Global-Local Temporal Representations For Video Person Re-Identification

Jianing Li[1], Jingdong Wang[1], Qi Tian[2], Wen Gao[1], Shiliang Zhang[1]

[1]School of Electronics Engineering and Computer Science, Peking University

[2]Huawei Noahs Ark Lab

Abstract

This paper proposes the Global-Local Temporal Representation (GLTR) to exploit the multi-scale temporal cues in video sequences for video person Re-Identification (ReID). GLTR is constructed by first modeling the short-term temporal cues among adjacent frames, then capturing the long-term relations among inconsecutive frames. Specifically, the short-term temporal cues are modeled by parallel dilated convolutions with different temporal dilation rates to represent the motion and appearance of pedestrian. The long-term relations are captured by a temporal self-attention model to alleviate the occlusions and noises in video sequences. The short and long-term temporal cues are aggregated as the final GLTR by a simple single-stream CNN. GLTR shows substantial superiority to existing features learned with body part cues or metric learning on four widely-used video ReID datasets. For instance, it achieves Rank-1 Accuracy of 87.02% on MARS dataset without re-ranking, better than current state-of-the art.

1. Introduction

Person Re-Identification aims to identify a probe person in a camera network by matching his/her images or video sequences and has many real applications, including smart surveillance and criminal investigation. Image person ReID has achieved significant progresses in terms of both solutions [38, 20, 24] and large benchmark dataset construction [23, 57, 44]. Recently, video person ReID, the interest of this paper, has been attracting a lot of attention because the availability of video data is easier than before, and video data provides richer information than image data. Being able to explore plenty of spatial and temporal cues, video person ReID has potentials to address some challenges in image person ReID, *e.g.*, distinguishing different persons wearing visually similar clothes.

The key focus of existing studies for video person ReID lies on the exploitation of temporal cues. Existing works can be divided into three categories according to their ways of temporal feature learning: (i) extracting dynamic fea-
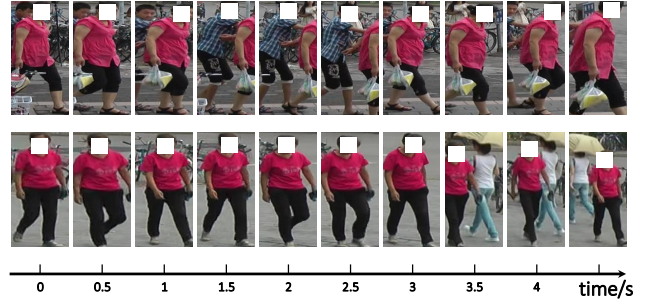

Figure 1. Illustrations of two video sequences from two different pedestrians with similar appearance on MARS dataset (we cover the face for privacy purpose). Local temporal cues among adjacent frames, *e.g.*, motion pattern or speed helps to differentiate those two pedestrians. The global contextual cues among adjacent frames can be applied to spot occlusions and noises, *e.g.*, occluded frames show smaller similarity to other frames.

tures from additional CNN inputs, *e.g.*, through optical flow [30, 5]; (ii) extracting spatial-temporal features by regarding videos as 3-dimensional data, *e.g.*, through 3D CNN [27, 19]. (iii) learning robust person representations by temporally aggregating frame-level features, *e.g.*, through Recurrent Neural Networks (RNN) [50, 30, 5], and temporal pooling or weight learning [26, 59, 22];

The third category, which our work belongs to, is currently dominant in video person ReID. The third category exhibits two advantages: (i) person representation techniques developed for image ReID can be easily explored compared to the first category; (ii) it avoids the estimation of optical flows, which is still not reliable enough due to misalignment errors between adjacent frames. Current studies have significantly boosted the performance on existing datasets, however they still show certain limitations in the aspects of either efficiency or the capability of temporal cues modeling. For instance, RNN model is complicated to train for long sequence videos. Feature temporal pooling could not model the order of video frames, which also conveys critical temporal cues. It is appealing to explore more efficient and effective way of acquiring spatial-temporal feature through end-to-end CNN learning.

This work targets to learn a discriminative Global-Local

Temporal Representation (GLTR) from a sequence of frame features by embedding both short and long-term temporal cues. As shown in Fig. 1, the short-term temporal cue among adjacent frames helps to distinguish visually similar pedestrians. The long-term temporal cue helps to alleviate the occlusions and noises in video sequences. Dilated spatial pyramid convolution [4, 51] is commonly used in image segmentation tasks to exploit the spatial contexts. Inspired by its strong and efficient spatial context modeling capability, this work generalizes the dilated spatial pyramid convolution to Dilated Temporal Pyramid (DTP) convolution for local temporal context learning. To capture the global temporal cues, a Temporal Self-Attention (TSA) model is introduced to exploit the contextual relations among inconsecutive frames. DTP and TSA are applied on frame-level features to learn the GLTR through end-to-end CNN training. As shown in our experiments and visualizations, GLTR presents strong discriminative power and robustness.

We test our approach on a newly proposed Large-Scale Video dataset for person ReID (LS-VID) and four widely used video ReID datasets, including PRID [14], iLIDS-VID [43], MARS [56], and DukeMTMC-VideoReID [47, 34], respectively. Experimental results show that GLTR achieves consistent performance superiority on those datasets. It achieves Rank-1 Accuracy of 87.02% on MARS without re-ranking, 2% better than the recent PBR [39] that uses extra body part cues for video feature learning. It achieves Rank-1 Accuracy of 94.48% on PRID and 96.29% on DukeMTMC-VideoReID, which also beat the ones achieved by current state-of-the art.

GLTR representation is extracted by simple DTP and TSA models posted on a sequence of frame features. Although simple and efficient to compute, this solution outperforms many recent works that use complicated designs like body part detection and multi-stream CNNs. To our best knowledge, this is an early effort that jointly leverages dilated convolution and self-attention for multi-scale temporal feature learning in video person ReID.

2. Related Work

Existing person ReID works can be summarized into image based ReID [43, 38, 31, 49, 55] and video based ReID [56, 35, 39, 19], respectively. This part briefly reviews four categories of temporal feature learning in video person ReID, which are closely related with this work.

Temporal pooling is widely used to aggregate features across all time stamps. Zheng *et al*. [56] apply max and mean pooling to get the video feature. Li *et al*. [22] utilize part cues and learn a weighting strategy to fuse features extracted from video frames. Suh *et al*. [39] propose a two-stream architecture to jointly learn the appearance feature and part feature, and fuse the image level features through a pooling strategy. Average pooling is also used in

recent works [21, 47], which apply unsupervised learning for video person ReID. Temporal pooling exhibits promising efficiency, but extracts frame features independently and ignores the temporal orders among adjacent frames.

Optical flow encodes the short-term motion between adjacent frames. Many works utilize optical flow to learn temporal features [36, 8, 5]. Simonyan *et al*. [36] introduce a two-stream network to learn spatial feature and temporal feature from stacked optical flows. Feichtenhofer *et al*. [7] leverage optical flow to learn spatial-temporal features, and evaluate different types of motion interactions between two streams. Chung *et al*. [5] introduce a two stream architecture for appearance and optical flow, and investigate the weighting strategy for those two streams. Mclaughlin *et al*. [30] introduce optical flow and RNN to exploit long and short term temporal cues. One potential issue of optical flow is its sensitive to spatial misalignment errors, which commonly exist between adjacent person bounding boxes.

Recurrent Neural Network (RNN) is also adopted for video feature learning in video person ReID. Mclaughlin *et al*. [30] first extract image level features, then introduce RNN to model temporal cues cross frames. The outputs of RNN are then combined through temporal pooling as the final video feature. Liu *et al*. [29] propose a recurrent architecture to aggregate the frame-level representations and yield a sequence-level human feature representation. RNN introduces a certain number of fully-connected layers and gates for temporal cue modeling, making it complicated and difficult to train.

3D convolution directly extracts spatial-temporal features through end-to-end CNN training. Recently, deep 3D CNN is introduced for video representation learning. Tran *et al*. [41] propose C3D networks for spatial-temporal feature learning. Qiu *et al*. [32] factorize the 3D convolutional filters into spatial and temporal components, which yield performance gains. Li *et al*. [19] build a compact Multiscale 3D (M3D) convolution network to learn multi-scale temporal cues. Although 3D CNN has exhibited promising performance, it is still sensitive to spatial misalignments and needs to stack a certain number of 3D convolutional kernels, resulting in large parameter overheads and increased difficult for CNN optimization.

This paper learns GLTR through posting DTP and TSA modules on frame features. Compared with existing temporal pooling strategies, our approach jointly captures global and local temporal cues, hence exhibits stronger temporal cue modeling capability. It is easier to optimize than RNN and presents better robustness to misalignment errors than optical flow. Compared with 3D CNN, our model has a more simple architecture and could easily leverage representations developed for image person ReID. As shown in our experiments, our approach outperforms the recent 3D CNN model M3D [19] and the recurrent model STMP [29].
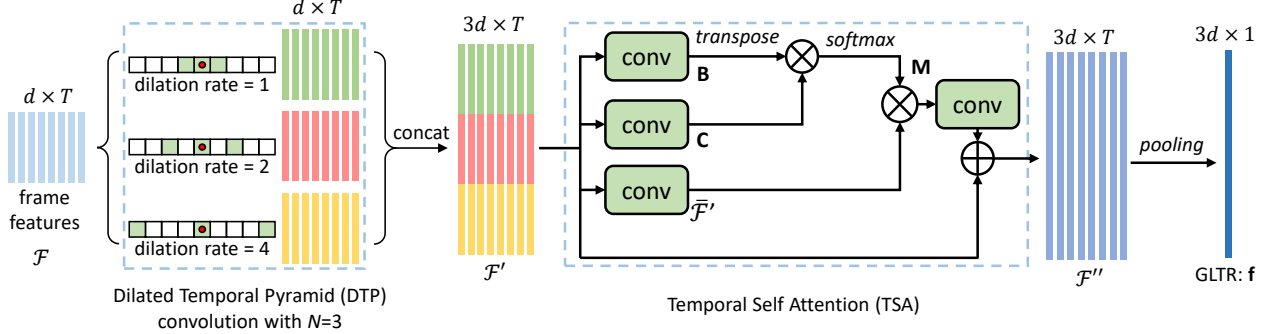
Figure 2. Illustration of our frame feature aggregation subnetwork for GLTR extraction, which consists of Dilated Temporal Pyramid (DTP) convolution for local temporal context learning and Temporal Self-Attention (TSA) model to exploit the global temporal cues.

3. Proposed Methods

3.1. Formulation

Video person ReID aims to identify a gallery video that is about the same person with a query video from a gallery set containing $\mathcal{K}$ videos. A gallery video is denoted by $\mathcal{G}^k = \{I_1^k, I_2^k, ..., I_{T^k}^k\}$ with $k \in \{1, 2, ..., \mathcal{K}\}$, and the query video is denoted by $\mathcal{Q} = \{I_1^q, I_2^q, ..., I_{T^q}^q\}$, where T^k (T^q) denotes the number of frames in the sequence and I_t^k (I_t^q) is the t-th frame. A gallery video $\mathcal{G}$ will be identified as true positive, if it has the closest distance to the query based on a video representation, i.e.,

$$\mathcal{G} = \arg\min_k \text{dist}(\mathbf{f}^{\mathcal{G}^k}, \mathbf{f}^{\mathcal{Q}}), \tag{1}$$

where $\mathbf{f}^{\mathcal{G}^k}$ and $\mathbf{f}^{\mathcal{Q}}$ are the representations of the gallery video $\mathcal{G}^k$ and the query video $\mathcal{Q}$, respectively.

Our approach consists of two subnetworks to learn a discriminative video representation $\mathbf{f}$, i.e., image feature extraction subnetwork and frame feature aggregation subnetwork, respectively. The first subnetwork extracts features of T frames, i.e., $\mathcal{F} = \{f_1, f_2, \ldots, f_T\}$, where $f_t \in \mathcal{R}^d$. The second subnetwork aggregates the T frame features into a single video representation vector. We illustrate the second subnetwork, which is the focus of this work in Fig. 2. We briefly demonstrate the computation of DTP and TSA in the following paragraphs.

The DTP is designed to capture the local temporal cues among adjacent frames. As shown in Fig. 2, DTP takes frame features in $\mathcal{F}$ as input and outputs the updated frame feature $\mathcal{F}' = \{f_1', f_2', \ldots, f_T'\}$. Each $f_t' \in \mathcal{F}'$ is computed by aggregating its adjacent frame features, i.e.,

$$f_t' = \mathcal{M}_{DTP}(f_{t-i}, ..., f_{t+i}), \tag{2}$$

where $\mathcal{M}_{DTP}$ denotes the DTP model, and f_t' is computed from $2 \times i$ adjacent frames.

The TSA model exploits the relation among inconsecutive frames to capture the global temporal cues. It takes $\mathcal{F}'$ as input and outputs the temporal feature $\mathcal{F}'' = \{f_1'', f_2'', \ldots, f_T''\}$. Each $f_t'' \in \mathcal{F}''$ is computed by considering the contextual relations among features inside $\mathcal{F}'$, i.e.,

$$f_t'' = \mathcal{M}_{TSA}(\mathcal{F}', f_t'), \tag{3}$$

where $\mathcal{M}_{TSA}$ is the TSA model.

Each f_t'' aggregates both local and global temporal cues. We finally apply average pooling on $\mathcal{F}''$ to generate the fixed length GLTR $\mathbf{f}$ for video person ReID, i.e.,

$$\mathbf{f} = \frac{1}{T} \sum_{t=1}^{T} f_t''. \tag{4}$$

Average pooling is also commonly applied in RNN [30] and 3DCNN [19] to generate fixed-length video feature. The global and local temporal cues embedded in each f_t'' guarantee the strong discriminative power and robustness of $\mathbf{f}$. The following parts introduce the design of DTP and TSA.

3.2. Dilated Temporal Pyramid Convolution

Dilated Temporal Convolution: Dilated spatial convolution has been widely used in image segmentation for its efficient spatial context modeling capability [52]. Inspired by dilated spatial convolution, we implement dilated temporal convolution for local temporal feature learning. Suppose the $W \in \mathcal{R}^{d \times w}$ is a convolutional kernel with temporal width w. With input frame features $\mathcal{F} = \{f_1, f_2, \ldots, f_T\}$, the output $\mathcal{F}^{(r)}$ of dilated convolution with dilation rate r can be defined as,

$$\mathcal{F}^{(r)} = \{ f_1^{(r)}, f_2^{(r)}, ..., f_T^{(r)} \},$$
$$f_t^{(r)} = \sum_{i=1}^{w} f_{[t+r \cdot i]} \times W_{[i]}^{(r)}, f_t^{(r)} \in \mathcal{R}^d, \tag{5}$$

where $\mathcal{F}^{(r)}$ is the collection of output features containing $f_t^{(r)}$. $W^{(r)}$ denotes dilated convolution with dilation rate r.

The dilation rate r indicates the temporal stride for sampling frame features. It decides the temporal scales covered by dilated temporal convolution. For instance, with
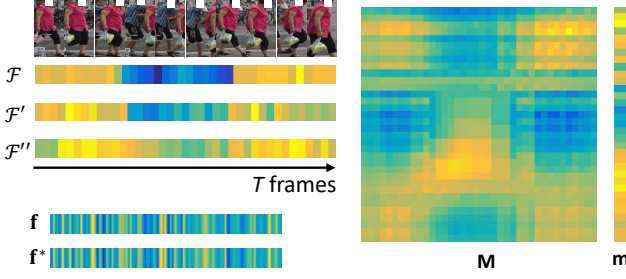
Figure 3. Visualization of $\mathcal{F}$, $\mathcal{F}'$, $\mathcal{F}''$, $\mathbf{M}$ and $\mathbf{f}$ computed on a tracklet with occlusions. Dimensionality of $\mathcal{F}$, $\mathcal{F}'$, $\mathcal{F}''$ is reduced to $1 \times T$ by PCA for visualization. It is clear that, occlusion affects the baseline feature $\mathcal{F}$, i.e., feature substantially changes as occlusion happens. DTP and TSA progressively alleviate the occlusions, i.e., features of occluded frames in $\mathcal{F}'$ and $\mathcal{F}''$ appear similar to the others. $\mathbf{f}^*$ is generated after manually removing occluded frames. $\mathbf{f}$ is quite close to $\mathbf{f}^*$, indicating the strong robustness of GLTR to occlusion.

$r = 2, w = 3$, each output feature corresponds to a temporal range of five adjacent frames. Standard convolution can be regarded as a special case with $r = 1$, which covers three adjacent frames. Compared with standard convolution, dilated temporal convolution with $r \geq 2$ has the same number of parameters to learn, but enlarges the receptive field of neurons without reducing the temporal resolution. This property makes dilated temporal convolution an efficient strategy for multi-scale temporal feature learning.

Dilated Temporal Pyramid Convolution: Dilated temporal convolutions with different dilation rates model temporal cues at different scales. We hence use parallel dilated convolutions to build the DTP convolution to enhance its local temporal cues modeling ability.

As illustrated in Fig. 2, DTP convolution consists of N parallel dilated convolutions with dilation rates increasing progressively to cover various temporal ranges. For n-th dilated temporal convolution, we set its dilation rate r_n as $r_n = 2^{n-1}$ to efficiently enlarge the temporal receptive fields. We concatenate the outputs from N branches as the updated temporal feature $\mathcal{F}'$, i.e., we compute $f'_t \in \mathcal{F}'$ as

$$f'_t = concat(f_t^{(r_1)}, f_t^{(r_2)}, ..., f_t^{(r_N)}), f'_t \in \mathcal{R}^{Nd}, \quad (6)$$

where r_i is the dilation rate of i-th dilated temporal convolutions.

3.3. Temporal Self Attention

Self-Attention: The self-attention module is recently used to learn the long-range spatial dependencies in image segmentation [10, 15, 53], action recognition [42] and image person ReID [16, 1]. Inspired by its promising performance in spatial context modeling, we generalize self-attention to to capture the contextual temporal relations among inconsecutive frames.

Temporal Self-Attention: The basic idea of TSA is to compute an $T \times T$ sized attention mask $\mathbf{M}$ to store the contextual relations among all frame features. As illustrated in Fig. 2, given the input $\mathcal{F}' \in R^{Nd \times T}$, TSA first applies two convolution layers followed by Batch Normalization and ReLU to generate feature maps $\mathbf{B}$ and $\mathbf{C}$ with size $(Nd/\alpha) \times T$, respectively. Then, it performs a matrix multiplication between $\mathbf{C}$ and the transpose of $\mathbf{B}$, resulting in a $T \times T$ sized temporal attention mask $\mathbf{M}$.

$\mathbf{M}$ is applied to update the $\mathcal{F}'$ to embed extra global temporal cues. $\mathcal{F}'$ is fed into a convolution layer to generate a new feature map $\bar{\mathcal{F}}'$ with size $(Nd/\alpha) \times T$. $\bar{\mathcal{F}}'$ is hence multiplied by $\mathbf{M}$ and is fed into a convolution layer to recover its size to $Nd \times T$. The resulting feature map is fused with the original $\mathcal{F}'$ by residual connection, leading to the updated temporal feature $\mathcal{F}''$. The computation of TSA can be denoted as

$$\mathcal{F}'' = W * (\bar{\mathcal{F}}' \cdot \mathbf{M}) + \mathcal{F}', \mathcal{F}'' \in \mathcal{R}^{Nd \times T}, \quad (7)$$

where W denotes the last convolutional kernel. W is initialized as 0 to simplify the optimization of residual connection. α controls the parameter size in TSA. We experimentally set α as 2. $\mathcal{F}''$ is processed with average pooling to generate the final GLTR $\mathbf{f} \in \mathcal{R}^{Nd}$.

We visualize the $\mathcal{F}$, $\mathcal{F}'$, $\mathcal{F}''$, $\mathbf{M}$, and $\mathbf{f}$ computed on a tracklet with occlusion in Fig. 3. DTP reasonably alleviates occlusion by applying convolutions to adjacent features. TSA alleviates occlusion mainly by computing the attention mask $\mathbf{M}$, which stores the global contextual relations as shown in Fig. 3. With $\mathbf{M}$, average pooling on $\mathcal{F}''$ can be conceptually expressed as:

$$\sum_{t=1}^{T} \mathcal{F}''(:, t) \doteq \sum_{t=1}^{T} \mathcal{F}'(:, t) \times \mathbf{m}(t) + \sum_{t=1}^{T} \mathcal{F}'(:, t), \quad (8)$$

where $\mathbf{m} = \sum_{t=1}^{T} \mathbf{M}(:, t)$ is a T-dim weighting vector. Note that, Eq. (8) omits the convolutions before and after $\bar{\mathcal{F}}'$ to simplify the expression. $\mathbf{m}$ is visualized in Fig. 3, where occluded frames presents lower weights, indicating their features are depressed during average pooling. Combining DTP and TSA, GLTR presents strong robustness.

4. Experiment

4.1. Dataset

We test our methods on four widely used video ReID datasets, and a novel large-scale dataset. Example images are depicted in Fig. 5 and statistics are given in Table 1.

PRID-2011 [14]. There are 400 sequences of 200 pedestrians captured by two cameras. Each sequence has a length between 5 and 675 frames.

iLIDS-VID [43]. There are 600 sequences of 300 pedestrians from two cameras. Each sequence has a variable length

Table 1. The statistics of our LS-VID dataset and other video person ReID datasets.

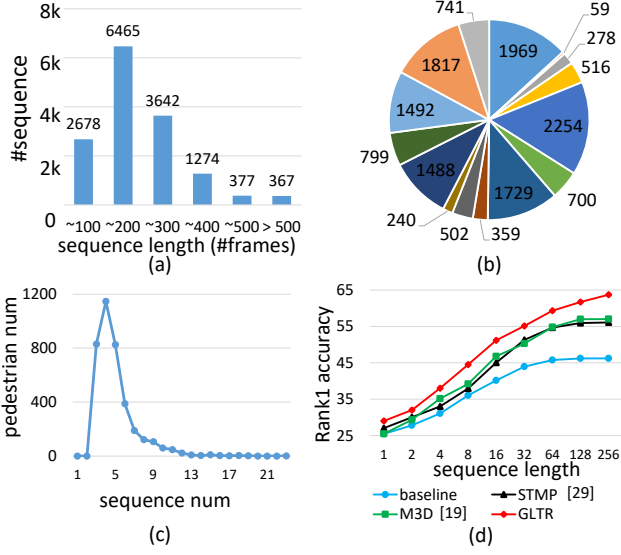| dataset	#identity	#sequence	#boxes	#frame	#indoor cam.	#outdoor cam.	detector	val. set	evaluation
DukeMTMC	1,404	4,832	815,420	168	0	8	Hand	×	CMC + mAP
MARS	1,261	20,715	1,067,516	58	0	6	DPM	×	CMC + mAP
PRID	200	400	40,033	100	0	2	Hand	×	CMC
iLIDS-VID	300	600	42,460	73	2	0	Hand	×	CMC
LS-VID	**3,772**	**14,943**	**2,982,685**	**200**	**3**	**12**	**Faster R-CNN**	**✓**	**CMC + mAP**

Figure 4. Some statistics on LS-VID dataset: (a) the number of sequences with different length; (b) the number of sequences in each of the 15 cameras; (c) the number of identities with different sequence number; (d) the ReID performance with different testing sequence length.
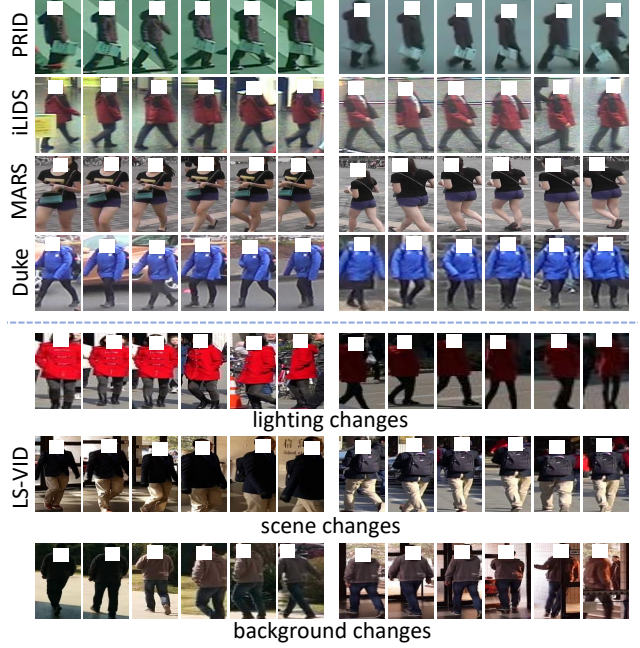

Figure 5. Frames evenly sampled from person tracklets. Each row shows two sequences of the same person under different cameras. Compared with existing datasets, LS-VID presents more substantial variations of lighting, scene, and background, *etc*. We cover the face in for privacy purpose.

between 23 and 192 frames. Following the implementation in previous works [43, 22], we randomly split this two datasets into train/test identities. This procedure is repeated 10 times for computing averaged accuracies.

MARS [56]. This dataset is captured by 6 cameras. It consists of 17,503 sequences of 1,261 identities and 3,248 distractor sequences. It is split into 625 identities for training and 636 identities for testing. The bounding boxes are detected with DPM detector [9], and tracked using the GMMCP tracker [6]. we follow the protocol of MARS and report the Rank1 accuracy and mean Average Precision (mAP).

DukeMTMC-VideoReID [47, 34]. There are 702 identities for training, 702 identities for testing, and 408 identities as distractors. The training set contains 369,656 frames of 2,196 tracklets, and test set contains 445,764 frames of 2,636 tracklets.

LS-VID. Besides the above four datasets, we collect a novel Large-Scale Video dataset for person ReID (LS-VID).

Raw video capture: We utilize a 15-camera network and

select 4 days for data recording. For each day, 3 hours of videos are taken in the morning, noon, and afternoon, respectively. Our final raw video contains 180 hours videos, 12 outdoor cameras, 3 indoor cameras, and 12 time slots.

Detection and tracking: Faster RCNN [33] is utilized for pedestrian detection. After that, we design a feature matching strategy to track each detected pedestrian in each camera. After discarding some sequences with too short length, we finally collect 14,943 sequences of 3,772 pedestrians, and the average sequence length is 200 frames.

Characteristics: Example sequences in LS-VID are shown in Fig. 5, and statistics are given in Table 1 and Fig. 4. LS-VID shows the following new features: (1)Longer sequences. (2) More accurate pedestrian tracklets. (3) Currently the largest video ReID dataset. (4) Define a more realistic and challenging ReID task.

Evaluation protocol: Because of the expensive data annotation, we randomly divide our dataset into training set
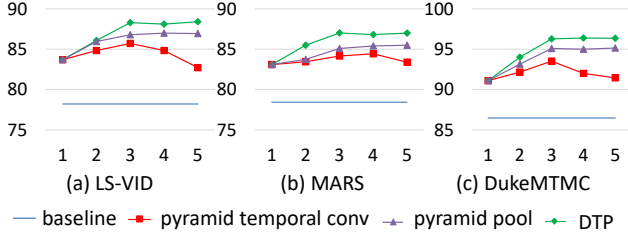
Figure 6. Rank1 accuracy of DTP and two competitors on three datasets with different numbers of branches, *i.e.*, parameter N.

Table 2. Performance of individual components in GLTR.

Dataset	LS-VID		MARS		DukeMTMC		PRID	iLIDS
Method	mAP	rank1	mAP	rank1	mAP	rank1	rank1	rank1
baseline	30.72	46.18	65.45	78.43	82.08	86.47	83.15	62.67
DTP	41.78	59.92	75.90	85.74	89.98	93.02	93.26	84.00
TSA	40.01	58.73	75.62	85.40	89.26	92.74	92.14	83.33
GLTR	**44.32**	**63.07**	**78.47**	**87.02**	**93.74**	**96.29**	**95.50**	**86.00**

Table 3. Performance of GLTR with different backbones on LS-VID test set.

method	backbone	mAP	rank1	rank5	rank10	rank20
baseline	Alexnet [17]	15.98	24.23	43.52	53.45	62.13
	Inception [40]	22.77	35.70	55.88	64.89	73.12
	ResNet50 [12]	30.72	46.18	67.41	74.71	82.33
GLTR	Alexnet [17]	22.57	35.45	56.59	66.01	75.06
	Inception [40]	35.75	51.83	71.66	79.19	84.79
	ResNet50 [12]	**44.43**	**63.07**	**77.22**	**83.81**	**88.41**

and test set with 1:3 ratio to encourage more efficient training strategies. We further divide a small validation set. Finally, the training set contains 550,419 bounding boxes of 842 identities, the validation set contains 155,191 bounding boxes of 200 identities, and the test set contains 2,277,075 bounding boxes of 2,730 identities. Similar to existing video ReID datasets [56, 47], LS-VID utilizes the Cumulated Matching Characteristics (CMC) curve and mean Average Precision (mAP) as evaluation metric.

4.2. Implementation Details

We employ standard ResNet50 [12] as the backbone for frame feature extraction. All models are trained and fine-tuned with PyTorch. Stochastic Gradient Descent (SGD) is used to optimize our model. Input images are resized to 256×128. The mean value is subtracted from each (B, G, and R) channel. For 2D CNN training, each batch contains 128 images. The initial learning rate is set as 0.01, and is reduced ten times after 10 epoches. The training is finished after 20 epoches. For DTP and TSA training, we sample 16 adjacent frames from each sequence as input for each training epoch. The batch size is set as 10. The initial learning rate is set as 0.01, and is reduced ten times after 120 epoches. The training is finished after 400 epoches. All models are trained with only softmax loss.

During testing, we use 2D CNN to extract a d=128-dim feature from each video frame, then fuse frame features into GLTR using the network illustrated in Fig. 2. The video feature is finally used for person ReID with Euclidean distance. All of our experiments are implemented with GTX TITAN X GPU, Intel i7 CPU, and 128GB memory.

4.3. Ablation Study

Comparison of DTP and other local temporal cues learning strategies: Besides DTP, we also implement the following strategies to learn temporal cues among adjacent frames: (i) pyramid temporal convolution without dilation, and (ii) temporal pyramid pooling [54]. As explained in Sec. 3.2, the dilation rate of i-th pyramid branch in DTP is $r_i = 2^{i-1}$. To make a fair comparison, we set three methods have the same number of branches, where each has the same size of receptive field. For instance, we set the convolution

kernel size as $d \times 9$ for the 3-rd branch of pyramid temporal convolution without dilation. The experiment results on MARS, DukeMTMC-VideoReID, and the **validation** set of LS-VID are summarized in Fig. 6.

Fig. 6 also compares average pooling as the baseline. It is clear that, three methods perform substantially better than baseline, indicating that average pooling is not effective in capturing the temporal cues among frame features. With N=1, the three methods perform equally, *i.e.*, apply a $d \times 3$ sized convolution kernel to frame feature $\mathcal{F}$. As we increase N, the performances of three algorithms are boosted. This means that introducing multiple convolution scales benefits the learned temporal feature.

It is also clear that, DTP consistently outperforms the other two strategies on three datasets. The reason may be because the temporal pyramid pooling loses certain temporal cues as it down-samples the temporal resolution. The traditional temporal convolution introduces too many parameters, leading to difficult optimization. The dilated convolutions in DTP efficiently enlarge the temporal respective fields hence performs better for local temporal feature learning. With $N \geq 3$, the performance boost slows down for DTP. Further introducing more branches increases the size of parameters and causes more difficult optimization. We select $N = 3$ for DTP in the following experiments.

Validity of combining DTP and TSA: This part proceeds to evaluate that combining DTP and TSA results in the best video feature. We compare several variants of our methods and summarize the results on four datasets and the **test set** of LS-VID in Table 2. In the table, "baseline" denotes the ResNet50 + average pooling. "DTP" and "TSA" denote aggregating frame feature only with DTP or TSA, respectively. "GLTR" combines DTP and TSA.

Table 2 shows that either DTP or TSA performs substantially better than the baseline, indicating modeling extra local and global temporal cues results in better video feature.

Table 4. Comparison with recent works on LS-VID test set.

Method	mAP	rank1	rank5	rank10	rank20
ResNet50 [12]	30.72	46.18	67.41	74.71	82.33
GLAD [45]	33.98	49.34	70.15	77.14	83.59
HACNN [24]	36.65	53.93	72.41	80.88	85.27
PBR [39]	37.58	55.34	74.68	81.56	86.16
DRSA [22]	37.77	55.78	74.37	81.06	86.81
Two-stream [36]	32.12	48.23	68.66	75.06	83.56
LSTM [50]	35.92	52.11	72.57	78.91	85.50
I3D [2]	33.86	51.03	70.08	78.08	83.65
P3D [32]	34.96	53.37	71.15	78.08	83.65
STMP [29]	39.14	56.78	76.18	82.02	87.12
M3D [19]	40.07	57.68	76.09	83.35	88.18
GLTR	**44.32**	**63.07**	**77.22**	**83.81**	**88.41**

Table 5. Comparison with recent works on MARS.

Method	mAP	rank1	rank5	rank20
BoW+kissme [56]	15.50	30.60	46.20	59.20
IDE+XQDA [56]	47.60	65.30	82.00	89.00
SeeForest [59]	50.70	70.60	90.00	97.60
QAN [28]	51.70	73.70	84.90	91.60
DCF [18]	56.05	71.77	86.57	93.08
TriNet [13]	67.70	79.80	91.36	-
MCA [37]	71.17	77.17	-	-
DRSA [22]	65.80	82.30	-	-
DuATM [35]	67.73	81.16	92.47	-
MGCAM [37]	71.17	77.17	-	-
PBR [39]	75.90	84.70	92.80	95.00
CSA [3]	76.10	86.30	94.70	98.20
STMP [29]	72.70	84.40	93.20	96.30
M3D [19]	74.06	84.39	93.84	97.74
STA [11]	**80.80**	86.30	95.70	98.10
GLTR	78.47	**87.02**	**95.76**	**98.23**

DTP model achieves rank1 accuracy of 85.74% on MARS dataset, outperforming the baseline by large margin. Similarly, TSA also performs substantially better than the baseline. By combining DTP and TSA, the GLTR consistently achieves the best performance on five datasets. We hence conclude that, jointly learning local and global temporal cues results in the best video feature.

Different backbones: We further evaluate the effectiveness of GLTR with different backbone networks, including Alexnet [17], Inception [40] and ResNet50 [12]. Experimental results on the test set of LS-VID are summarized in Table 3. Table 3 shows that, implemented on different backbones, GLTR consistently outperforms baselines, indicating that our methods work well with different frame feature extractors. GLTR thus could leverage strong image representations and serve as a general solution for video person ReID. Since ResNet50 achieves best performance in Table 3, we adopt ResNet50 in the following experiments.

4.4. Comparison With Recent Works

LS-VID: This section compares several recent methods with our approach on LS-VID **test set**. To make a comparison on LS-VID, we implement several recent works with code provided by their authors, including temporal feature learning methods for person reid: M3D [19] and STMP [29], other temporal feature learning methods: two-stream CNN with appearance and optical flow [36], LSTM [50], 3D convolution: I3D [2] and P3D [32], as well as recent person ReID works: GLAD [45], HACNN [24], PBR [39] and DRSA [22], respectively. Video features of GLAD [45] and HACNN [24] are extracted by average pooling. We repeat PBR [39] and DRSA [22] by referring to their implantations on MARS. Table 4 summarizes the comparison.

Table 4 shows that, GLAD [45] and HACNN [24] get promising performance in image person ReID, but achieve lower performance than temporal feature learning strategies, *e.g.*, M3D [19] and STMP [29]. This indicates the importance of learning temporal cues in video person ReID.

Among those compared temporal feature learning methods, the recent M3D achieves the best performance. In Table 4, the proposed GLTR achieves the best performance. It outperforms the recent video person ReID work STMP [29] and M3D [19] by large margins, *e.g.*, 6.29% and 5.39% in rank1 accuracy, respectively.

MARS: Table 5 reports the comparison with recent works on MARS. GLTR achieves the rank1 accuracy of 87.02% and mAP of 78.47%, outperforming most of the recent work, *e.g.*, STMP [29], M3D [19] and STA [11] by 2.62%, 2.63%, and 0.72% in rank1 accuracy, respectively. Note that, STMP [29] introduces a complex recurrent network and uses part cues and triplet loss. M3D [19] use 3D CNN to learn the temporal cues, hence requires higher computational complexity. STA [11] achieves competitive performance on MARS dataset, and outperform GLTR on mAP. Note that, STA introduces multi-branches for part feature learning and uses triplet loss to promote the performance. Compared with those works, our method achieves competitive performance with simple design., *e.g.*, we extract global feature with basic backbone and train only with the softmax loss. GLTR can be further combined with a re-ranking strategy [58], which further boosts its mAP to 85.54%.

PRID and iLIDS-VID: The comparisons on PRID and iLIDS-VID datasets are summarized in Table 6. It shows that, our method presents competitive performance on rank1 accuracy. M3D [19] also gets competitive performance on both datasets. The reason may be because the M3D jointly learns the multi-scale temporal cues form video sequences, and introduces a two-stream architecture to learn the spatial and temporal representations respectively. With a single feature extraction stream design, our method still outperforms M3D on both datasets. Table 6 also compares with several temporal feature learning methods, *e.g.*, RFA-Net [50], SeeForest [59], T-CN [48], CSA [3] and STMP [29]. Our method outperforms those works by large margins in rank1 accuracy.
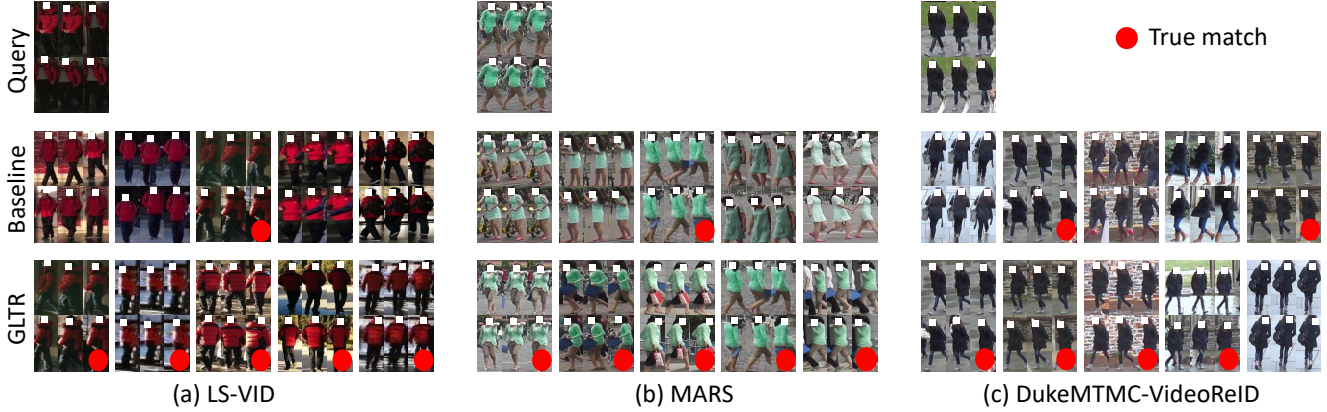
Figure 7. Illustration of person ReID results on LS-VID, MARS and DukeMTMC-VideoReID datasets. Each example shows the top-5 retrieved sequences by baseline method (first tow) and GLTR (second tow), respectively. The true match is annotated by the red dot. We cover the face for privacy purpose.

Table 6. Comparison with recent works on PRID and iLIDS-VID.

Dataset	PRID		iLIDS-VID	
Method	rank1	rank5	rank1	rank5
BoW+XQDA [56]	31.80	58.50	14.00	32.20
IDE+XQDA [56]	77.30	93.50	53.00	81.40
DFCP [25]	51.60	83.10	34.30	63.30
AMOC [26]	83.70	98.30	68.70	94.30
QAN [28]	90.30	98.20	68.00	86.80
DRSA [22]	93.20	-	80.20	-
RCN [30]	70.00	90.00	58.00	84.00
DRCN [46]	69.00	88.40	46.10	76.80
RFA-Net [50]	58.20	85.80	49.30	76.80
SeeForest [59]	79.40	94.40	55.20	86.50
T-CN [48]	81.10	85.00	60.60	83.80
CSA [3]	93.00	99.30	85.40	96.70
STMP [29]	92.70	98.80	84.30	96.80
M3D [19]	94.40	100.00	74.00	94.33
GLTR	**95.50**	**100.00**	**86.00**	**98.00**

Table 7. Comparison on DukeMTMC-VideoReID.

Method	mAP	rank1	rank5	rank20
ETAP-Net [47]	78.34	83.62	94.59	97.58
STA [11]	**94.90**	96.20	99.30	99.60
GLTR	93.74	**96.29**	**99.30**	**99.71**

lower than the 87.02% on MARS. This shows that, even though LS-VID collects longer sequences to provide more abundant spatial and visual cues, it still presents a more challenging person ReID task.

We show some person ReID results achieved by GLTR and ResNet50 baseline on LS-VID, MARS [56] and DukeMTMCVideoReID [47, 34] in Fig. 7. For each query, we show the top5 returned video sequences by those two methods. It can be observed that, the proposed GLTR is substantially more discriminative for identifying persons with similar appearance.

5. Conclusion

This paper proposes the Global Local Temporal Representation (GLTR) for video person ReID. Our proposed network consists of the DTP convolution and TSA model, respectively. The DTP consists of parallel dilated temporal convolutions to model the short-term temporal cues among adjacent frames. TSA exploits the relation among inconsecutive frames to capture global temporal cues. Experimental results on five benchmark datasets demonstrate the superiority of the proposed GLTR over current state-of-the-art methods.

DukeMTMC-VideoReID: Comparisons on this dataset are shown in Table 7. Because DukeMTMC-VideoReID is a recently proposed video ReID dataset, a limited number of works have reported performance on it. We compare with ETAP-Net [47] and STA [11] in this section. The reported performance of ETAP-Net [47] in Table 7 is achieved with a supervised baseline. As shown in Table 7, GLTR achieves 93.74% mAP and 96.29% rank1 accuracy, outperforming ETAP-Net [47] by large margins. The STA [11] also achieves competitive performance on this dataset. GLTR still outperforms STA [11] on rank1, rank5, and rank20 accuracy, respectively. Note that, STA [11] utilizes extra body part cues and triplet loss.

Summary: The above comparisons on five datasets could indicate the advantage of GLTR in video representation learning for person ReID, *i.e.*, it achieves competitive accuracy with simple and concise model design. We also observe that, the ReID accuracy on LS-VID is substantially lower than the ones on the other datasets. For example, the best rank1 accuracy on LS-VID is 63.07%, substantially

References

[1] Jean-Paul Ainam, Ke Qin, and Guisong Liu. Self attention grid for person re-identification. *arXiv preprint arXiv:1809.08556*, 2018.

[2] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017.

[3] Dapeng Chen, Hongsheng Li, Tong Xiao, Shuai Yi, and Xiaogang Wang. Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding. In *CVPR*, 2018.

[4] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. PAMI*, 40(4):834–848, 2017.

[5] Dahjung Chung, Khalid Tahboub, and Edward J Delp. A two stream siamese convolutional neural network for person re-identification. In *ICCV*, 2017.

[6] Afshin Dehghan, Shayan Modiri Assari, and Mubarak Shah. Gmmcp tracker: Globally optimal generalized maximum multi clique problem for multiple object tracking. In *CVPR*, 2015.

[7] Christoph Feichtenhofer, Axel Pinz, and Richard P Wildes. Spatiotemporal multiplier networks for video action recognition. In *CVPR*, 2017.

[8] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *CVPR*, 2016.

[9] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE Trans. PAMI*, 32(9), 2010.

[10] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *CVPR*, 2019.

[11] Yang Fu, Xiaoyang Wang, Yunchao Wei, and Thomas Huang. Sta: Spatial-temporal attention for large-scale video-based person re-identification. In *AAAI*. 2019.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[13] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

[14] Martin Hirzer, Csaba Beleznai, Peter M Roth, and Horst Bischof. Person re-identification by descriptive and discriminative classification. In *SCIA*, 2011.

[15] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Crisscross attention for semantic segmentation. *arXiv preprint arXiv:1811.11721*, 2018.

[16] Minyue Jiang, Yuan Yuan, and Qi Wang. Self-attention learning for person re-identification. In *BMVC*, 2018.

[17] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[18] Dangwei Li, Xiaotang Chen, Zhang Zhang, and Kaiqi Huang. Learning deep context-aware features over body and latent parts for person re-identification. In *CVPR*, 2017.

[19] Jianing Li, Shiliang Zhang, and Tiejun Huang. Multiscale 3d convolution network for video based person re-identification. In *AAAI*, 2019.

[20] Jianing Li, Shiliang Zhang, Qi Tian, Meng Wang, and Wen Gao. Pose-guided representation learning for person re-identification. *IEEE Trans. PAMI*, 2019.

[21] Minxian Li, Xiatian Zhu, and Shaogang Gong. Unsupervised person re-identification by deep learning tracklet association. In *ECCV*, 2018.

[22] Shuang Li, Slawomir Bak, Peter Carr, and Xiaogang Wang. Diversity regularized spatiotemporal attention for video-based person re-identification. In *CVPR*, 2018.

[23] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, 2014.

[24] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *CVPR*, 2018.

[25] Youjiao Li, Li Zhuo, Jiafeng Li, Jing Zhang, Xi Liang, and Qi Tian. Video-based person re-identification by deep feature guided pooling. In *CVPR Workshops*, 2017.

[26] Hao Liu, Zequn Jie, Karlekar Jayashree, Meibin Qi, Jianguo Jiang, Shuicheng Yan, and Jiashi Feng. Video-based person re-identification with accumulative motion context. *IEEE Trans. CSVT*, 28(10):2788–2802, 2017.

[27] Jiawei Liu, Zheng-Jun Zha, Xuejin Chen, Zilei Wang, and Yongdong Zhang. Dense 3d-convolutional neural network for person re-identification in videos. *ACM Trans. on TOMM*, 15(1s):8, 2019.

[28] Yu Liu, Junjie Yan, Wanli Ouyang, and Wanli Ouyang. Quality aware network for set to set recognition. In *CVPR*, 2017.

[29] Yiheng Liu, Zhenxun Yuan, Wengang Zhou, and Houqiang Li. Spatial and temporal mutual promotion for video-based person re-identification. In *AAAI*, 2019.

[30] Niall McLaughlin, Jesus Martinez del Rincon, Paul Miller, and Paul Miller. Recurrent convolutional network for video-based person re-identification. In *CVPR*, 2016.

[31] Sateesh Pedagadi, James Orwell, Sergio Velastin, and Boghos Boghossian. Local fisher discriminant analysis for pedestrian re-identification. In *CVPR*, 2013.

[32] Zhaofan Qiu, Ting Yao, and Tao Mei. Learning spatiotemporal representation with pseudo-3d residual networks. In *ICCV*, 2017.

[33] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015.

[34] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *ECCV*, 2016.

[35] Jianlou Si, Honggang Zhang, Chun-Guang Li, Jason Kuen, Xiangfei Kong, Alex C Kot, and Gang Wang. Dual attention matching network for context-aware feature sequence based person re-identification. In *CVPR*, 2018.

[36] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, 2014.

[37] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. Mask-guided contrastive attention model for person re-identification. In *CVPR*, 2018.

[38] Chi Su, Jianing Li, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. Pose-driven deep convolutional model for person re-identification. In *ICCV*, 2017.

[39] Yumin Suh, Jingdong Wang, Siyu Tang, Tao Mei, and Kyoung Mu Lee. Part-aligned bilinear representations for person re-identification. In *ECCV*, 2018.

[40] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016.

[41] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, 2015.

[42] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, 2018.

[43] Xiaogang Wang and Rui Zhao. Person re-identification: System design and evaluation overview. In *Person Re-Identification*, pages 351–370. Springer, 2014.

[44] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *CVPR*, 2018.

[45] Longhui Wei, Shiliang Zhang, Hantao Yao, Wen Gao, and Qi Tian. Glad: global-local-alignment descriptor for pedestrian retrieval. In *ACM MM*, 2017.

[46] Lin Wu, Chunhua Shen, Anton van den Hengel, and Anton van den Hengel. Deep recurrent convolutional networks for video-based person re-identification: An end-to-end approach. *arXiv preprint arXiv:1606.01609*, 2016.

[47] Yu Wu, Yutian Lin, Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi Yang. Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning. In *CVPR*, 2018.

[48] Yang Wu, Jie Qiu, Jun Takamatsu, and Tsukasa Ogasawara. Temporal-enhanced convolutional network for person re-identification. In *AAAI*, 2018.

[49] Fei Xiong, Mengran Gou, Octavia Camps, and Mario Sznaier. Person re-identification using kernel-based metric learning methods. In *ECCV*. 2014.

[50] Yichao Yan, Bingbing Ni, Zhichao Song, Chao Ma, Yan Yan, and Xiaokang Yang. Person re-identification via recurrent feature aggregation. In *ECCV*, 2016.

[51] Maoke Yang, Kun Yu, Chi Zhang, Zhiwei Li, and Kuiyuan Yang. Denseaspp for semantic segmentation in street scenes. In *CVPR*, 2018.

[52] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.

[53] Yuhui Yuan and Jingdong Wang. Ocnet: Object context network for scene parsing. *arXiv preprint arXiv:1809.00916*, 2018.

[54] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017.

[55] Liming Zhao, Xi Li, Yueting Zhuang, and Jingdong Wang. Deeply-learned part-aligned representations for person re-identification. In *ICCV*, 2017.

[56] Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, and Qi Tian. Mars: A video benchmark for large-scale person re-identification. In *ECCV*, 2016.

[57] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, 2015.

[58] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In *CVPR*, 2017.

[59] Zhen Zhou, Yan Huang, Wei Wang, Liang Wang, and Tieniu Tan. See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification. In *CVPR*, 2017.