Video-based Person Re-identification with Spatial and Temporal Memory Networks

Chanho Eom Geon Lee Junghyup Lee Bumsub Ham*

School of Electrical and Electronic Engineering, Yonsei University
https://cvlab-yonsei.github.io/projects/STMN

Abstract

Video-based person re-identification (reID) aims to retrieve person videos with the same identity as a query person across multiple cameras. Spatial and temporal distractors in person videos, such as background clutter and partial occlusions over frames, respectively, make this task much more challenging than image-based person reID. We observe that spatial distractors appear consistently in a particular location, and temporal distractors show several patterns, e.g., partial occlusions occur in the first few frames, where such patterns provide informative cues for predicting which frames to focus on (i.e., temporal attentions). Based on this, we introduce a novel Spatial and Temporal Memory Networks (STMN). The spatial memory stores features for spatial distractors that frequently emerge across video frames, while the temporal memory saves attentions which are optimized for typical temporal patterns in person videos. We leverage the spatial and temporal memories to refine frame-level person representations and to aggregate the refined frame-level features into a sequence-level person representation, respectively, effectively handling spatial and temporal distractors in person videos. We also introduce a memory spread loss preventing our model from addressing particular items only in the memories. Experimental results on standard benchmarks, including MARS, DukeMTMC-VideoReID, and LS-VID, demonstrate the effectiveness of our method.

1. Introduction

Person re-Identification (reID) aims at retrieving a person of interest from a set of pedestrian images/videos taken from non-overlapping cameras. Convolutional neural networks (CNNs) have made remarkable advances in image-based person reID [44, 29, 21, 19, 4, 46] over the last decade. Video-based person reID has recently attracted increasing attention in accordance with the prevalence of video capturing systems. Video frames provide rich information to specify a particular person, but they often contain spatial distractors, *e.g.*, trees, bicycles, and concrete pavers. In particular, person videos, typically cropped by off-the-shelf object detectors from a whole sequence, also have temporal distractors, *e.g.*, misaligned persons across video frames or partial occlusions within a sequence.

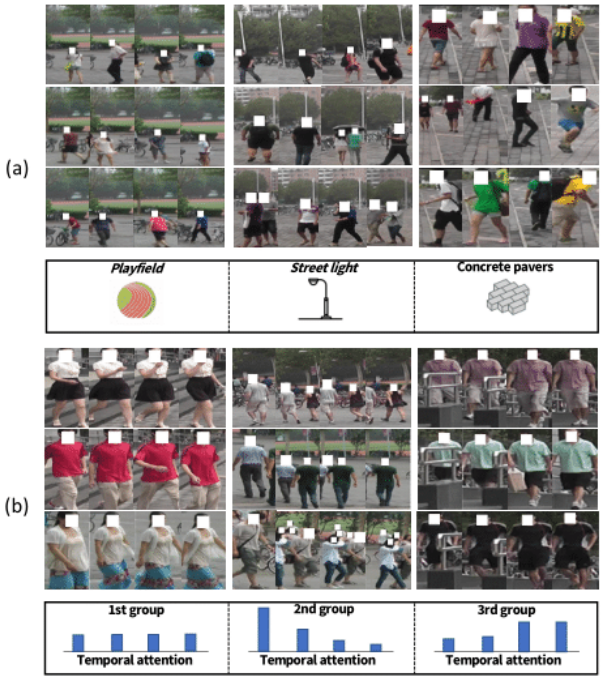

Figure 1. Examples of (a) spatial distractors that appear frequently in surveillance videos and (b) prototypes of temporal patterns that provide important clues to predict temporal attentions.

Recent video reID methods [18, 5] attempt to tackle these issues by exploiting spatial and temporal attention modules, which are useful for extracting person representations robust to noisy regions (*e.g.*, background clutter) and temporal variations (*e.g.*, partial occlusions). They, however, do not consider a global view in a sequence [43, 11], suggesting that these approaches may focus on less discriminative parts or video frames. Several works [17, 16, 20, 30, 40, 41] instead propose to use non-local [36] or graph convolutional networks [13] to capture co-attention

*Corresponding author.

over frames. They focus on the shared information across multiple frames to obtain a person representation from a video, taking into account the temporal context. The co-attention, however, may concentrate on distracting scene details or partial occlusions, which are often shared in successive video frames, producing an incorrect video representation.

We present in this paper Spatial and Temporal Memory Networks (STMN) to extract person representations robust to spatial and temporal distractors for video-based person reID. The main idea is based on the following observations: 1) Since video sequences are captured by a stationary camera, it is likely that they constantly contain background clutter such as a playfield, a street light, or concrete pavers in a particular location (Fig. 1(a)); 2) Temporal patterns, *e.g.*, a person of interest disappears at the end of the sequence (Fig. 1(b) center) or partial occlusions occur in the first few frames (Fig. 1(b) right), provide crucial clues to determine which frames we have to focus on (*i.e.*, temporal attentions).

Based on the observations, we propose to exploit two external memories called spatial and temporal memories. The spatial memory is trained to store spatial distractors that frequently appear across video frames, while the temporal memory is trained to memorize attentions which are optimized for typical temporal patterns in person videos. At test time, we leverage the memories as look-up tables, and ease the difficulty of handling the spatial and temporal distractors from videos of unseen identities. Specifically, we exploit the spatial memory to suppress features for distracting scene details from each frame-level person representation, and the temporal one to aggregate the frame-level person representations focusing more on discriminative frames. We also propose a memory spread loss that encourages our model to access all items in the memories during training. We demonstrate the effectiveness of our method on the MARS [45], DukeMTMC-VideoReID [38], and LS-VID [16] datasets. To our best knowledge, this is an early effort that jointly leverages multiple types of memories. The main contributions of our work can be summarized as follows:

- We introduce a simple yet effective method for video-based person reID, dubbed STMN, which extracts a robust video representation to spatial and temporal distractors using spatial and temporal memories.
- We propose a memory spread loss that prevents our model from accessing few items repeatedly, encouraging all items in the memories to be used.
- We achieve the state of the art on standard video reID benchmarks. Ablation studies further validate the effectiveness of our method.

2. Related Work

Here, we briefly introduce representative works closely related to ours, and clarify their differences from ours.

Video-based person reID. The key for video-based reID is to extract person representations robust to spatial and temporal distractors. Many methods [22, 18, 5] propose to use attention modules for video-based person reID. QAN [22] uses a temporal attention to aggregate frame-level features, focusing on discriminative frames. DRSA [18] and STA [5] additionally use a spatial attention to suppress features for spatial distractors. They, however, assign attentions to each frame without considering whole frames in a sequence, indicating that they may aggregate less discriminative parts or frames in the sequence [43, 11]. Recent methods [17, 16, 20, 30, 40] propose to use co-attention modules between frames by adopting non-local [36] or graph convolutional networks [13]. Specifically, GLTR [16] adds a co-attention module at the end of backbone CNNs, while M3D [17], STE-NVAN [20] and COSAM [30] insert multiple co-attention modules into different levels of backbone CNNs, to refine frame-level person representations, considering contextual temporal relations between frames. The work of [40, 41] introduces a hierarchical co-attention module, dividing frames into multiple granularities, to capture discriminative spatial and temporal features from different semantic levels. These approaches highlight shared information between frames, suppressing features from distracting scene details and occlusion, which is useful only when such distractors appear in a few frames. When similar backgrounds and/or occlusion are shared across frames, the features from these distractors are propagated, which rather interferes with retrieving persons. The works of [24, 39, 49] propose to use recurrent neural networks (RNNs) for aggregating frame-level person representations robust to temporal distractors. The hidden states of RNNs store the temporal context in previous frames, and allow to aggregate the person representations selectively, based on the context. We also exploit RNNs in STMN, but we do not use them directly to aggregate frame-level representations, which may be suboptimal, since RNNs do not consider a temporal context in whole frames (except at the last time step). We instead leverage RNNs to encode a temporal pattern of a sequence for accessing a temporal memory.

Previous works overlook the fact that several scene details and temporal patterns repeatably appear in surveillance videos, which may provide important cues to handle spatial and temporal distractors. STMN stores the scene details and attentions for the temporal patterns in the spatial and temporal memories, respectively, providing person representations robust against the spatial and temporal distractors.

Memory network. The work of [37] first introduces memory networks to handle long-term dependencies for

question and answering. They, however, require extra supervisory signals to access the memory, and are not able to be trained end-to-end. The soft addressing technique [32] addresses these problems by using attention maps to access the memory. Key-value memory networks [25] propose to adopt different encodings for accessing and reading operations, where they address relevant memory items by keys, and their corresponding values are subsequently returned. Recently, many computer vision methods exploit memory networks for, *e.g.*, one-shot learning [1], video object segmentation [26], domain adaptation [48], image colorization [42], and anomaly detection [6, 27]. Our work also leverages memory networks but for recording features for distracting scene details and temporal attentions. By using the memory networks, we are able to extract person representations robust against spatial and temporal distractors. In addition, we propose a memory spread loss to penalize our model when it keeps accessing particular items only, while other items remain unused.

3. Approach

In this section, we provide a brief overview of our approach to exploiting spatial and temporal memories for video-based reID (Sec. 3.1). We then present a detailed description for a network architecture (Sec. 3.2) and training losses (Sec. 3.3).

3.1. Overview

STMN mainly consists of three components: an encoder, a spatial memory (Fig. 2), and a temporal memory (Fig. 3). For each frame, the encoder extracts a person representation and two query maps, where each query is used to access either spatial or temporal memories. The spatial memory stores features for scene details, frequently appearing across video frames, such as street lights, trees, and concrete pavers. We extract such features from the spatial memory using the corresponding query map, and use them to refine the person representation, removing information that interferes with identifying persons. The temporal memory saves attentions optimized for typical temporal patterns that repeatably occur in person videos. We access the temporal memory with the corresponding query map, and use the output to aggregate the refined frame-level features into a sequence-level person representation. We train our model end-to-end using memory spread, triplet, and cross-entropy terms.

3.2. Network architecture

Encoder. The encoder takes a video sequence $F_i|_{i=1}^{L}$ as an input, where F_i is the i-th frame of the sequence, and L is the total number of frames. We exploit ResNet [7] cropped at `conv4` layer as our backbone network, where the network parameters are pre-trained for ImageNet classification [14]. We add three heads on top of the backbone network to extract feature maps for each frame: a frame-level person representation $\mathbf{f}_i^{\text{o}}$, and query maps, $\mathbf{q}_i^{\text{s}}$ and $\mathbf{q}_i^{\text{t}}$, for accessing the spatial and temporal memories, respectively. Each feature map has a size of $D \times H \times W$, where D, H, and W are the number of channels, height, and width, respectively. We denote by $\mathbf{f}_{i,k}^{\text{o}}$, $\mathbf{q}_{i,k}^{\text{s}}$, and $\mathbf{q}_{i,k}^{\text{t}}$ individual features of size D at position k, where $k \in \{1, 2, ..., K\}$ and $K = H \times W$.

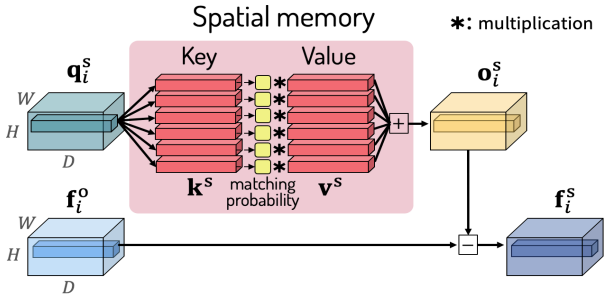

Figure 2. The spatial memory takes a person representation $\mathbf{f}_i^{\text{o}} \in \mathbb{R}^{D \times K}$ and a query map $\mathbf{q}_i^{\text{s}} \in \mathbb{R}^{D \times K}$ of the i-th frame as inputs. We access the memory based on the matching probability between the query feature $\mathbf{q}_{i,k}^{\text{s}} \in \mathbb{R}^{D}$ and keys $\mathbf{k}^{\text{s}}$, and use the output to refine the input representation $\mathbf{f}_{i,k}^{\text{o}} \in \mathbb{R}^{D}$. (Best viewed in color.)

Spatial memory. Frame-level person representations extracted by the encoder may contain features for distracting scene details (*e.g.*, trees, concrete paver, bicycles, or cars), which may prevent distinguishing different pedestrians in similar scenes. To handle this problem, we refine the frame-level person representations using a spatial memory (Fig. 2).

The spatial memory has a key-value structure, and contains M items. The values $\mathbf{v}^{\text{s}} \in \mathbb{R}^{D \times M}$ encode distracting scene details over the video sequence, while the keys $\mathbf{k}^{\text{s}} \in \mathbb{R}^{D \times M}$ are used to access corresponding values. We denote by $\mathbf{k}_n^{\text{s}} \in \mathbb{R}^{D}$ and $\mathbf{v}_n^{\text{s}} \in \mathbb{R}^{D}$ each key and value in the memory, respectively, where $n \in \{1, 2, ..., M\}$. The spatial memory takes a person representation $\mathbf{f}_i^{\text{o}} \in \mathbb{R}^{D \times K}$ and a query map $\mathbf{q}_i^{\text{s}} \in \mathbb{R}^{D \times K}$ of the frame F_i as inputs. Since different parts of the input frame may contain distinct scene details, we access the memory with individual components of the input query map, $\mathbf{q}_{i,k}^{\text{s}} \in \mathbb{R}^{D}$. Specifically, we compute cosine similarities between the query $\mathbf{q}_{i,k}^{\text{s}}$ and all keys $\mathbf{k}^{\text{s}}$ in the memory, resulting in a correlation map of size $1 \times M$. We then normalize it as follows:

$$a_{i,k,n}^{\text{s}} = \frac{\exp((\mathbf{q}_{i,k}^{\text{s}})^{\text{T}} \mathbf{k}_n^{\text{s}})}{\sum_{n'=1}^{M} \exp((\mathbf{q}_{i,k}^{\text{s}})^{\text{T}} \mathbf{k}_{n'}^{\text{s}})}. \quad (1)$$

The matching probability $a_{i,k,n}^{\text{s}}$ represents a likelihood that the scene detail recorded in the n-th memory item exists in the k-th position of the i-th frame. The memory outputs a weighted average of values $\mathbf{v}_n^{\text{s}}$ using the corresponding probabilities $a_{i,k,n}^{\text{s}}$ as follows:

$$\mathbf{o}_{i,k}^{\text{s}} = \sum_{n=1}^{M} a_{i,k,n}^{\text{s}} \mathbf{v}_n^{\text{s}}, \quad (2)$$

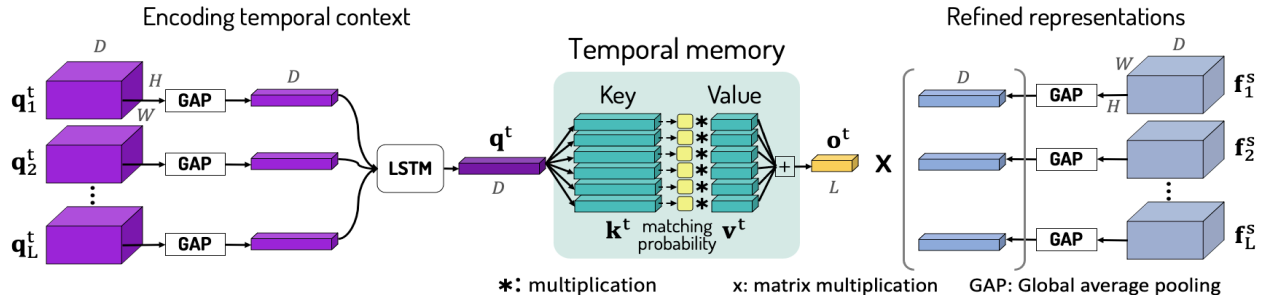

Figure 3. The temporal memory takes a sequence of query maps $\mathbf{q}_i^{\mathrm{t}}|_{i=1}^{L}$ and the person representations $\mathbf{f}_i^{\mathrm{s}}|_{i=1}^{L}$ that are refined by the spatial memory as inputs. We aggregate the query maps by using global average pooling and LSTM modules, and use the output to address the memory. The memory outputs temporal attentions $\mathbf{o}^{\mathrm{t}}$, and the attentions are used to aggregate the frame-level representations into a sequence-level one. (Best viewed in color.)

where the output of the spatial memory, $\mathbf{o}_{i,k}^{\mathrm{s}}$, contains uninformative features, that interfere with identifying persons, for the k-th position of the i-th frame. We use the output of the spatial memory to refine the person representation as follows:

$$\mathbf{f}_{i,k}^{\mathrm{s}} = \mathbf{f}_{i,k}^{\mathrm{o}} - \mathrm{BN}(\mathbf{o}_{i,k}^{\mathrm{s}}). \tag{3}$$

Motivated by [36], we use a batch normalization (BN) layer to adjust the distribution gap between outputs from the encoder and the spatial memory. Note that our spatial memory is similar to non-local networks [36] in that they both refine input features in a residual manner. However, ours is clearly different from the non-local networks. Keys and values in our method are external parameters stored in the memory, and they are updated by backpropagation during training in order to memorize the scene details. On the contrary, keys, queries, and values in the non-local networks are computed from input features, similar to a self-attention method [33].

Temporal memory. The refinement process using the spatial memory operates on each frame independently, which is not capable of capturing temporal contexts in video sequences. This may lead our framework susceptible to occlusion or misalignment between frames. To address this problem, we propose to use an additional temporal memory network (Fig. 3).

The temporal memory also has a key-value structure, and contains N items, where the keys $\mathbf{k}^{\mathrm{t}} \in \mathbb{R}^{D\times N}$ encode prototypes of temporal patterns that repeatably appear in person videos, and the values $\mathbf{v}^{\mathrm{t}} \in \mathbb{R}^{L\times N}$ memorize temporal attentions which are optimized for the corresponding temporal patterns. We denote by $\mathbf{k}_n^{\mathrm{t}} \in \mathbb{R}^{D}$ and $\mathbf{v}_n^{\mathrm{t}} \in \mathbb{R}^{L}$ each key and value in the memory, respectively, where $n \in \{1, 2, ..., N\}$. The temporal memory takes a sequence of query maps $\mathbf{q}_i^{\mathrm{t}}|_{i=1}^{L}$ and the person representations refined by the spatial memory $\mathbf{f}_i^{\mathrm{s}}|_{i=1}^{L}$ as inputs. We first encode a temporal context of a given sequence, *e.g.*, the occlusion arises in the middle frame, using the query maps. Concretely, we spatially aggregate the input query maps by a global average pooling (GAP), and feed them into a long short-term memory (LSTM) [10] as follows:

$$\mathbf{q}^{\mathrm{t}} = \mathrm{LSTM}([\mathrm{GAP}(\mathbf{q}_1^{\mathrm{t}}), \mathrm{GAP}(\mathbf{q}_2^{\mathrm{t}}), ..., \mathrm{GAP}(\mathbf{q}_L^{\mathrm{t}})]), \tag{4}$$

where $\mathbf{q}^{\mathrm{t}} \in \mathbb{R}^{D}$ is an output of the last time step, representing the temporal context of the sequence. We then use the temporal context $\mathbf{q}^{\mathrm{t}}$ to access the temporal memory in a similar way to the spatial one as follows:

$$a_n^{\mathrm{t}} = \frac{\exp((\mathbf{q}^{\mathrm{t}})^{\mathrm{T}}\, \mathbf{k}_n^{\mathrm{t}})}{\sum_{n'=1}^{N} \exp((\mathbf{q}^{\mathrm{t}})^{\mathrm{T}}\, \mathbf{k}_{n'}^{\mathrm{t}})}, \tag{5}$$

where a_n^{t} represents a probability that the encoded temporal context $\mathbf{q}^{\mathrm{t}}$ belongs to the temporal pattern stored in the n-th memory item $\mathbf{k}_n^{\mathrm{t}}$. We synthesize a temporal attention specific for the given sequence by taking weighted average over the values with the corresponding probability a_n^{t} as follows:

$$\mathbf{o}^{\mathrm{t}} = \sum_{n=1}^{N} a_n^{\mathrm{t}} \mathbf{v}_n^{\mathrm{t}}, \tag{6}$$

where the memory output $\mathbf{o}^{\mathrm{t}} \in \mathbb{R}^{L}$ represents the temporal attention, and o_i^{t}, the i-th element of the output, indicates the relative importance of the i-th frame in the sequence. We then apply a softmax function on the temporal attention $\mathbf{o}^{\mathrm{t}}$, and use it to aggregate the refined frame-level features $\mathbf{f}_i^{\mathrm{s}}$ as follows:

$$\mathbf{f}^{\mathrm{t}} = \sum_{i=1}^{L} \hat{o}_i^{\mathrm{t}}\, \mathrm{GAP}(\mathbf{f}_i^{\mathrm{s}}), \tag{7}$$

where $\hat{o}_i^{\mathrm{t}} = \exp(o_i^{\mathrm{t}}) / \sum_{i'=1}^{L} \exp(o_{i'}^{\mathrm{t}})$, and $\mathbf{f}^{\mathrm{t}}$ is our final person representation for the input video sequence $F_i|_{i=1}^{L}$.

Note that previous methods, *e.g.*, [49, 18, 5, 17, 16, 20, 30, 40], decide which frames to focus on during temporal fusion based on person representations. This may enforce the representations to encode temporal contexts as well as identity-related cues, preventing the representations from being discriminative, particularly when video sequences of different identities contain similar temporal contexts. In our framework, on the contrary, person representations are decoupled from encoding temporal contexts, where query

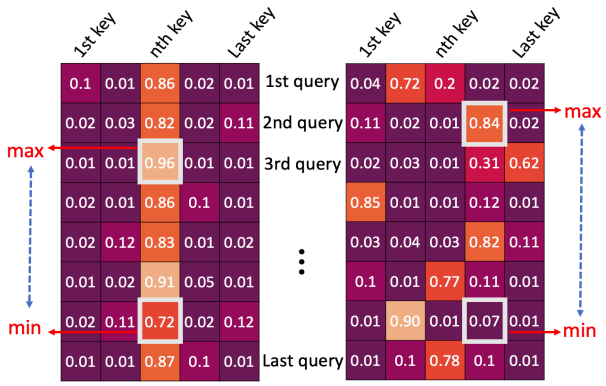

Figure 4. Example of matching probability maps for the case when our model addresses a particular memory item only (left) and the case when it uses all items in the memory (right). (Best viewed in color.)

maps $\mathbf{q}_i^{\mathrm{t}}$ and keys in the temporal memory, $\mathbf{k}^{\mathrm{t}}$, encode such contexts. This encourages our model to extract person representations focusing on information that is useful for discriminating different identities, leading to performance gains on the reID task.

3.3. Training loss

We use two terms to train our model end-to-end as follows:

$$\mathcal{L}_{total} = \mathcal{L}_S + \mathcal{L}_{ID}, \tag{8}$$

where we denote by $\mathcal{L}_S$ and $\mathcal{L}_{ID}$ memory spread and identification losses, respectively. The memory spread term penalizes our model when it accesses a particular memory item only, while the identification term allows to extract discriminative person representations from video sequences. The detailed descriptions of each loss are presented in the following.

Memory spread term. We denote by $\mathbf{A}^{\mathrm{s}} \in \mathbb{R}^{LKB \times M}$ and $\mathbf{A}^{\mathrm{t}} \in \mathbb{R}^{B \times N}$ matching probability maps for the spatial and temporal memories, respectively, in a mini-batch, where B is the number of sequences in the mini-batch. Note that we address the spatial and temporal memories LKB and B times for each mini-batch, respectively. Since we do not have extra supervisory signals except identification labels, we do not know which key should be matched to the input query. In this context, our model may address particular keys continually, while others are left unused (Fig. 4 left). This causes memories to produce similar outputs regardless of input frames or sequences. To address this problem, we propose a memory spread loss as follows:

$$\begin{aligned}\mathcal{L}_S = \sum_{n=1}^{M} &[\, min(\mathbf{a}_n^{\mathrm{s}}) - max(\mathbf{a}_n^{\mathrm{s}}) + \alpha \,]_+ \\ &+ [\, min(\mathbf{a}_n^{\mathrm{t}}) - max(\mathbf{a}_n^{\mathrm{t}}) + \alpha \,]_+,\end{aligned} \tag{9}$$

where $\mathbf{a}_n^{\mathrm{s}} \in \mathbb{R}^{LKB}$ and $\mathbf{a}_n^{\mathrm{t}} \in \mathbb{R}^{B}$ are the n-th column vector of $\mathbf{A}^{\mathrm{s}}$ and $\mathbf{A}^{\mathrm{t}}$, respectively, representing matching probabilities of the n-th key in each memory w.r.t all queries in a mini-batch. $min(\cdot)$ and $max(\cdot)$ return the minimum and maximum values of an input vector. The memory spread loss enforces the minimum and maximum values of $\mathbf{a}_n^{\mathrm{s}}$ and $\mathbf{a}_n^{\mathrm{t}}$ to differ by at least a pre-defined margin α. This prevents the case when our model keeps addressing a particular memory item (Fig. 4 left), while encouraging it to access all memory items during training (Fig. 4 right).

Identification term. Following other person reID methods [40, 43, 11, 3], we exploit a combination of cross-entropy and batch-hard triplet [8] terms, with identification labels as a supervisory signal. The former encourages our model to learn a person representation $\mathbf{f}^{\mathrm{t}}$ by focusing on identity-related cues, while the latter enforces the representations of the same identity to be closer to each other than those of different identities in the embedding space. Motivated by a deep supervision technique [15, 34], we also use the frame-level representations $\mathbf{f}_i^{\mathrm{s}}|_{i=1}^{L}$ to compute the cross-entropy and triplet losses, where global and temporal average pooling are used to aggregate frame-level representations into a sequence-level one.

4. Experiments

In this section, we provide implementation details of STMN (Sec. 4.1), and show ablation studies and visual analysis on spatial and temporal memories to validate the effectiveness of STMN (Sec. 4.2). Lastly, we compare our method with the state of the art (Sec. 4.3).

4.1. Implementation details

Dataset and evaluation metric. We evaluate our model on MARS [45], DukeMTMC-VideoReID [28, 38] (abbreviated as "DukeV"), and LS-VID [16], following the standard protocol of each dataset. Note that we do not use PRID [9] and iLIDS-VID [35] for evaluation, since they contain few sequences captured with two cameras only. We report cumulative matching characteristics at rank-1 and mean average precision (mAP) for quantitative comparisons.

Training. We train our model end-to-end for 200 epochs using the Adam [12] optimizer, where β_1 and β_2 are set to 0.9 and 0.999, respectively. The learning rate, initially set to 1e-4, is reduced by a factor of 10 for every 50 epochs. To train our model, we randomly choose 8 identities, and sample 4 sequences for each identity. Following the restricted random sampling (RRS) strategy [18], we then divide each sequence into L chunks, and randomly choose one frame from each chunk. We resize input frames into the size of 256×128, and augment them with horizontal flipping and random erasing [47].

Hyperparameter. To set the sizes of spatial and temporal memories, M and N, the pre-defined margin α in the memory spread loss, and the length of an input sequence L, we divide the training set of MARS [45] into two subsets. Specifically, we randomly divide identities in the training

Methods	MARS		DukeV		LS-VID	
	R-1	mAP	R-1	mAP	R-1	mAP
① Baseline	87.3	79.1	95.0	92.7	71.6	55.9
② + SM (w/o $\mathcal{L}_S$)	88.7	81.6	95.4	93.6	78.8	64.7
③ + SM	89.3	<u>82.5</u>	<u>96.2</u>	<u>94.2</u>	79.6	<u>65.8</u>
④ + TM (w/o $\mathcal{L}_S$)	88.5	81.9	95.2	93.3	77.8	63.0
⑤ + TM	<u>89.5</u>	82.0	95.4	93.7	78.9	64.4
⑥ + SM + TM (w/o $\mathcal{L}_S$)	89.1	81.9	95.6	93.9	<u>79.9</u>	65.4
⑦ + SM + TM	**89.9**	**83.7**	**96.7**	**94.6**	**80.6**	**66.6**

Table 1. Quantitative comparison for variants of our model on MARS [45], DukeV [38] and LS-VID [16]. Numbers in bold indicate the best performance and underscored ones are the second best. SM: spatial memory; TM: temporal memory.

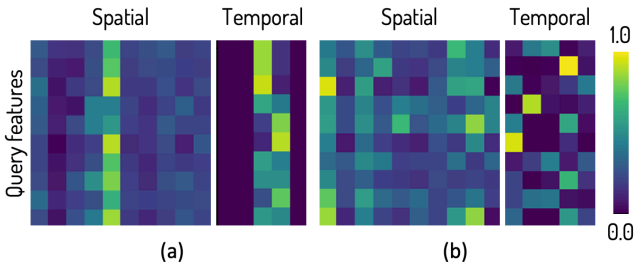

Figure 5. Matching probability maps of spatial and temporal memories, when they are trained (a) without and (b) with the memory spread loss. We randomly select 10 query features from a gallery set of MARS [45]. We can see that the memory spread loss encourages our model to access all items in the memories. (Best viewed in color.)

set into two subsets of sizes $500/125$, and use corresponding $7075/1223$ sequences as training/validation splits. For query sequences, we randomly select 200 sequences from the validation split. For the sizes of memories, we perform a grid search over (M, N) pairs, where $M, N \in \{5, 10, 20\}$. We choose a pair of $M = 10$ and $N = 5$ for our final model, which shows the best result in terms of the mean and standard deviation of rank-1 accuracy and mAP for five trials. For the margin α and the sequence length L, we also use a grid search over $\alpha \in \{0.1, 0.3, 0.5, 0.7, 1.0\}$ and $L \in \{4, 6, 8, 10\}$, respectively, setting $\alpha = 0.3$ and $L = 6$. We fix all hyperparameters, and train our model on the training splits of MARS [45], DukeV [38] and LS-VID [16]. Please refer to the supplementary materials for the details.

4.2. Discussion

Ablation study. We show in Table 1 an ablation study of our model on MARS [45], DukeV [38], and LS-VID [16] in terms of rank-1 accuracy(%) and mAP(%). For the baseline, we use the same network architecture as the encoder, while removing two heads for query maps, and exploit global and temporal average pooling to aggregate person representations. From ① and ③, we can clearly see that the feature refinement process using a spatial memory boosts the reID performance, while ① and ⑤ demonstrate that using a temporal memory for aggregating frame-level representations gives better results. ③, ⑤, and ⑦ further show

Figure 6. Top-5 retrieved frames from a gallery set of MARS [45], whose query features have high matching probabilities with a key of the spatial memory.

Figure 7. The magnitude difference of person representations before and after the refinement using the spatial memory.

the spatial and temporal memories are complementary to each other. Note that LS-VID provides person videos with more diverse spatial and temporal distractors than the other datasets. It contains videos of three times larger number of identities than MARS and DukeV, which are captured under two times larger number of cameras. Our memories help the baseline model to handle such distractors, giving the significant performance gains on LS-VID. The performance gains by the memories are relatively small on DukeV, since it contains person videos that are manually annotated by humans, *i.e.*, with less distractors, where the simple baseline already gives 95% rank-1 accuracy. By comparing ② to ③, ④ to ⑤, and ⑥ to ⑦, we can see that enforcing our model to address all memory items during training by a memory spread loss consistently enhances the performance.

To further verify the effectiveness of the memory spread loss, we visualize matching probability maps of spatial and temporal memories on MARS, when the memories are trained without (Fig. 5(a)) and with (Fig. 5(b)) the loss. We randomly choose frames or sequences from a gallery set of MARS, and extract query features, $\mathbf{q}^{\mathrm{s}}_{i,k}$ and $\mathbf{q}^{\mathrm{t}}$, from them. We then compute matching probabilities with keys of the spatial and temporal memories using Eq. (1) and Eq. (5), respectively. We can see that the memory spread loss encourages our model to leverage all items in the memories, while preventing it from accessing particular items only. This enables our spatial and temporal memories to produce diverse outputs depending on both frame-level scene details and sequence-level temporal contexts.

Spatial memory. In Fig. 6, we visualize video frames whose query features $\mathbf{q}^{\mathrm{s}}_{i,k}$ have high matching probabilities with randomly chosen keys from the spatial memory (see Eq. (1)). We can observe that each key retrieves the video frames that share similar scene details such as a play field (1st row), a street light (2nd row), or concrete pavers (3rd row). This verifies that our model accesses the spatial memory depending on the scene details for each video frame. The spatial memory aggregates the features

Figure 8. Comparison of top-10 retrieval results on the test split of MARS [45] using the original frame-level features $\mathbf{f}_i^{\mathrm{o}}$ (top) and refined ones $\mathbf{f}_i^{\mathrm{s}}$ (bottom). Results with green boxes have the same identity as the query, while those with red boxes do not. We show the first frame of sequences for the purpose of visualization. (Best viewed in color.)

Figure 9. Examples of sequences from a gallery set of MARS [45], whose query features show high matching probabilities with a particular key in the temporal memory. We also visualize temporal attentions stored in corresponding values of the memory. (Best viewed in color.)

for scene details, and we use them to refine frame-level person representations (see Eq. (3)). To see the effect of the refinement, we visualize in Fig. 7 the magnitude difference of person representations, overlaid on input images, using bilinear interpolation, before and after the refinement, *i.e.*, $\left\{\|\mathbf{f}_{i,k}^{\mathrm{s}}\|_2 - \|\mathbf{f}_{i,k}^{\mathrm{o}}\|_2 \middle| k \in H \times W\right\}$. We can observe that the differences mainly occur from distracting scene details, *e.g.*, concrete pavers, playfield, or street lights, implying that the memory suppresses features from them. Note that the video frames in the 1st row of Fig. 7 share the same background while pedestrians appear in different positions. However, regardless of the person's position, the memory removes features from background clutter. Figure 8 compares retrieval results when we use initial person representations $\mathbf{f}_i^{\mathrm{o}}$ (top) and refined ones $\mathbf{f}_i^{\mathrm{s}}$ (bottom). Note that we use global and temporal average pooling to obtain the person representations, instead of exploiting the temporal memory, to see the effect of the refinement by the spatial memory. We can see that the initial representations retrieve person sequences of different identities from the query but with similar scene details (*e.g.*, a play field). On the other hand, the refined ones retrieve person sequences with the same identity as the query correctly, regardless of background clutter in each frame. This also suggests that the refinement process using the spatial memory suppresses information of the scene details in person representations.

Temporal memory. We visualize in Fig. 9 person sequences whose query features $\mathbf{q}^{\mathrm{t}}$ show high matching probabilities (see Eq. (5)) with randomly chosen keys from the temporal memory. We also visualize the corresponding values of the memory in the below. We can observe that each key retrieves the sequences with similar temporal patterns, *e.g.*, persons disappear at the end of the sequence (left) or appear in all frames with similar appearances (right), and the values highlight discriminative frames in each sequence. This verifies that the keys encode prototypes of temporal patterns in person videos, and the values of the memory store temporal attentions which are optimized for the corresponding temporal patterns. Note that we aggregate individual values of the memory with matching probabilities between keys and input query features to synthesize temporal attentions specific for input person sequences (see Eq. (6)). Figure 10 shows examples of the aggregated temporal attentions. When the temporal memory takes a sequence with less temporal distractors as an input, the memory generates similar attentions for all frames (1st row). Namely, the memory works similarly to the temporal average pooling which fuses video frames with equal probabilities. On the other hand, in case of a sequence with severe temporal distractors, *e.g.*, misalignments between frames (2nd row) or occlusions (3rd row), the memory lowers attentions for the frames where such variations occur, suggesting that the temporal memory allows our model to extract person representations robust to the temporal variations. Note that we can replace the temporal memory with multilayer perceptrons (MLPs) by directly regressing attentions from the encoded context $\mathbf{q}^{\mathrm{t}}$ [18]. To compare this approach with ours, we use two-layer perceptrons whose sizes are $2048 \times N$ and $N \times L$, respectively, which makes the number of parameters the same as ours. We found that MLP often produce attentions that focus more on few certain frames, ignoring features from the other frames (see the last row of Fig. 10), and this leads to large performance drops, 1.3/1.5 (R-1/mAP) on MARS. The result are similar, even the size of MLPs increase (*e.g.*, 2048x512 and 512x6). These show the effectiveness of our approach that predicts

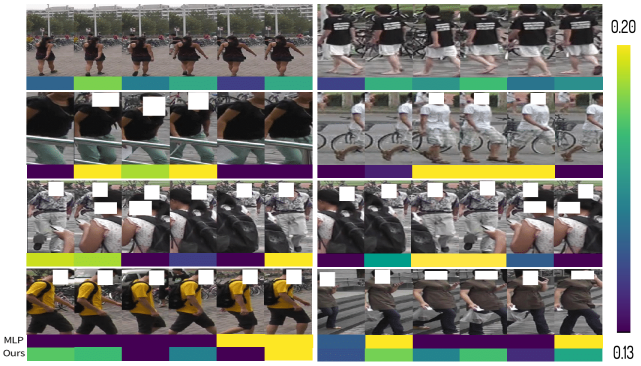

Figure 10. Examples of temporal attentions generated by the temporal memory on the test split of MARS [45]. Note that the sequence on the right side of the 3rd row is made by reordering the sequence on the left side. (Best viewed in color.)

attentions by discovering repetitive temporal patterns in a dataset and searching the most relevant patterns to the context of an input video.

4.3. Comparison with the state of the art

We compare in Table 2 STMN with the state of the art in terms of rank-1 accuracy and mAP on MARS [45], DukeV [38] and LS-VID [16]. We found that previous methods compare their performance using different test strategies. For fair comparisons, we classify them into two groups, depending on whether they follow RRS or all-frames strategies for evaluation. The methods, *e.g.*, [18, 20], which follow the RRS strategy [18], divide an input video into L chunks of equal length. They then sample the first frame of each chunk to obtain a sequence of L frames, regardless of the total number of frames. On the other hand, several works use all frames in an input video by grouping them into multiple sequences of length L. They extract a person representation from each sequence independently, and average all the representations to represent the input video. Note that we reproduce TCLNet [11] and MGH [40] to evaluate them on the both strategies. Using all frames in given videos to extract person representations does give performance gains for TCLNet, MGH, and STMN. This, however, is far from practical usages in that it runs, *e.g.*, 35 times slower than the RRS strategy on LS-VID, requiring more than three hours for evaluation using a Titan RTX 2080Ti GPU. Furthermore, the time for searching persons increases linearly as the number of video frames increases.

From Table 2, we have following observations: 1) On the RRS setting, STMN sets a new state of the art on the three benchmarks. The results of STMN using the RRS even surpass those of previous methods, *e.g.*, COSAM [30], M3D [17], and GLTR [16] on the all-frames setting. This suggests that STMN already extracts essential information for identifying a person with sampled frames only, showing its efficiency over the previous methods. This characteristic is crucial for massive surveillance systems which need to search for a person of interest from lots of videos in a very short time; 2) DRSA [18] leverages attention modules for handling spatial and temporal distractors in videos, while STMN exploits spatial and temporal memories instead. The performance gap between these two methods demonstrates the superiority of our framework over the attention-based method; 3) Co-attention-based methods [20, 16, 40] may propagate non-discriminative features across frames when multiple frames share common background clutter or occlusion. As a result, there are large performance gaps between these methods and STMN on LS-VID, the most challenging dataset, which contains sequences captured under various conditions (*e.g.*, lighting/background changes, indoor/outdoor changes) with frequent occlusions; 4) TCLNet [11] and MGH [40] are the most recently introduced video reID methods. They boost the reID performance using a temporal saliency erasing module and a multi-granular hypergraph, respectively. They, however, give results worse than STMN on the RRS setting. By using all frames, they may show comparable results to STMN, however note that the size of a person representation is much larger than that of STMN (TCLNet:$4,096$, MGH:$5,120$ vs. STMN:$2,048$).

	Methods	MARS		DukeV		LS-VID	
		rank-1	mAP	rank-1	mAP	rank-1	mAP
RRS	EUG [38]	62.7	42.5	72.8	63.2	-	-
	SeeForest [49]	70.6	50.7	-	-	-	-
	QAN [22]	73.7	51.7	-	-	-	-
	DRSA [18]	82.3	65.8	-	-	-	-
	CSA [2]	86.3	76.1	-	-	-	-
	STE-NVAN [20]	88.9	81.2	95.2	<u>93.5</u>	(72.1)	(56.6)
	TCLNet [11]	(88.5)	(80.9)	(95.0)	(92.8)	(75.0)	<u>(60.2)</u>
	MGH [40]	<u>(89.2)</u>	<u>(83.4)</u>	<u>(95.3)</u>	(93.4)	<u>(75.3)</u>	(58.9)
	STMN	**89.9**	**83.7**	**96.7**	**94.6**	**80.6**	**66.6**
All frames	COSAM [30]	83.7	77.2	94.4	94.0	-	-
	STMP [23]	84.4	72.7	-	-	56.8	39.1
	M3D [17]	84.4	74.1	-	-	57.7	40.1
	Part-Aligned [31]	84.7	75.9	-	-	-	-
	STA [5]	86.3	80.8	96.0	95.0	-	-
	GLTR [16]	87.0	78.5	96.3	93.7	63.1	44.3
	TCLNet [11]	(89.1)	(83.4)	<u>(96.7)</u>	<u>(95.6)</u>	<u>(81.0)</u>	<u>(67.2)</u>
	MGH [40]	<u>(89.4)</u>	(**85.3**)	(95.0)	(94.6)	(79.6)	(61.8)
	STMN	**90.5**	<u>84.5</u>	**97.0**	**95.9**	**82.1**	**69.2**

Table 2. Comparison with the state of the art on MARS [45], DukeV [38], and LS-VID [16] in terms of rank-1 accuracy(%) and mAP(%). Numbers in bold indicate the best performance and underscored ones are the second best. Results in brackets are obtained with the source codes provided by the authors.

5. Conclusion

We have presented a novel video-based person reID method, dubbed STMN, that extracts robust person representations against spatial and temporal distractors in videos. To this end, we have proposed to exploit two external memory networks, spatial and temporal memories, to refine frame-level representations and to aggregate them into a sequence one, focusing on discriminative frames. We have also proposed a memory spread loss that prevents certain memory items from remaining redundant. We have shown that STMN achieves state-of-the-art performance on standard video-based reID benchmarks, and demonstrated the effectiveness of each component of our method with an extensive ablation study.

Acknowledgments This research was supported in part by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (NRF-2019R1A2C2084816) and the Yonsei University Research Fund of 2021 (2021-22-0001).

References

[1] Qi Cai, Yingwei Pan, Ting Yao, Chenggang Yan, and Tao Mei. Memory matching networks for one-shot image recognition. In *CVPR*, 2018. 3

[2] 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. 8

[3] Guangyi Chen, Yongming Rao, Jiwen Lu, and Jie Zhou. Temporal coherence or temporal motion: Which is more critical for video-based person re-identification? In *ECCV*, 2020. 5

[4] Chanho Eom and Bumsub Ham. Learning disentangled representation for robust person re-identification. In *NeurIPS*, 2019. 1

[5] Yang Fu, Xiaoyang Wang, Yunchao Wei, and Thomas Huang. STA: Spatial-temporal attention for large-scale video-based person re-identification. In *AAAI*, 2019. 1, 2, 4, 8

[6] Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *ICCV*, 2019. 3

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3

[8] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv:1703.07737*, 2017. 5

[9] Martin Hirzer, Csaba Beleznai, Peter M Roth, and Horst Bischof. Person re-identification by descriptive and discriminative classification. In *SCIA*, 2011. 5

[10] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*. 4

[11] Ruibing Hou, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Temporal complementary learning for video person re-identification. In *ECCV*, 2020. 1, 2, 5, 8

[12] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 5

[13] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *ICLR*, 2017. 1, 2

[14] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet classification with deep convolutional neural networks. In *NIPS*, 2012. 3

[15] Chen-Yu Lee, Saining Xie, Patrick Gallagher, Zhengyou Zhang, and Zhuowen Tu. Deeply-supervised nets. In *Artificial Intelligence and Statistics*, 2015. 5

[16] Jianing Li, Jingdong Wang, Qi Tian, Wen Gao, and Shiliang Zhang. Global-local temporal representations for video person re-identification. In *ICCV*, 2019. 1, 2, 4, 5, 6, 8

[17] Jianing Li, Shiliang Zhang, and Tiejun Huang. Multi-scale 3D convolution network for video based person re-identification. In *AAAI*, 2019. 1, 2, 4, 8

[18] Shuang Li, Slawomir Bak, Peter Carr, and Xiaogang Wang. Diversity regularized spatiotemporal attention for video-based person re-identification. In *CVPR*, 2018. 1, 2, 4, 5, 7, 8

[19] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *CVPR*, 2018. 1

[20] Chih-Ting Liu, Chih-Wei Wu, Yu-Chiang Frank Wang, and Shao-Yi Chien. Spatially and temporally efficient non-local attention network for video-based person re-identification. In *BMVC*, 2019. 1, 2, 4, 8

[21] Xihui Liu, Haiyu Zhao, Maoqing Tian, Lu Sheng, Jing Shao, Shuai Yi, Junjie Yan, and Xiaogang Wang. Hydraplus-net: Attentive deep features for pedestrian analysis. In *ICCV*, 2017. 1

[22] Yu Liu, Junjie Yan, and Wanli Ouyang. Quality aware network for set to set recognition. In *CVPR*, 2017. 2, 8

[23] Yiheng Liu, Zhenxun Yuan, Wengang Zhou, and Houqiang Li. Spatial and temporal mutual promotion for video-based person re-identification. In *AAAI*, 2019. 8

[24] Niall McLaughlin, Jesus Martinez del Rincon, and Paul Miller. Recurrent convolutional network for video-based person re-identification. In *CVPR*, 2016. 2

[25] Alexander Miller, Adam Fisch, Jesse Dodge, Amir-Hossein Karimi, Antoine Bordes, and Jason Weston. Key-value memory networks for directly reading documents. In *EMNLP*, 2016. 3

[26] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *ICCV*, 2019. 3

[27] Hyunjong Park, Jongyoun Noh, and Bumsub Ham. Learning memory-guided normality for anomaly detection. In *CVPR*, 2020. 3

[28] 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. 5

[29] 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. 1

[30] Arulkumar Subramaniam, Athira Nambiar, and Anurag Mittal. Co-segmentation inspired attention networks for video-based person re-identification. In *ICCV*, 2019. 1, 2, 4, 8

[31] Yumin Suh, Jingdong Wang, Siyu Tang, Tao Mei, and Kyoung Mu Lee. Part-aligned bilinear representations for person re-identification. In *ECCV*, 2018. 8

[32] Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. End-to-end memory networks. In *NeurIPS*, 2015. 3

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 4

[34] Liwei Wang, Chen-Yu Lee, Zhuowen Tu, and Svetlana Lazebnik. Training deeper convolutional networks with deep supervision. *arXiv:1505.02496*, 2015. 5

[35] Taiqing Wang, Shaogang Gong, Xiatian Zhu, and Shengjin Wang. Person re-identification by video ranking. In *ECCV*, 2014. 5

[36] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, 2018. 1, 2, 4

[37] Jason Weston, Sumit Chopra, and Antoine Bordes. Memory networks. In *ICLR*, 2015. 2

[38] 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. 2, 5, 6, 8

[39] Yichao Yan, Bingbing Ni, Zhichao Song, Chao Ma, Yan Yan, and Xiaokang Yang. Person re-identification via recurrent feature aggregation. In *ECCV*, 2016. 2

[40] Yichao Yan, Jie Qin, Jiaxin Chen, Li Liu, Fan Zhu, Ying Tai, and Ling Shao. Learning multi-granular hypergraphs for video-based person re-identification. In *CVPR*, 2020. 1, 2, 4, 5, 8

[41] Jinrui Yang, Wei-Shi Zheng, Qize Yang, Ying-Cong Chen, and Qi Tian. Spatial-temporal graph convolutional network for video-based person re-identification. In *CVPR*, 2020. 1, 2

[42] Seungjoo Yoo, Hyojin Bahng, Sunghyo Chung, Junsoo Lee, Jaehyuk Chang, and Jaegul Choo. Coloring with limited data: Few-shot colorization via memory augmented networks. In *CVPR*, 2019. 3

[43] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. Multi-granularity reference-aided attentive feature aggregation for video-based person re-identification. In *CVPR*, 2020. 1, 2, 5

[44] Haiyu Zhao, Maoqing Tian, Shuyang Sun, Jing Shao, Junjie Yan, Shuai Yi, Xiaogang Wang, and Xiaoou Tang. SpindleNet: Person re-identification with human body region guided feature decomposition and fusion. In *CVPR*, 2017. 1

[45] 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*. 2, 5, 6, 7, 8

[46] Zhedong Zheng, Xiaodong Yang, Zhiding Yu, Liang Zheng, Yi Yang, and Jan Kautz. Joint discriminative and generative learning for person re-identification. In *CVPR*, 2019. 1

[47] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. *arXiv:1708.04896*, 2017. 5

[48] Zhun Zhong, Liang Zheng, Zhiming Luo, Shaozi Li, and Yi Yang. Invariance matters: Exemplar memory for domain adaptive person re-identification. In *CVPR*, 2019. 3

[49] 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. 2, 4, 8