M3D}(S),
\]

where \( F_{M3D} \) denotes the M3D CNN. It takes video sequence as input and outputs temporal feature \( f_t \).

The 2D CNN and M3D CNN compose a two-stream feature fusion architecture illustrated in Fig. 4. The following sections describe our design of the M3D CNN, which is constructed by inserting several local M3D layers and one global M3D layer into a 2D CNN.

B. Multi-Scale 3D Convolution

As illustrated in Fig. 4, the global and local M3D layers are inserted into a 2D CNN to enable it to process video sequences and learn extra temporal cues. Before introducing M3D layer, we first briefly review 3D convolution and dilated convolution, which inspire the design of M3D layer.

1) 3D Convolution: A video clip can be represented as a 4D tensor with the size of \( C \times T \times H \times W \), where \( C \), \( T \), \( H \), and \( W \) denote the number of input channels, temporal length, height and width of each frame, respectively. A 3D convolution kernel can be formulated as a 3D tensor with size of \( t \times h \times w \) (the channel dimension is omitted for the brief description), where \( t \) is the temporal depth of the kernel, while \( h \) and \( w \) are the spatial sizes. 3D convolution encodes the spatial-temporal cues through sliding along both spatial and temporal dimensions of the video clip.

3D convolution kernel only captures the short-term temporal cues, e.g., the 3D kernels in Fig. 3 (b-e) capture the temporal relations across 3 frames. To model longer-term temporal cues, multiple 3D convolution kernels have to be concatenated as a deep network. A deep 3D CNN involves a large amount of parameters. Moreover, 3D CNNs can not leverage the 2D images in ImageNet [53] for model pre-training, making it hard to be initialized and optimized. Further more, the small spatial receptive field makes existing 3D convolution kernels sensitive to misalignment errors between adjacent frames.

2) Dilated Convolution: The spatial dilated convolution has been widely used in image segmentation for its strong and efficient multi-scale spatial context modeling capability [54]. Fig. 5 illustrates traditional convolution and dilated convolution. Considering input signals \( x \in \mathcal{R}^{H \times W} \), for each location \([h, w]\) on the output \( y \) and a filter \( W \in \mathcal{R} \), the spatial
Fig. 4. Illustration of our two-stream feature fusion architecture for spatial and temporal features extraction. The temporal stream, i.e., M3D CNN, is constructed by inserting several local M3D layers and one Global M3D layer into a 2D CNN.

3) Local M3D Layer: Shortcomings of 3D CNNs motivate us to design a compact convolution kernel that captures longer-term temporal cues. Inspired by dilated convolution [54], we propose to capture temporal cues through parallel dilated convolutions on the temporal dimension.

A local M3D layer contains a spatial convolution kernel and n parallel temporal kernels with different temporal ranges. Given an input feature map \( x \in \mathbb{R}^{C \times T \times H \times W} \), we define the output of local M3D layer as:

\[
y = S(x) + \sum_{i=1}^{n} T^{(i)}(S(x)),
\]

where \( S \) is a standard spatial convolution and \( T^{(i)} \) is the \( i \)-th temporal convolution with dilation rate \( r^{(i)} \). The dilation rate \( r^{(i)} = 2^{i-1} \) exponentially increases with \( i \) to get larger temporal receptive fields. The computation of \( S \) follows the ones in 2D convolution. The computation of temporal kernel \( T^{(i)} \) is similar to the spatial one in Eq. (4), i.e.,

\[
y^{(i)} = T^{(i)}(x),
\]

\[
y^{(i)}[t, h, w] = \sum_{a=-1}^{1} x[t + a \times r^{(i)}, h, w] \times W^{(i)}[a],
\]

where \( W^{(i)} \) denotes the \( i \)-th temporal kernel weight.

Fig. 6 illustrates the detailed structure of local M3D layer with \( n = 3 \). As shown in Fig. 6, \( n \) controls the maximum receptive field size in the temporal dimension. If we
Fig. 7 illustrates the detailed structure of global M3D layer with \( n = 3 \). It can be observed that, global M3D layer has similar structure with local M3D layer. Differently, global M3D layer is computed on the global feature vectors, hence learns the temporal cues on a larger spatial scale. This makes the global M3D layer more robust to misalignment errors and improves the feature robustness.

### C. Video Feature Generation

The final video feature is the concatenation of spatial feature and temporal feature, as shown in Eq. (1). The spatial feature is generated by pooling 2D CNN features with Eq. (2). The temporal feature is generated based on the output \( \tilde{F} \in R^{(n+1)d \times T \times 1 \times 1} \) of global M3D layer. Temporal average pooling is applied to generate a fixed length temporal feature, i.e.,

\[
  f_t = \text{avgpool}(\tilde{F}) = \frac{1}{T} \sum_{i=1}^{T} \tilde{F}[r, 1, 1]. \tag{10}
\]

With the designed local M3D and global M3D layers, our M3D CNN can jointly learn the multi-scale temporal feature \( f_t \) with a compact architecture. The spatial and temporal features are concatenated as the final video feature \( f \) for person ReID. Fig. 4 illustrates the framework of proposed two-stream feature fusion network, as well as the extraction of spatial and temporal features. The following section tests the validity of the proposed algorithms.

### IV. Experiment

#### A. Dataset

We use four video person ReID datasets to conduct experiments, including PRID-2011 [27], iLIDS-VID [28], MARS [8] and DukeMTMC-VideoReID [25], [26], respectively.

PRID-2011 [27] consists of 400 sequences of 200 pedestrians from two cameras. Each sequence has a length between 5 and 675 frames. Following the implementation in previous works [10], [28], we randomly split this dataset into train/test identities. This procedure is repeated for 10 times to report the accuracy.

iLIDS-VID [28] consists of 600 sequences of 300 pedestrians from two non-overlapping cameras. Each sequence has a variable length between 23 and 192 frames. We also follow the implementation in [10], [28] to randomly split this dataset into train/test identities for 10 times.

MARS [8] consists of 1261 pedestrians and 20,715 sequences under 6 cameras. Each pedestrian is captured by at least 2 cameras. This dataset provides pre-defined training and testing sets, which contain 630 and 631 identities, respectively.

DukeMTMC-VideoReID [25], [26] is a manually annotated dataset. It contains 702 identities for training, 702 identities for testing, and 408 identities as distractors. The training set contains 369,656 frames of 2,196 sequences, and the test set contains 445,764 frames of 2,636 sequences. The evaluation protocol of this dataset is also provided.
TABLE I

| Dataset          | Method | MARS mAP | rank1 | DukeMTMC mAP | rank1 |
|------------------|--------|----------|-------|--------------|-------|
| ResNet50         |        | 62.54    | 76.43 | 79.08        | 84.47 |
| Local M3D (n=1)  |        | 67.35    | 78.63 | 84.04        | 88.32 |
| Local M3D (n=2)  |        | 69.96    | 81.41 | 87.20        | 91.03 |
| Local M3D (n=3)  |        | 72.84    | 82.98 | 90.04        | 93.16 |
| Local M3D (n=4)  |        | 72.04    | 82.42 | 89.68        | 93.30 |
| Local M3D (n=5)  |        | 72.01    | 82.22 | 89.87        | 92.59 |

C. Ablation Study

1) Temporal Branches in Local M3D Layer: This section tests the influence of temporal branches in local M3D layer. The global M3D layer is not considered. We directly use average pooling to generate the final temporal feature vector. The max branch number is set as $n = 5$. The experimental results achieved by temporal feature on the MARS and DukeMTMC-VideoReID datasets are summarized in Table I.

As shown in Table I, with $n \geq 1$, the local M3D captures extra temporal cues among adjacent frames and outperforms ResNet50 baseline by large margins, e.g., with $n = 1$, the local M3D boosts the mAP of ResNet50 baseline from 62.54% to 67.35%. Note that, the local M3D layer is equal to P3D-C layer with $n = 1$. We only insert 4 local M3D layers into the ResNet50 backbone, while P3D-C CNN inserts 16 P3D-C layers. Therefore, with $n = 1$ M3D CNN gets lower performance than P3D-C CNN in Sec. IV-D. With the increase of $n$, local M3D layer brings further performance gains. With $n \geq 3$, the performance boost slows down. This could be because 3 branches correspond to long temporal ranges, which are enough for multi-scale temporal feature learning. We hence set $n = 3$ for local M3D layer in the following experiments as a reasonable tradeoff between accuracy and efficiency.

2) Temporal Branches in Global M3D Layer: This section evaluates the influence of temporal branches in global M3D layer. To clearly verify the effectiveness of global M3D layer, local M3D layer is not considered in this section. The max branch number is also set to 5. Experiment results on MARS and DukeMTMC-VideoReID datasets are summarized in Table II.

It can be observed from Table II that, with $n \geq 1$, the global M3D layer captures temporal cues among adjacent frame features, and substantially outperforms the ResNet50 baseline. This comparison also shows that learning temporal feature with global M3D layer outperforms the simple feature average pooling strategy. It is also clear that, the increase of temporal branches $n$ brings consistent performance gains on two datasets. Therefore, introducing multiple branches in global M3D layer is beneficial to temporal feature learning. Similar to the observation for local M3D layer, $n \geq 3$ does not bring substantial performance gains. We hence also fix $n = 3$ for global M3D layer in the following experiments.

3) Evaluation of Individual Components of M3D

We further verify the effectiveness of each component in the proposed two-stream feature fusion architecture, including local M3D layer, global M3D layer, and the fused spatial-temporal feature. Experimental results are shown in Table III. In the table, “ResNet50” denotes the 2D ResNet50 baseline, which doesn’t
consider any temporal cues and aggregate the frame features through average pooling. “Local M3D” denotes replacing four 2D convolution layers in ResNet50 with local M3D layers and applying average pooling for final feature generation. “Global M3D” denotes imposing a global M3D layer on frame feature vectors learned by ResNet50. “M3D” denotes the M3D CNN including local and global M3D layers, as shown in Fig. 4. “Two-stream” denotes the fused spatial-temporal feature from M3D CNN and ResNet50.

Table III shows that, the ResNet50, which doesn’t consider any temporal cues, gets the worst performance. After inserting 4 local M3D layers into ResNet50, the Local M3D gets a substantial performance boost, e.g., 10.30% mAP on MARS dataset over with the ResNet50. Similarly, after imposing the global M3D layer, the Global M3D also brings a large performance promotion. It is interesting to observe that, Global M3D outperforms Local M3D. This may be because global M3D layer is computed on global feature vectors, hence is more robust to spatial misalignment errors which commonly exist in video ReID datasets. Meanwhile, the global M3D layer also contains smaller number of parameters, hence is easier to be optimized. By considering both global and local M3D layers, the M3D CNN gets further performance promotions. This implies that, global and local M3D layers can learn complementary temporal cues. Moreover, when combined with spatial feature from 2D ResNet50, the two-stream architecture achieves the best performance.

We could conclude from the above experiments that, each component in our method works well for discriminative video feature learning. By considering all of those components, the two-stream feature fusion model exhibits the best performance. We visualize the saliency maps from M3D and 2D CNN models in Fig. 8. It can be observed from Fig. 8(b) that, the 2D CNN tends to focus on specific area, e.g., the upper body. This is reasonable because the upper body is discriminative for differentiating persons. Fig. 8(c) also shows that, M3D CNN tends to focus on areas with motions, e.g., the moving legs and arms. This implies that, the M3D CNN is more effective to capture the temporal cues in video sequences. The illustration also shows that, the 2D CNN and M3D CNN learn complementary cues, hence they have potential to achieve better performance when combined. As shown in Table III, their combination gets the best performance.

D. Comparison With Other Temporal Feature Learning Algorithms

To further verify the effectiveness of the proposed M3D layers, we build a M3D CNN and compare it against several widely used temporal feature extraction methods, including:

I3D [20] inflates 2D kernels into 3D version to acquire the temporal cues learning ability. Fig. 3 (a-b) show the inflation process. 2D kernels are typically square, therefore they are inflated cubically, e.g., $N \times N$ to $N \times N \times N$, which introduces a large amount of parameters to I3D.

P3D [21] factorizes the 3D kernels into separate spatial and temporal ones, e.g., factorizes a $N \times N \times N$ kernel to a $1 \times N \times N$ spatial kernel and a $N \times 1 \times 1$ temporal kernel to reduce the number of parameters. Fig. 3(c-e) shows 3 ways of factorizations, which are named as P3D-A, P3D-B and P3D-C, respectively. The factorization substantially decreases the parameter number in 3D CNNs. Real implementations still need to stack many P3D kernels to capture long-term temporal cues.

LSTM [16] is widely used in video recognition tasks to model the long-short term temporal relationship. Usually the recurrent layer learns temporal relation on image-level features, and temporal pooling is further imposed to aggregate the frame-level feature and yield a sequence-level feature vector.

**Appearance and Optical flow (A&O)** Two-stream [47] is also widely used to learn spatial and temporal features. The optical flow stream directly models the short-term temporal cues, while the appearance stream models the spatial appearance. The two streams are final combined to aggregate the spatial and temporal features.

We apply ResNet50 as the 2D CNN baseline. All of the models mentioned above are implemented based on ResNet50. 3D CNNs are implemented by replacing 2D convolution layers with corresponding 3D versions. The LSTM is implemented by imposing LSTM layer on frame features. For the A&O Two-stream [47] model, we use two ResNet50 networks to extract features from the original frame and optical flow, respectively. The comparison is summarized in Table IV.

As shown in Table IV, I3D achieves similar performance with 2D CNN in Table IV. The reason might be because I3D model has too many parameters, which makes it hard to be optimized on the relatively small video person ReID training sets. The P3D-A shows poor performance compared with 2D CNN. This could be caused by the serial connection between spatial kernel and temporal kernel, which increases the non-linearity of the CNN model and leads to the difficulty of optimization on small training set. The P3D-B and P3D-C connect the spatial and temporal kernels through parallel or residual connections which are relatively easy to optimize. They get substantial performance improvement over the 2D CNN. This shows that consider temporal information is helpful for video recognition tasks.
person ReID task. The LSTM based method, also gets poor performance compared with 2D CNN. The reason may be because the LSTM contains several fully connected layers which are hard to optimize on video ReID datasets. The A&O two-stream gets slight performance promotion over 2D CNN. We could infer that, the optical flow is helpful for person ReID, but could be sensitive to noises.

The experimental results also show that, our M3D CNN constantly outperforms 2D CNNs and other 3D CNNs, e.g., the M3D CNN outperforms 2D CNN and P3D-C by large margins, i.e., 16.10% and 9.58% in mAP, respectively. Meanwhile, the M3D CNN is also more compact than the compared 3D CNNs. As shown in Table IV, among compared 3D CNN methods, the M3D model contains the least number of parameters, meanwhile achieves the best performance and highest speed, i.e., 766 frames/s, which is quite close the 781 frames/s speed by 2D CNN baseline. M3D CNN also saves about 40% parameters of 13D CNN. We further embed the SAL attention model [19] of the conference version into M3D CNN. The SAL increases the number of parameters and reduces the speed, but doesn’t bring obvious promotion on ReID performance. The reason may be because the global M3D and local M3D layers have learned the robust spatial-temporal feature, and the SAL can’t work well on such strong baseline. Therefore, we discard SAL in journal version to chase better efficiency.

We further replace all the 2D convolution layers in ResNet50 with local M3D layers, which introduces large parameter overheads but doesn’t bring further performance improvement. This implies that a small number of local M3D layers already captures discriminative temporal cues in video sequences. We hence could conclude that, M3D layers present promising capability in learning multi-scale temporal cues.

E. Comparison With Recent Work

1) Comparison on PRID and iLIDS-VID: We first compare our approach with recent ones on PRID and iLIDS-VID datasets, and summarize the comparison in Table V. As shown in the table, our proposed method presents competitive performance on rank1 accuracy. DRSA [10] and AFDTA [56] also achieve competitive performance on both datasets. The reason may be because PRID has a relatively small training set. DRSA and AFDTA alleviates the insufficiency of training data through multi-task learning, e.g., DRSA imposes body part cues and Online Instance Matching loss (OIM) for training, and AFDTA introduces extra attribute datasets for pre-train. While our method uses standard softmax loss and extracts global feature from whole frames without extra annotations. With more simple design, our method still outperforms DRSA and AFDTA on both PRID and iLIDS-VID. We also compare with recent temporal feature learning methods, e.g., RFA-Net [16], DRCN [57], RCN [14], SeeForest [58], T-CN [59], CSA [60] and STMP [29]. Our method outperforms these works in rank1 and rank5 accuracies. This competitive performance demonstrates the superiority of M3D in spatial-temporal feature learning for video person ReID.

2) Comparison on MARS: Table VI reports the comparison of our approach with recent works on MARS. It can be observed from Table VI that, our method outperforms most of the compared methods. Our method achieves the rank1
accuracy of 88.63%, outperforming all of compared methods, including the recent works AFDTA [56] and VRSTC [66].

Note that, AFDTA [56] leverages extra attribute recognition dataset for pre-training. VRSTC [66] introduces extra global and part losses for training and adopts temporal attention layer to refine the video feature. VRSTC hence requires higher computational complexity and multiple losses for optimization. Compared with those works, our method is more concise and efficient, e.g., we directly extract global feature with M3D model and match features with simple Euclidean distance. We only use softmax loss and don’t introduce any attention mechanism. When we combine two-stream M3D with re-ranking strategy [73], we get further performance boost and achieve 85.46% mAP.

3) Comparison on DukeMTMC-VideoReID: The comparison on DukeMTMC-VideoReID dataset is shown in Table VII. Because DukeMTMC-VideoReID is a newly proposed video person ReID dataset, a limited number of works have reported performance on it. The reported performance of ETAP-Net [25] in Table VII is achieved with a supervised baseline. Besides the ETAP-Net [25], we add ResNet50 [24] and VRSTC [66] for comparison. As shown in Table VII, based on ResNet50 backbone, the two-stream M3D model achieves 93.67% mAP and 95.49% rank1 accuracy, outperforms the supervised ETAP-Net [25] by large margins. It also outperforms the recent method VRSTC [66]. Compared with VRSTC [66], our method only uses softmax loss and don’t introduce extra attention mechanism. We hence could conclude that our approach achieves competitive performance with efficient model design.

We show some person ReID results achieved by our approach and ResNet50 baseline on MARS [8] and DukeMTMC-VideoReID [25] in Fig. 9. For each query, we show the top-5 returned video sequences by those two methods. It can be observed that, our approach is more discriminative for identifying persons.

F. Discussions

The proposed M3D layer does share certain similarity with C3D, P3D, and I3D, because it also uses 3D convolution to
process video sequences. C3D, P3D, and I3D are initially designed for action recognition. Differently, M3D layer has some designs that make it more suitable for person ReID than video action recognition. 1) our global M3D layer is designed to relieve the misalignment issues, which are common in ReID datasets because of human bounding box detection errors. 2) As shown in Fig. 1, temporal and appearance cues are equally important to identify a person. However, the motion cue is more important for action recognition, e.g., different persons with distinct clothes can perform the same action. Our two-stream M3D CNN hence combines both temporal and appearance features for ReID. 3) As shown in Fig. 1, different frames in person tracklets many exhibit different poses and viewpoints. Those cues provide important clues to boost the robustness for viewpoint and pose variances. Our model hence inserts several embedding layers and one global M3D layer to aggregate the meaningful cues learned in different video frames. Most of action recognition frameworks do not consider such feature aggregation design and directly perform action recognition after the last convolution layer. Comparisons with C3D, P3D, and I3D on person ReID also show the promising performance of our M3D. For instance, M3D achieves rank1 accuracy of 88.63% on MARS significantly outperforming 76.62%, 52.17%, and 79.86% of I3D, C3D, and P3D, respectively.

V. CONCLUSION

This paper proposes a novel two-stream convolution network to explicitly leverage spatial and temporal cues for video person ReID. Two types of M3D layers are proposed to learn multi-scale temporal cues in video sequences. The local M3D layer is implemented by parallel dilated temporal convolutions to learn spatial-temporal cues among adjacent 2D feature maps. The global M3D layer is computed on adjacent frame feature vectors to learn their global temporal relations. The local and global M3D layers are complementary to each other. Their combination boosts the temporal feature learning capability in the resulting M3D CNN. The temporal feature learned by M3D CNN is fused with a spatial feature for video person ReID. Experimental results on four widely used video person ReID datasets demonstrate the superiority of our approach over the recent ones.

REFERENCES

[1] H. Yao, S. Zhang, R. Hong, Y. Zhang, C. Xu, and Q. Tian, “Deep representation learning with part loss for person re-identification,” IEEE Trans. Image Process., vol. 28, no. 6, pp. 2860–2871, Jun. 2019.
[2] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian, “Pose-driven deep convolutional model for person re-identification,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 3960–3969.
[3] Y. Yan, B. Ni, J. Liu, and X. Yang, “Multi-level attention model for person re-identification,” Pattern Recognit. Lett., vol. 127, pp. 156–164, Nov. 2019.
[4] J. Li, S. Zhang, Q. Tian, M. Wang, and W. Gao, “Pose-guided representation learning for person re-identification,” IEEE Trans. Pattern Anal. Mach. Intell., to be published.
[5] J. Lin, L. Ren, J. Lu, J. Feng, and J. Zhou, “Consistent-aware deep learning for person re-identification in a camera network,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 5771–5780.
[6] X. Liu, S. Zhang, X. Wang, R. Hong, and Q. Tian, “Group-group loss-based global-regional feature learning for vehicle re-identification,” IEEE Trans. Image Process., vol. 29, pp. 2638–2652, 2020.
[7] S. Mao, S. Zhang, and M. Yang, “Resolution-invariant person re-identification,” 2019, arXiv:1906.09748. [Online]. Available: http://arxiv.org/abs/1906.09748
[8] L. Zheng et al., “Mars: A video benchmark for large-scale person re-identification,” in Proc. ECCV, 2016, pp. 868–884.
[9] G. Chen, J. Lu, M. Yang, and J. Zhou, “Spatial-temporal attention-aware learning for video-based person re-identification,” IEEE Trans. Image Process., vol. 28, no. 9, pp. 4192–4205, Sep. 2019.
[10] S. Li, S. Bak, P. Carr, and X. Wang, “Diversity regularized spatiotemporal attention for video-based person re-identification,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 369–378.
[11] Y. Rao, J. Lu, and J. Zhou, “Learning discriminative aggregation network for video-based face recognition and person re-identification,” Int. J. Comput. Vis., vol. 127, pp. 701–718, Jun. 2019.
[12] G. Chen, J. Lu, J. Feng, and J. Zhou, “Localized multi-kernel discriminative canonical correlation analysis for video-based person re-identification,” in Proc. IEEE Int. Conf. Image Process. (ICIP), Sep. 2017, pp. 111–115.
[13] J. Li, J. Wang, Q. Tian, W. Gao, and S. Zhang, “Global-local temporal representations for video person re-identification,” in Proc. ICCV, 2019, pp. 3958–3967.
[14] N. Mclaughlin, J. M. D. Rincon, and P. Miller, “Recurrent convolutional network for video-based person re-identification,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 1325–1334.
[15] D. Chung, K. Tabboub, and E. J. Delp, “A two stream siamese convolutional neural network for person re-identification,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 1983–1991.
[16] Y. Yan, B. Ni, Z. Song, C. Ma, Y. Yan, and X. Yang, “Person re-identification via recurrent feature aggregation,” in Proc. ECCV, 2016, pp. 701–716.
[17] Z. Song, B. Ni, Y. Yan, Z. Ren, Y. Xu, and X. Yang, “Deep cross-modality alignment for multi-shot person re-identification,” in Proc. ACM Multimedia Conf. (MM), 2017, pp. 645–653.
[18] J. Liu, Z.-J. Xua, X. Chen, Z. Wang, and Y. Zhang, “Dense 3D-convolutional neural network for person re-identification in videos,” ACM Trans. Multimedia Comput. Commun. Appl., vol. 15, no. 1s, pp. 1–19, Jan. 2019.
[19] J. Li, S. Zhang, and T. Huang, “Multi-scale 3D convolution network for video based person re-identification,” AAAI, vol. 33, pp. 8618–8625, Aug. 2019.
[20] J. Carreira and A. Zisserman, “Quo vadis, action recognition? A new model and the kinetics dataset,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 6299–6308.
[21] Z. Qiu, T. Yao, and T. Mei, “Learning spatio-temporal representation with pseudo-3D residual networks,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 5533–5541.
[22] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, “Learning spatiotemporal features with 3D convolutional networks,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 4489–4497.
[23] S. Ji, W. Xu, M. Yang, and K. Yu, “3D convolutional neural networks for human action recognition,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 1, pp. 221–231, Jan. 2013.
[24] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 770–778.
[25] Y. Wu, Y. Lin, X. Dong, Y. Yan, W. Ouyang, and Y. Yang, “Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2018, pp. 5177–5186.
[26] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, “Performance measures and a data set for multi-target, multi-camera tracking,” in Proc. Eur. Conf. Comput. Vis. Workshops, 2016, pp. 17–35.
[27] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof, “Person re-identification by descriptive and discriminative classification,” in Proc. Scand. Conf. Image Anal. Berlin, Germany: Springer, 2011, pp. 91–102.
[28] X. Wang and R. Zhao, “Person re-identification: System design and evaluation overview,” in Person Re-Identification, London, U.K.: Springer, 2014, pp. 351–370.
[29] Y. Liu, Z. Yuan, W. Zhou, and H. Li, “Spatial and temporal mutual promotion for video-based person re-identification,” 2018, arXiv:1812.10305. [Online]. Available: http://arxiv.org/abs/1812.10305
[30] X. Liu, S. Zhang, Q. Huang, and W. Gao, “TRAM: A region-aware deep model for vehicle re-identification,” in Proc. IEEE Int. Conf. Multimedia Expo (ICME), Jul. 2018, pp. 1–6.
Shiliang Zhang (Member, IEEE) received the Ph.D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences, in 2012. He was a Postdoctoral Scientist with NEC Laboratories America and a Postdoctoral Research Fellow with The University of Texas at San Antonio. He is currently a tenure-track Assistant Professor with the Department of Computer Science, School of Electronic Engineering and Computer Science, Peking University. He has authored or coauthored over 70 articles in journals and conferences, including the IEEE TRANSACTIONS ON PAMI, the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON CSVT, ACM Multimedia, ICCV, CVPR, and ECCV. His research interests include large-scale image retrieval and computer vision. He was a recipient of many important awards like the Outstanding Doctoral Dissertation Awards from the Chinese Academy of Sciences and the Chinese Computer Federation, the President Scholarship from the Chinese Academy of Sciences, the NEC Laboratories America Spot Recognition Award, and the Microsoft Research Fellowship. He was also a recipient of the Top 10% Paper Award at the IEEE MMSP 2011.

Tiejun Huang (Senior Member, IEEE) received the bachelor’s and master’s degrees in computer science from the Wuhan University of Technology, Wuhan, China, in 1992 and 1995, respectively, and the Ph.D. degree in pattern recognition and intelligent system from the Huazhong (Central China) University of Science and Technology, Wuhan, in 1998. He is currently a Professor with the School of Electronic Engineering and Computer Science, Peking University, Beijing, China, where he is also the Director of the Institute for Digital Media Technology. He has authored or coauthored over 100 peer-reviewed articles and three books. His research areas include video coding, image understanding, digital right management, and digital library. He is also a member of the Board of Directors for Digital Media Project, the Advisory Board of the IEEE Computing Society, and the Board of the Chinese Institute of Electronics.