Temporal Complementary Learning for Video Person Re-Identification

Ruibing Hou[1,2], Hong Chang[1,2], Bingpeng Ma[2], Shiguang Shan[1,2,3], and Xilin Chen[1,2]

[1] Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing 100190, China
[2] University of Chinese Academy of Sciences, Beijing 100049, China
[3] CAS Center for Excellence in Brain Science and Intelligence Technology, China
ruibing.hou@vipl.ict.ac.cn, {changhong, sgshan,xlchen}@ict.ac.cn,
bpma@ucas.ac.cn

Abstract. This paper proposes a *Temporal Complementary Learning Network* that extracts complementary features of consecutive video frames for video person re-identification. Firstly, we introduce a *Temporal Saliency Erasing* (TSE) module including a saliency erasing operation and a series of ordered learners. Specifically, for a specific frame of a video, the saliency erasing operation drives the specific learner to mine new and complementary parts by erasing the parts activated by previous frames. Such that the diverse visual features can be discovered for consecutive frames and finally form an integral characteristic of the target identity. Furthermore, a *Temporal Saliency Boosting* (TSB) module is designed to propagate the salient information among video frames to enhance the salient feature. It is complementary to TSE by effectively alleviating the information loss caused by the erasing operation of TSE. Extensive experiments show our method performs favorably against state-of-the-arts. The source code is available at https://github.com/blue-blue272/VideoReID-TCLNet.

Keywords: Video Person Re-identification, Complementary Learning, Feature Enhancing

1 Introduction

Person re-identification (reID) aims at retrieving particular persons from non-overlapping camera views. It plays a significant role in video surveillance analysis. Image person reID has achieved great progress in term of the methods [35, 12, 45] and large benchmarks construction [48, 50, 39]. Recently, with the emergence of large video benchmarks [47, 40] and the growth of computational resource, video person reID has been attracting a significant amount of attention. The video data contain richer spatial appearance information and temporal cues, which can be exploited for more robust reID.

Current state-of-the-art approaches for video person reID are based on deep neural networks. The mainstream approaches usually consider a sequence of

(a) The pair of Input video sequences

(b) The activation maps of existing methods
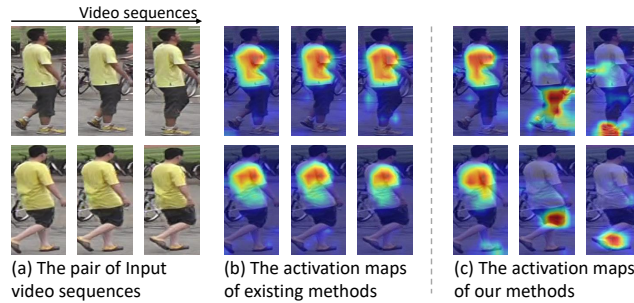
(c) The activation maps of our methods

Fig. 1: An example of class activation maps [52] of a pair of input video sequences of existing method [26] and our method. Warmer color with higher value

frames as input, and utilize convolutional neural networks (CNNs) to extract the feature for each frame independently, followed by temporal feature aggregation, *e.g.*, through recurrent layer [28, 41, 53] or temporal attention layer [26, 32, 41]. With the powerful deep networks and large-scale labeled benchmarks, these methods achieve favorable performance and efficiency.

Despite the significant progress in video person reID, most existing methods do not take full advantage of the rich spatial-temporal clues in the video. To be specific, since the pedestrian frames of a video are highly similar and the existing methods perform the same operation on each frame, these methods typically produce highly redundant features for the frames of a video. The redundant features typically attend to the same local salient parts [11], which are difficult to distinguish the persons with similar appearance. For example, as illustrated in Fig. 1(b), the upper clothes of the sequence pair attract the most attention, but are difficult to distinguish the two pedestrians. Therefore, it is appealing to explore a way of fully mining the spatial-temporal clues in the video, which can discover diverse visual cues for different frames of a video to form a full characteristic of each identify.

In this paper, we propose a *Temporal Complementary Learning Network* (TCLNet) for fully exploiting the spatial-temporal information of the video data. Firstly, we introduce a *Temporal Saliency Erasing* (TSE) module consisting of a saliency erasing operation and a series of ordered adversary learners. The key idea is to extract complementary features for consecutive frames of a video by the adversary learners. In particular, the first learner is firstly leveraged to extract the most salient feature for the first frame of a video sequence. Then, for the feature map of the second frame, the *saliency erasing operation* utilizes the *temporal cues* to erase the region attended by the first learner. Then we feed the feature excluding the erased region into the second learner for discovering new and complementary parts. By recursively erasing all previous discovered parts, the ordered learners can mine complementary parts for the consecutive frames and finally obtain an integral characteristic of the target person. As shown in

Fig. 1(c), with TSE, the features of consecutive frames can focus on diverse parts, covering the whole body of the target identity.

However, as TSE recursively erases the most salient part of the second and subsequent frames of the input video sequence, the representation of the most salient part is less powerful. To this end, we propose a *Temporal Saliency Boosting* (TSB) module to enhance the representational power of the most salient part. Concretely, TSB utilizes the *temporal cues* to propagate the most salient information among the video frames. In this way, the most salient features can capture the visual cues across all frames of video, hence exhibit stronger representational capability.

The proposed two modules can be inserted in any deep CNNs to extract complementary features for video frames. To the best of our knowledge, it is the first attempt to extract complementary features for the consecutive frames of a video. We demonstrate the effectiveness of the proposed method on three challenging video reID benchmarks, and our method outperforms the state-of-the-art methods under multiple evaluation metrics.

2   Related Work

Person Re-identification. Person reID for still images has been extensively studied [35, 5, 45, 49, 2, 23, 12, 7]. Recently, researchers start to pay attention to video reID [26, 21, 41, 13, 8]. The existing methods can be divided into two categories, *i.e.*, image-set based methods and temporal-sequence based methods.

Image-set based methods consider a video as a set of disordered images. These methods usually extract the features of each frame independently, then use a specific temporal pooling strategy to aggregate the frame-level features. For example, Zheng *et al.* [47] apply an average pooling across all frames to obtain the video feature. The works [26, 32, 21, 46] further use a temporal attention mechanism that assigns a quality score to each frame for weighted average pooling. These methods exhibit promising efficiency, but totally ignore the temporal cues of the video data.

Temporal-sequence based methods exploit the temporal cues for video representation learning. The early works [28, 41, 53] use the optical flows to encode the short-term motion information among adjacent frames. Mclaughlin *et al.* [38] propose a recurrent architecture [14] to aggregate the frame-level representations and yield a sequence-level feature representations. Zhang *et al.* [44] argue that the recurrent structure may not be optimal to learn temporal dependencies and propose to learn orderless ensemble ranking. Liao *et al.* [22] propose to use 3D convolution for spatial-temporal feature learning. Recently, some works [22, 13, 19] apply non-local blocks [37] to capture long-term temporal cues.

Nevertheless, these methods perform the same operation on each frame, leading to that the features of different frames are highly redundant. Therefore in this work, we propose a temporal complementary learning mechanism to extract complementary features for consecutive frames of a video, which is able to obtain an integral characteristic of the target identity for better reID.

Erasing Pixels or Activations. *Image based erasing* has been widely applied as a data augmentation technique. For instance, the works [51, 4] randomly erase a rectangle region of the input images during training. Singh *et al.* [31] propose to divide the image into a grid with fixed patch size and randomly mask each patch. *Feature based erasing* typically drops the feature activations. Dropout [33] drops the feature units randomly, which is a widely used regularization technique to prevent overfitting. DropBlock [6] randomly drops a contiguous region of the convolutional features for CNNs. In the work [3], all feature maps in the same batch are dropped in a consistent way for better metric learning.

Our work is fundamentally different from the existing erasing methods in two folds. Firstly, **the basic idea is different**. The above methods typically use the erasing strategy to regularize the *training* of networks to prevents overfitting. Differently, our method uses the erasing operation to extract complementary features for video frames during both training and testing phases. Secondly, **the erasing mechanism is different.** The existing methods usually randomly erase the pixels or activations without any high-level guidance. Our method erases the regions for each frame guided by the activated parts of previous frames, which guarantees the frame pays attention to new person parts thus is more efficient.

Recently, Liu *et al.* [24, 25] also propose to use saliency to guide erasing to learn complementary features. However, the purpose and implementation of these methods are different from ours. Specifically, these methods perform *image erasing* on *a singe image input*. On the contrary, our method performs *feature erasing* on *a video sequence* that allows for extracting complementary features for consecutive video frames.

3   Temporal Complementary Learning Network

The proposed Temporal Complementary Learning Network for video person reID includes two novel components, *i.e.* TSE for complementary feature mining and TSB for enhancing the most salient feature.

3.1   Temporal Saliency Erasing Module

Compared to images, video data contain richer spatial-temporal information, which should be fully exploited for more robust feature representation. However, most existing methods perform the same operation on each frame, resulting in highly redundant features of different frames that only highlight a local part [49, 42], as shown in Fig. 1 (b). To this end, we design a *temporal saliency erasing module* to mine complementary parts from consecutive frames of a video to form an integral characteristic of the target person.

TSE Overview. As shown in Fig. 2 (a), TSE iteratively performs two operations on each frame: adversarially erasing the parts discovered by previous frames with a *saliency erasing operation* (SEO) and learning a specific *learner* for discovering new part to extract complementary feature. Concretely, the input of TSE is the set of frame-level feature maps $\{F_n\}_{n=1}^N (F_n \in \mathbb{R}^{H \times W \times D})$ of
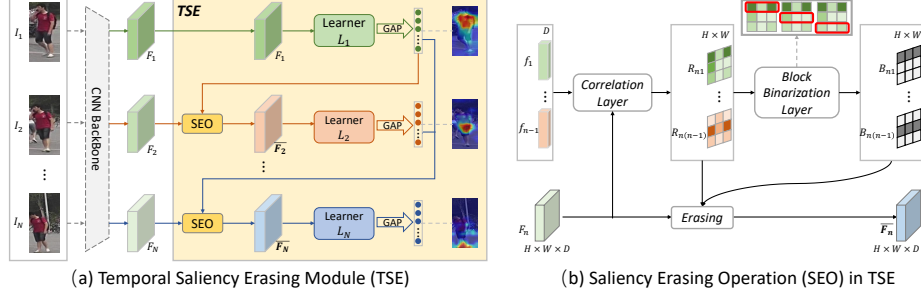
(a) Temporal Saliency Erasing Module (TSE)   (b) Saliency Erasing Operation (SEO) in TSE

Fig. 2: (a) The architecture of TSE. (b) *Saliency erasing operation* in TSE for feature map F_n $(n > 1)$

a video segment $\{I_n\}_{n=1}^N$ after a CNN Backbone, where the video segment contains N consecutive frames and n is the index of the video frame, and H, W and D denote the height, width and channel number of the feature map respectively. Firstly, TSE uses the learner L_1 followed by a GAP (global average pooling) layer to extract the most salient feature $f_1 \in \mathbb{R}^{D_1}$ for I_1, which is denoted as:

$$f_1 = \mathrm{GAP}(L_1(F_1)). \tag{1}$$

Then, SEO erases the feature map F_2 of frame I_2 guided by the mined discriminative part of f_1. The erased feature map is then fed into learner L_2. As the part attended by learner L_1 has been removed, the learner L_2 is naturally driven to discover new discriminative parts for identifying the target person. Recursively, for the frame feature map F_n $(n > 1)$, TSE firstly applies SEO to erases all parts discovered by previous frames to form the erased feature map $\overline{F_n}$, and then uses its specific learner L_n to mine new parts and obtain the feature vector f_n, which can be formulated as:

$$\overline{F_n} = \mathrm{SEO}(F_n; f_1, \ldots, f_{n-1}), \quad f_n = \mathrm{GAP}(L_n(\overline{F_n})) \ (\ 1 < n \leq N). \tag{2}$$

The *saliency erasing operation* and *learners* repeatedly perform on the N consecutive frames of the input segment. Finally, the integral characteristic of target person can be obtained by combining the features produced by these frames.

Correlation Layer of SEO. The implementation of SEO (Eq. 2) is illustrated in Fig. 2 (b). For the feature map F_n of frame I_n to be erased, a *correlation layer* is firstly designed to obtain the correlation maps between previous-frame feature vectors $f_k(k < n)$ and F_n. In particular, we firstly consider the feature vector at every spatial location (i, j) of F_n as a D dimensional local descriptor $F_n^{(i,j)}$. Then the *correlation layer* computes the semantic relevance between f_k and all the local descriptors of F_n with dot-produce similarity [37] to get the corresponding correlation map $R_{nk} \in \mathbb{R}^{H \times W}$ as:

$$R_{nk}^{(i,j)} = (F_n^{(i,j)})^T \left(w^T f_k \right) \quad (1 \leq i \leq H, 1 \leq j \leq W, 1 \leq k \leq n-1). \tag{3}$$

Here $w \in \mathbb{R}^{D_1 \times D}$ projects f_k to the feature space of F_n, matching the number of channels to that of F_n. Eq. 3 shows that the local descriptors that describe the part activated by f_k tend to present higher relevance values in R_{nk}. Thus R_{nk} can localize the regions in F_n of the parts activated by the previous-frame feature vector f_k.

Block Binarization Layer of SEO. The correlation maps are then used to generate the binary masks to identify the regions to be erased. A valiant approach is to conduct a threshold on the correlation maps. However, it usually produces noncontinuous regions. As pointed by [6], since the convolutional feature units are correlated spatially, when erasing the feature units discontinuously, information about the erased units can still be transmitted to the next layer. To this end, we design a *block binarization layer* to generate the binary mask which can erase a contiguous region of a feature map. As shown in Fig. 2 (b), we search the most highlighted continuous area in the correlation map using a *sliding block*. Formally, for a correlation map of size $H \times W$ and the sliding block of size $h_e \times w_e$, when we move the block with horizontal and vertical strides s_w and s_h respectively, the total number of block position can be computed as $N_{pos} = \left(\lfloor \frac{H-h_e}{s_h} \rfloor + 1 \right) \times \left(\lfloor \frac{W-w_e}{s_w} \rfloor + 1 \right)$. Thus N_{pos} candidate blocks can be obtained for each correlation map. We then define the correlation value of a block as the sum of the correlation values of the items in the block. Finally, we select the candidate block with the highest correlation value as the block to be erased, *i.e.*, the binary mask $B_{nk} \in \mathbb{R}^{H \times W}$ of correlation map R_{nk} is generated by setting the values of the units in the selected block to 0 and others to 1. We then merge the masks $\{B_{nk}\}_{k=1}^{n-1}$ to a fused mask B_n for the feature map F_n, which is calculated as:

$$B_n = B_{n1} \odot B_{n2} \odot \cdots \odot B_{n(n-1)}, \tag{4}$$

where $\odot$ is element-wise product operation.

Erasing Operation in SEO. In order to make TSE end-to-end trainable, we employ a gate mechanism to erase the feature map F_n. In particular, we apply a softmax layer to the fused correlation map, from which we erase the selected block using B_n to obtain a gate map $G_n \in \mathbb{R}^{H \times W}$:

$$G_n = \text{softmax} \left(R_{n1} \odot R_{n2} \odot \cdots \odot R_{n(n-1)} \right) \odot B_n. \tag{5}$$

F_n is erased based on B_n and G_n to generate the erased feature map $\overline{F_n}$. For consistency, we also apply the erasing operation on F_1 with a binary mask filling with 1. In this way, the gradients can propagate to the parameter w of SEO and TSE can be trained by back-propagation.

3.2   Temporal Saliency Boosting Module

Although TSE can extract complementary features for input segment frames, the *saliency erasing operation* inevitably leads to information loss of the most salient part. To address this problem, we propose a *temporal saliency boosting module*. Since before the erasing operation, the intermediate feature maps of all
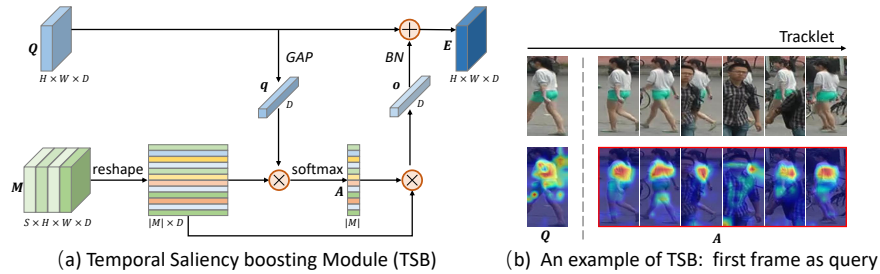
(a) Temporal Saliency boosting Module (TSB)   (b) An example of TSB: first frame as query

Fig. 3: (a) The architecture of TSB. (b) Visualization of probability map A computed on a video sequence, where we take the first frame as the query

high quality frames usually focus on the most salient parts. TSB is proposed to propagate the most salient information among the intermediate frame-level feature maps. In this way, the most salient feature can fully capture the visual cues in all frames thus presents strong discriminative power.

The structure of TSB is illustrated in Fig. 3 (a), which takes a set of frame-level feature maps of a video as the input. TSB is based on a *query-memory* attention mechanism, where we respectively consider the feature map of each frame as the query $Q \in \mathbb{R}^{H \times W \times D}$, and the memory $M \in \mathbb{R}^{S \times H \times W \times D}$ containing a collection of feature maps from the remaining S frames of the video is used to enhance the representational power of the query. Specifically, we firstly squeeze Q to a descriptor which can describe the query statistics. This is achieved by using GAP (global average pooling) to generate the channel-wise statistics $q \in \mathbb{R}^{D}$. Then we reshape M to $\mathbb{R}^{|M| \times D}$ ($|M| = S \times H \times W$) which can be viewed as a set of D-dimensional local descriptors. Corresponding, a probability map $A \in \mathbb{R}^{|M|}$ can be obtained regarding how well the query vector matches each descriptor of the memory through cosine similarity:

$$A_i = \frac{\exp(\tau \bar{q}^T \bar{M}_i)}{\sum_{j=1}^{|M|} \exp(\tau \bar{q}^T \bar{M}_j)}, \tag{6}$$

where $M_i \in \mathbb{R}^D$ denotes the ith local descriptor of M, $\bar{q}$ and $\bar{M}_i$ is the normalized q and M_i with L_2 norm, and τ is the temperature hyperparameter. The output $o \in \mathbb{R}^D$ is then calculated as the sum of all the items in the memory weighted by their probabilities as $o = M^T A$. In this way, the descriptors of M that are similar to the query present higher weights, which can avoid the corruption of low quality frames, *e.g.*, the occluded frames shown in Fig. 3 (b).

At last, we propagate the weighted descriptor o to the query Q with a residual learning scheme, which is defined as $E = \text{BN}(o) + Q$. Here, BN is a batch normalization [15] layer to adjust the scale of o to the query Q. Notably, before entering the BN layer, $o \in \mathbb{R}^D$ is duplicated along the spatial dimensions to $\mathbb{R}^{H \times W \times D}$ to be compatible with the size of Q.
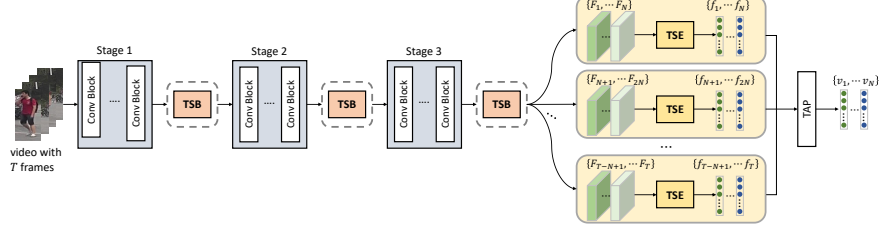
Fig. 4: The architecture of TCLNet. TAP denotes temporal average pooling layer

Taking the first frame of a video sequence as the query example, Fig. 3 (b) visualizes its initial feature map Q and corresponding probability map A. We can observe that Q roughly focuses on the most salient part (*i.e.*, shirts), and the probability map A can localize this part in other frames. Notably, through the *query-memory* matching, the corrupted frames presents lower weights in A, which indicates their features are suppressed during propagation and the output feature is robust to corruption. With the information propagation, the output feature fully captures the most salient visual cues in all frames of the video thus presents stronger representational power.

3.3   Overall Architecture

The architecture of Temporal Complementary Learning Network (TCLNet) that integrates TSE and TSB modules is illustrated in Fig. 4. Our network is built on ResNet-50 [9] pretrained on ImageNet [17]. ResNet-50 consists of four consecutive stages, *i.e.*, $stage1{\sim}4$, which respectively contains 3, 4, 6 and 3 residual blocks. We adopt the first three stages ($stage1{\sim}3$) as the backbone and the last stage ($stage4$) as the learners of TSE. TSB can be inserted into the backbone to any stage and TSE is added to the end of the backbone. In order to reduce the network complexity, the N learners of TSE share the same parameters for the first two residual blocks and have their own parameters in the last block.

Formally, given an video consisting of T consecutive frames, the backbone with inserted TSB firstly extracts features for each frame, which is denoted as $\mathcal{F} = \{F_1, F_2, \ldots, F_T\}$. Since the number of discriminative visual cues is usually finite, we only extract complementary features for N ($N < T$) consecutive frames. In particular, we equally divide $\mathcal{F}$ into L segments $\{C_k\}_{k=1}^L$ where each segment contains N consecutive feature maps, *i.e.*, $C_k = \{F_{(k-1)N+1}, \ldots, F_{kN}\}$. We then respectively feed each segment into TSE to extract complementary features for the segment frames:

$$c_k = \{f_{(k-1)N+1}, \ldots, f_{kN}\} = \mathrm{TSE}(F_{(k-1)N+1}, \ldots, F_{kN}) = \mathrm{TSE}(C_k). \qquad (7)$$

We finally apply temporal average pooling on $\{c_k\}_{k=1}^L$ that aggregates the set of segment features to generate the video features $\{v_1, \ldots v_N\}$. In the test stage,

the final video representation v can be obtained by concatenating the feature vectors extracted by all learners, *i.e.*, $v = [\hat{v}_1, \ldots, \hat{v}_N]$, where $\hat{v}_i$ denotes the L2-normalization of v_i.

Objective Function. Following the standard identity classification paradigm [35, 12], we add a classification layer to each video vector v_i. Cross entropy loss is then used for every v_i to guide the training of the corresponding learner. Recently, some works [34, 22] use the combination of cross entropy loss and batch triplet loss [10] to train the network. To fairly compare with this methods, we also explore a batch triplet loss during training.

4   Experiments

4.1   Dataset and Settings

Datasets. MARS [47], DukeMTMC-VideoReID [40] and iLIDS-VID [36] datasets are used for evaluation.

Implementation Details. Our method is implemented using the PyTorch framework [29]. During training, we sample four frames from each video sequence as input and each frame is resized to 256×128. We only adopt random flipping for data augmentation. The initial learning rate is set to 0.0003 with a decay factor 0.1 at every 40 epochs. Adam optimizer [18] is used with a mini-batch size of 32 for 150 epochs training. TSB is added to stage2 of the backbone. In TSE, the number of learners, *i.e*, the number of frames in each divided segment (N in Eq. 7), is set to 2, the height of erased block h_e is set to 3 and the width w_e is set to 8, and the strides s_h and s_w of sliding block are both set to 1. During testing, given an input of entire video, the video feature is extracted using the trained TCLNet for retrieval under cosine distance. Notably, *we use all the frames of a video to obtain the video feature in the testing phase.*

4.2   Comparison with State-of-the-art Methods

In Tab. 1, we compare our method with state-of-the-arts on MARS, DukeMTMC-VideoReID and iLIDS-VID datasets. Our method outperforms the best existing methods. It is noted that: **(1)** The gaps between our results and those that consider the video as a set of unordered images [47, 43, 26, 21, 40, 10, 46] are significant: about 5% mAP improvement on MARS. The significant improvements demonstrate that it is effective to employ the temporal cues for video reID. **(2)** Recent works [20, 22, 13] uses 3D CNN or non-local blocks to learn the temporal cues, which require high computational complexity. Our TCLNet puts much less overheads with a better performance on MARS: about 1% mAP improvement. We attribute this improvement to the complementary features learned from video frames which enhance the discriminative capability of reID models. **(3)** The works [43, 26, 53, 30, 22, 34] use the triplet loss to promote the performance. To fairly compare with them, we also adopt a triplet loss which further increases our performance by about 1%. It still outperforms the best performing work [34] by a large margin.

Table 1: Comparison with related methods on MARS, DukeMTMC-VideoReID and iLIDS-VID datasets. The methods are separated into two groups: image-set based methods (**IS**) and temporal-sequence based methods (**TS**). * denotes those trained with the combination of cross entropy loss and triplet loss

	Methods	MARS		Duke-Video		iLIDS-VID
		mAP	top-1	mAP	top-1	top-1
IS	Mars [47]	49.3	68.3	-	-	53.0
	SeqDecision* [43]	-	71.2	-	-	60.2
	QAN* [26]	51.7	73.7	-	-	68.0
	DRSA [21]	65.8	82.3	-	-	80.2
	EUG [40]	67.4	80.8	78.3	83.6	-
	AttDriven [46]	78.2	87.0	-	-	86.3
TS	ASTPN [41]	-	44.0	-	-	62.0
	SeeForest* [53]	50.7	70.6	-	-	55.2
	DuATM* [30]	67.7	81.2	-	-	-
	M3D [20]	84.4	74.1			84.4
	Snipped [1]	76.1	86.3	-	-	85.4
	V3D* [22]	77.0	84.3	-	-	81.3
	GLTP [19]	78.5	87.0	93.7	**96.3**	86.0
	COSAM* [34]	79.9	84.9	94.1	95.4	-
	VRSTC [13]	82.3	88.5	93.5	95.0	83.4
	TCLNet	**83.0**	**88.8**	**95.2**	**96.3**	84.3
	TCLNet-tri*	**85.1**	**89.8**	**96.2**	**96.9**	**86.6**

4.3   Ablation Study

We investigate the effectiveness of TSE and TSB modules by conducting a series of ablation studies on MARS dataset. We adopt ResNet-50 [9] with temporal average pooling as baseline (denoted as base.). In this part, all models are trained with only cross entropy loss.

Effectiveness of TSE. We firstly evaluate the effect of TSE by replacing the stage4 layer of baseline with TSE (**base.+TSE**). As shown in Tab. 2, TSE module improves the performance remarkably. Compared with the baseline, employing TSE brings 2.9% mAP and 1.4% top-1 accuracy gains respectively with negligible computational overhead. We argue that the learners of TSE work collaboratively to mine complementary parts so as to generate integral characteristic of the target identity, which helps to distinguish different identities with seemingly similar local parts. An example is shown in Fig. 5.

Effectiveness of *Saliency Erasing Operation* in TSE. It is noteworthy that the improvement of TSE dose not just come from the increased parameters by its learners. To see this, we introduce a variant of TSE, *i.e.*, **TSE-wo-SEO**, which adopts a series of ordered learners for consecutive frames without the *saliency erasing operation*. As shown in Tab. 2, TSE-wo-SEO brings only a small improvement over the baseline, indicating that the visual features captured by different learners are almost the same without the *saliency erasing operation*. While TSE performs significantly better than TSE-wo-SEO, which validates the

Table 2: Component analysis of the proposed network on MARS. We also report the number of floating-point operations (GFLOPs) for a four-frames sequence, and the parameter number (Params) of the models

Models	GFLOPs	Params	mAP	top-1
base.	16.246	23.5M	79.6	86.8
base.+TSE-wo-SEO	16.246	27.9M	80.9	87.2
base.+TSE	16.251	29.9M	**82.5**	**88.2**
base.+TSB	16.254	23.5M	82.3	87.6
TCLNet(TSE+TSB)	16.259	29.9M	**83.0**	**88.8**
TCLNet(TSE+TSB-stage1)	16.267	29.9M	82.2	87.4
TCLNet(TSE+TSB-stage2)	16.259	29.9M	**83.0**	**88.8**
TCLNet(TSE+TSB-stage3)	16.255	29.9M	82.6	88.2
TCLNet(TSE+TSB-stage23)	16.263	29.9M	82.7	88.2

Table 3: Impact of TSE hyper-parameters on MARS

(a) The number of ordered learners

N	GFLOPs	Params	mAP	top-1
1 (base.)	16.246	23.5M	79.6	86.8
2	16.251	29.9M	**82.5**	88.2
3	16.252	34.5M	82.4	**88.4**
4	16.253	38.9M	81.0	87.0

(b) Erased block height h_e

h_e	mAP	top-1
2	81.8	87.2
3	**82.5**	**88.2**
4	82.0	87.4
5	81.7	87.2

powerful capability of the *saliency erasing operation* to force different learners to focus on diverse image parts so as to discover integral visual features. Overall, we can see that the improvement of TSE mainly comes from the *saliency erasing operation* rather than the increased parameters.

TSE *w.r.t* Number of Ordered Learners. As shown in Fig. 2 (a), TSE contains N ordered learners that mine complementary parts for N consecutive frames. Tab. 3 (a) studies the impact of the number of learners N on the model base.+TSE. We can observe that the performance increases as more learners are considered to mine complementary parts. However, the performance drops largely when N reaches to 4. In this case, most discriminative parts have been erased in the last frame of the input segment, the fourth learner has to activate non-discriminative regions, *e.g.*, the background, which corrupts the final video representation. Considering the model complexity, we set N to 2 in our work.

TSE *w.r.t* Erased Block Size. In our network, TSE is applied for the frame-level feature map with size 16×8. Because of the spatial structure of the pedestrian images, we fix the erased width to the width of the feature map to erase entire rows of the feature map. Tab. 3 (b) studies the impact of the erased height h_e on the performance of TSE. We observe the best performance when $h_e = 3$, and it becomes worse when the erased height is larger or smaller. We can conclude: 1) too small erased size cannot effectively encourage the current

learner to discover the complementary parts; 2) too large erased size force the current learner to activate non-discriminative image parts, *e.g.*, background.

Effectiveness of TSB. We further assess the effectiveness of TSB module by adding it to the stage2 of baseline (**base.+TSB**) in Tab. 2. TSB individually brings 2.7% mAP and 0.8% top-1 accuracy gains with an extremely small increase in computational complexity. The improvements indicate that it is effective to enhance the feature representation power by propagating the salient features among the video frames. When we integrate TSE and TSB modules together to TCLNet, the performance can be further improved by about 1% on mAP and top-1 accuracy.

Efficient Positions to Place TSB. Tab. 2 compares a single TSB module added to different stages of ResNet50 in our TCLNet. The improvements of an TSB module in stage2 and stage3 are similar, but smaller in stage1. One possible explanation is that stage1 has a big spatial size 64×32 that is not very expressive and sufficient to provide precise semantic information. We also present the results of more TSB modules. In particular, we add a TSB module to stage2 and stage3 of the backbone respectively. However, we observe that adding more TSB modules does not bring improvement, indicating that one TSB module is enough for enhancing the salient features.

Complexity Comparisons. As shown in Tab. 2. We can observe that TSE and TSB introduce negligible computational overhead. In particular, TCLNet requires 16.259 GFLOPs, corresponding to only 0.08% relative increase over original model (base.). The increased computation cost mainly comes from the correlation map of TSE and probability map of TSB, which can be worked out by matrix multiplications thus occupy little time in GPU libraries. TCLNet introduces 6.4M parameters to baseline, which mainly come from the learners of TSE. Noting that only extending the baseline with a series of learners (base.+TES-wo-SEO) brings marginal improvement, showing the improvement of TCLNet is not just because of the added parameters.

Single shot reID. TSE is easy to generalize to single shot reID, where different learners discover diverse visual cues for the input image. We compare ResNet50 and ResNet50+TSE on Market1501 dataset [48]. ResNet50+TSE outperforms ResNet50 by 3.2% mAP (85.3%/80.3%), which indicates the good generality of our method.

4.4   Comparison with Related Approaches

Comparison TSE with Other Erasing Strategies. Tab. 4 (a) compares various erasing methods that are applied to baseline. DropBlock [6] *randomly* drops a contiguous regions of the convolution features during *training* to *prevent overfitting.* As shown in Tab. 4 (a), DropBlock brings marginal improvements while TSE significantly outperforms DropBlock by 2.7% mAP and 1.3% top-1 accuracy. The significant improvements show that it is more efficient to use our erasing strategy to extract complementary features for consecutive frames in video reID tasks. Random Erasing (RE) [51] randomly erases a rectangle region of the *input images* during training, which is a widely used data augmentation

Table 4: Compare TSB and TSE modules with related approaches on MARS

(a) TSE *vs.* other erasing methods

Models	mAP	top-1
base.	79.6	86.8
base.+DropBlock [6]	79.8	86.9
base.+RE [51]	81.5	86.6
base.+TSE	**82.5**	**88.2**
base.+TSE+RE	**82.9**	88.1

(b) TSB *vs.* other feature propagation methods

Models	GFLOPs	Params	mAP	top-1
base.+3D [16]	22.756	33.7M	80.0	86.1
base.+NL [37]	21.615	25.6M	**82.4**	**87.6**
base.+TSB	16.254	23.5M	82.3	**87.6**
base.+TSE+NL	21.619	32.1M	82.6	87.6
TCLNet	16.259	29.9M	**83.0**	**88.8**

technique. As shown in Tab. 4 (a), TSE still outperforms RE. Furthermore, as a data augmentation technique, RE is compatible with our method, in which the performance can be furthered lifted 0.4% mAP when combining with RE.

Comparison TSB with Other Feature Propagation Strategies. Tab. 4 (b) compares TSB with other feature propagation strategies. We can see that our TSB significantly outperforms 3D convolution. We argue that 3D convolution is prone to the corruption of low quality frames and hard to optimize because of the large parameter overhead. Compared with non-local (NL) [37] method, our TSB can achieve comparable performance under less computation budge and model size. More importantly, TSB is more effective to combine with TSE, where TCLNet (TSE+TSB) outperforms base.+TSE+NL by 1.2% top-1 accuracy. We argue that TSB only propagates the *salient information* among video frames which is more complementary to TSE module.

4.5 Visualization Analysis

Visualization of Feature Maps. For qualitative analysis, we compare the visualization results of feature maps extracted by baseline and TCLNet for some input video segments. As shown in Fig. 5 (b), the features of baseline only pay attention to some local regions, *i.e.*, the red T shirts, which are difficult to distinguish the different pedestrians in Fig. 5 (a). Instead, TCLNet is able to mine complementary parts for consecutive frames. As shown in Fig. 5 (c), for a video segment consisting of two consecutive frames, the feature of the first frame learned by the learner L_1 is most related to the red T shirts, while the learner L_2 activates the lower body for the second frame. With the complementary parts mined by learner L_2, different persons with similar local appearances become distinguishable. So the final ranking results can be improved significantly.

Visualization of Feature Distribution. Furthermore, we choose a number of person IDs with similar appearances from the test set of MARS to visualize the feature distribution by t-SNE [27]. These pedestrians wear blue shirts with small inter-person variation as shown in Fig. 6 (c). For the baseline that only focuses on local parts, the features belonging to these different identities are staggered. With TSE extracting complementary features and TSB enhancing salient features, the features of different identities extracted by TCLNet become
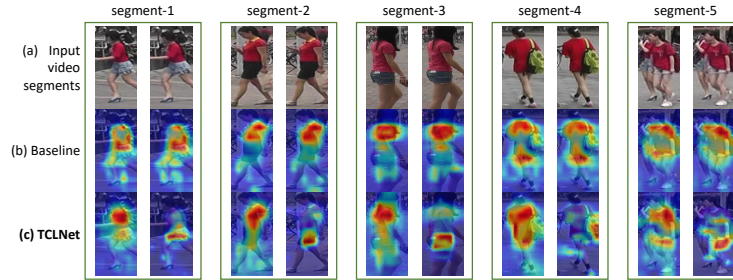
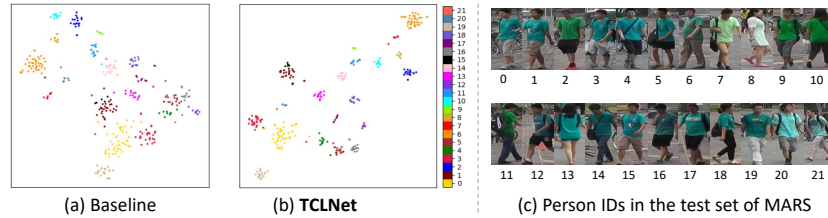Fig. 5: Feature map visualization of baseline and TCLNet

Fig. 6: tSNE visualization of feature distribution of baseline and TCLNe on MARS test set. Different colors indicate different identities

more separable. Specifically, by comparing Fig. 6 (a) and (b), we can observe that for some identities that are hard to be distinguished by baseline, the proposed model can better distinguish them, $e.g.$, the 1st, 3rd and 4th identities.

5 Conclusions

In this work, we propose a novel Temporal Complementary Learning Network for video person reID. Firstly, we introduce the Temporal Saliency Erasing module for complementary feature learning of video frames. TSE employs a saliency erasing strategy to progressively discover diverse and complementary visual cues for consecutive frames of a video. Furthermore, we propose the Temporal Saliency Boosting module to propagate the salient information among video frames. After propagation, the salient features capture the visual cues of all frames in a video and obtain stronger representational power. Extensive experiments demonstrate the superiority of our method over current state-of-the-art methods. In the future work, we will improve our method for longer-term temporal modeling. Also, we will combine our method with an efficient strategy to erase the noisy frames for more robust feature representation.

References

1. Chen, D., Li, H., Xiao, T., Yi, S., Wang, X.: Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding. In: CVPR. pp. 1169–1178 (2018)
2. Chen, D., Xu, D., Li, H., Sebe, N., Wang, X.: Group consistent similarity learning via deep crf for person re-identification. In: CVPR. pp. 8649–8658 (2018)
3. Dai, Z., Chen, M., Zhu, S., Tan, P.: Batch feature erasing for person re-identification and beyond. arXiv preprint arXiv:1811.07130 **1**(2), 3 (2018)
4. DeVries, T., Taylor, G.W.: Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552 (2017)
5. Ge, Y., Gu, X., Chen, M., Wang, H., Yang, D.: Deep multi-metric learning for person re-identification. In: ICME. pp. 1–6 (2018)
6. Ghiasi, G., Lin, T.Y., Le, Q.V.: Dropblock: A regularization method for convolutional networks. In: NeurIPS. pp. 10727–10737 (2018)
7. Gu, X., Ma, B., Chang, H., Shan, S., Chen, X.: Temporal knowledge propagation for image-to-video person re-identification. In: ICCV. pp. 9647–9656 (2019)
8. Gu, X., Ma, B., Chang, H., Zhang, H., Chen, X.: Appearance-preserving 3d convolution for video-based person re-identification. In: ECCV (2020)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770 – 778 (2016)
10. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person reidentification. arXiv preprint arXiv: 1703.07737 (2017)
11. Hou, R., Chang, H., Ma, B., Shan, S., Chen, X.: Cross attention network for few-shot classification. In: NeurIPS. pp. 4003–4014 (2019)
12. Hou, R., Ma, B., Chang, H., Gu, X., Shan, S., Chen, X.: Interaction-and-aggregation network for person re-identification. In: CVPR. pp. 9317–9326 (2019)
13. Hou, R., Ma, B., Chang, H., Gu, X., Shan, S., Chen, X.: Vrstc: Occlusion-free video person re-identification. In: CVPR. pp. 7183–7192 (2019)
14. Hou, R., Chang, H., Ma, B., Chen, X.: Video prediction with bidirectional constraint network. In: FG. pp. 1–8 (2019)
15. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167 (2015)
16. Ji, S., Xu, W., Yang, M., Yu, K.: 3d convolutional neural networks for human action recognition. IEEE transactions on pattern analysis and machine intelligence **35**(1), 221–231 (2012)
17. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: CVPR. pp. 1725–1732 (2014)
18. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
19. Li, J., Wang, J., Tian, Q., Gao, W., Zhang, S.: Global-local temporal representations for video person re-identification. In: ICCV (2019)
20. Li, J., Zhang, S., Huang, T.: Multiscale 3d convolution network for video based person reidentification. In: AAAI (2019)
21. Li, S., Bak, S., Carr, P., Hetang, C., Wang., X.: Diversity regularized spatiotemporal attention for video-based person re-identification. In: CVPR. pp. 369–378 (2018)
22. Liao, X., He, L., Yang, Z.: Video-based person re-identification via 3d convolutional networks and non-local attention. In: ACCV (2018)

23. Liu, F., Zhang, L.: View confusion feature learning for person re-identification. In: CVPR. pp. 6639–6648 (2019)
24. Liu, N., Zhao, Q., Zhang, N., Cheng, X., Zhu, J.: Pose-guided complementary features learning for amur tiger re-identification. In: ICCV Workshops. pp. 0–0 (2019)
25. Liu, S., Hao, X., Zhang, R., Zhang, Z., Durrani, T.S.: Adversarial erasing attention for person re-identification in camera networks under complex environments. IEEE Access **8**, 56469–56479 (2020)
26. Liu, Y., Yan, J., Ouyang, W.: Quality aware network for set to set recognition. In: CVPR. pp. 4694–4703 (2017)
27. Maaten, L.V.D., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research (2008)
28. McLaughlin, N., del Rincon, J.M., Miller, P.C.: Recurrent convolutional network for video-based person re-identification. In: CVPR. pp. 1325–1334 (2016)
29. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch. In: NeurIPS workshop (2017)
30. Si, J., Zhang, H., Li, C., Kuen, J., Kong, X., Kot, A.C., Wang, G.: Dual attention matching network for context-aware feature sequence based person re-identification. In: CVPR (2018)
31. Singh, K.K., Lee, Y.J.: Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In: ICCV (2017)
32. Song, G., Leng, B., Liu, Y., Hetang, C., Cai, S.: Region-based quality estimation network for large-scale person re-identification. arXiv preprint arXiv:1711.08766 (2017)
33. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research **15**(1), 1929–1958 (2014)
34. Subramaniam, A., Nambiar, A., Mittal, A.: Co-segmentation inspired attention networks for video-based person re-identification. In: ICCV (2019)
35. Sun, Y., Zheng, L., Yang, Y., Tian, Q., Wang, S.: Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In: ECCV. pp. 480–496 (2018)
36. Wang, T., Gong, S., Zhu, X., Wang, S.: Person reidentification by video ranking. In: ECCV. pp. 688–703 (2014)
37. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: CVPR. pp. 7794–7803 (2018)
38. Wang, X., Shrivastava, A., Gupta, A.: A-fast-rcnn: Hard positive generation via adversary for object detection. In: CVPR. pp. 2606–2615 (2017)
39. Wei, L., Zhang, S., Gao, W., Tian, Q.: Person trasfer gan to bridge domain gap for person re-identification. In: CVPR. pp. 79–88 (2018)
40. Wu, Y., Lin, Y., Dong, X., Yan, Y., Quyang, W., Yang, Y.: Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning. In: CVPR. pp. 5177–5186 (2018)
41. Xu, S., Cheng, Y., Gu, K., Yang, Y., Chang, S., Zhou, P.: Jointly attentive spatial-temporal pooling networks for video-based person re-identification. In: ICCV. pp. 4743–4752 (2017)
42. Yang, W., Huang, H., Zhang, Z., Chen, X., Huang, K., Zhang., S.: Towards rich feature discovery with class activation maps augmentation for person re-identification. In: CVPR (2019)

43. Zhang, J., Wang, N., Zhang, L.: Multi-shot pedestrian re-identification via sequential decision making. In: CVPR (2018)
44. Zhang, L., Shi, Z., Zhou, J.T., Cheng, M.M., Liu, Y., Bian, J.W., Zeng, Z., Shen, C.: Ordered or orderless: A revisit for video based person re-identification. IEEE Transactions on Pattern Analysis and Machine Intelligence (2020)
45. Zhang, Z., Lan, C., Zeng, W., Chen, Z.: Densely semantically aligned person re-identification. In: CVPR. pp. 667–676 (2019)
46. Zhao, Y., Shen, X., Jin, Z., Lu, H., Hua, X.: Attribute-driven feature disentangling and temporal aggregation for video person re-identification. In: CVPR (2019)
47. Zheng, L., Bie, Z., Sun, Y., Wang, J., Su, C., Wang, S., Tian, Q.: Mars: A video benchmark for large-scale person re-identification. In: ECCV. pp. 868–884 (2016)
48. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: A benchmark. In: ICCV. pp. 1116–1124 (2015)
49. Zheng, M., Karanam, S., Wu, Z., Radke, R.J.: Re-identification with consistent attentive siamese networks. In: CVPR. pp. 5735–5744 (2019)
50. Zheng, Z., Zheng, L., Yang, Y.: Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In: ICCV. pp. 3754–3762 (2017)
51. Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y.: Random erasing data augmentation. arXiv preprint arXiv:1708.04896 (2017)
52. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: CVPR. pp. 2921–2929 (2016)
53. Zhou, Z., Huang, Y., Wang, W., Wang, L., Tan., T.: See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification. In: CVPR. pp. 6776–6785 (2017)