Spatially and Temporally Efficient Non-local Attention Network for Video-based Person Re-Identification

Chih-Ting Liu[1, 2, 3]
jackieliu@media.ee.ntu.edu.tw

Chih-Wei Wu[1, 2, 3]
cwwu@media.ee.ntu.edu.tw

Yu-Chiang Frank Wang[2]
ycwang@ntu.edu.tw

Shao-Yi Chien[1, 2, 3]
sychien@ntu.edu.tw

[1] IOX Center
National Taiwan University, Taiwan

[2] Department of Electrical Engineering
National Taiwan University, Taiwan

[3] Graduate Institute of Electronics
Engineering
National Taiwan University, Taiwan

Abstract

Video-based person re-identification (Re-ID) aims at matching video sequences of pedestrians across non-overlapping cameras. It is a practical yet challenging task of how to embed spatial and temporal information of a video into its feature representation. While most existing methods learn the video characteristics by aggregating image-wise features and designing attention mechanisms in Neural Networks, they only explore the correlation between frames at high-level features. In this work, we target at refining the intermediate features as well as high-level features with non-local attention operations and make two contributions. (i) We propose a Non-local Video Attention Network (NVAN) to incorporate video characteristics into the representation at multiple feature levels. (ii) We further introduce a Spatially and Temporally Efficient Non-local Video Attention Network (STE-NVAN) to reduce the computation complexity by exploring spatial and temporal redundancy presented in pedestrian videos. Extensive experiments show that our NVAN outperforms state-of-the-arts by **3.8**% in rank-1 accuracy on MARS dataset and confirms our STE-NVAN displays a much superior computation footprint compared to existing methods. Codes are available at https://github.com/jackie840129/STE-NVAN.

1   Introduction

Person re-identification (Re-ID) tackles the problem of retrieving pedestrian images/videos across non-overlapping cameras. Previous approaches mostly focus on image-based Re-ID, where each pedestrian possesses multiple images for retrieval [4, 9, 11, 18, 22, 24, 34, 42]. Recently, video-based Re-ID has drawn significant attention in literature since retrieving pedestrian videos is more realistic and critical in real-world surveillance applications [12, 26, 32, 35]. With the emergence of large-scale video-based Re-ID datasets [26, 35], researchers design Deep Neural Networks to learn robust representation for videos [1, 8, 17, 23, 26, 45].

To perform video-based Re-ID, typical methods require learning a mapping function to project the video sequences to a low-dimensional feature space, where Re-ID can then be performed by comparing distances between samples. As demonstrated by numerous works, training the convolutional Neural Network (CNN) as a mapping function has dominated over classic methods with hand-crafted features [6, 15, 20]. Usually, they obtain features for a sequence by aggregating image features with average or maximum pooling [23, 26]. However, their approaches fail to handle occlusion or spatial misalignment in video sequences since it treats all images in a sequence with equal importance [0]. In order to distill relevant information for Re-ID, some works integrate Recurrent Neural Network to learn the spatial-temporal dependency in an end-to-end training manner [2, 23, 39]. Recently, several works propose attention mechanism to weight the importance of different frames or different spatial locations to aggregate a better representation [0, 8, 17]. While these methods successfully capture both the spatial and temporal characteristics of video sequences, they only explore the aggregation of high-level features for representation, which might not be sufficiently robust for fine-grained classification tasks such as Re-ID [19, 25, 40].

In this paper, we first aim to improve the representation for video sequences by exploiting spatial and temporal characteristics in both low-level and high-level features. Inspired by Wang *et al*. [33], we propose a Non-local Video Attention Network (NVAN) by introducing the non-local attention layer into an image classification CNN model. The non-local attention layer enriches the local image feature with global sequence information by generating attention masks according to features of different frames and different spatial locations. By inserting non-local attention layers at different feature levels, NVAN explores the spatial and temporal diversity of a sequence and alters its feature representation subsequently rather than combining individual image features with a set of weights as in previous works. Our NVAN model surpasses all state-of-the-art video-based Re-ID methods by a large margin on the challenging MARS [26] dataset, proving that exploiting global information for multi-level features is crucial for learning representation for video sequences.

While applying non-local attention layer to multi-level features significantly improves the Re-ID performance, it comes at a great cost in terms of computation complexity. In fact, it increases the total floating point operations (FLOP) by 99.3%, making it difficult to scale up to practical applications. To alleviate such challenge, we take advantage of the space-time redundancy in pedestrian videos and propose a Spatially and Temporally Efficient Non-local Video Attention Network (STE-NVAN). We first reduce the granularity of attention masks in non-local attention layers by exploiting the spatial redundancy exhibited in pedestrian images. On the other hand, we explore the temporal redundancy between video frames to aggregate image-wise information into a representative video feature with a hierarchical structure. By reducing the computation complexity both spatially and temporally, our STE-NVAN cut down 72.7% of FLOP compared to original NVAN with only 1.1% drop in rank-1 accuracy on MARS dataset. Our proposed STE-NVAN demonstrates a much superior trade-off between performance and complexity compared to existing video-based Re-ID methods. The contribution of our work can be summarized as follows:

- We introduce the non-local attention operation into the backbone CNN at multiple feature levels to incorporate both spatial and temporal characteristics of pedestrian videos into the representation.
- We significantly reduce the computation count for our Non-local Video Attention Network by exploring the spatial and temporal redundancy presented in pedestrian videos.
- Extensive experiments validate that our proposed model not only outperforms state-

of-the-art methods in Re-ID accuracy but also requires less computation count than existing attention methods for video-based Re-ID.

2 Related Work

In this section, we briefly review the related works regarding image-based person Re-ID, video-based person Re-ID and the usage of attention mechanisms for the Re-ID problem.

Image-based person Re-ID has been extensively studied over the years. With the success of CNNs [11, 19, 27, 37, 44], deep features learned from the networks has replaced hand-crafted features [4, 9, 20, 42] for representing pedestrian images. As suggested by Zheng *et al.* [43], these networks can be categorized into discriminative learning and metric learning. Discriminative learning learns deep features for identity classification with the help of the cross-entropy loss [27, 37, 44]. As for metric learning, Hermans *et al.* [11] use the triplet loss to teach the network to push together features of the same person and pull away features of different people. In this work, we utilize both loss functions to train our network for video-based person Re-ID.

Video-based person Re-ID is an extension of image-based person Re-ID. Zheng *et al.* [26] introduce a large-scale dataset to enable the learning of deep features for video-based Re-ID. They first train a CNN to extract image features then aggregate them into a sequence features with average/maximum pooling. Other works [23, 39, 45] adopt Recurrent Neural Networks to summarize image-wise features into a single feature by exploiting temporal relation within a sequence.

Recently, attention mechanisms are introduced for capturing spatial and temporal characteristics of pedestrian sequences within the deep features. Xu *et al.* [38] introduce the joint attentive spatial and temporal pooling network to extract sequence features by jointly considering the query and gallery pairs with an affinity matrix. Li *et al.* [17] learn attention weights to combine features of different spatial locations and different temporal frames into a sequence feature. Chen *et al.* [1] utilize techniques in [29] to perform self-attention on each video snippet and co-attention between video snippets for learning sequence features. Fu *et al.* [8] learn sequence features by mining features of discriminative regions and select important frames with a parameter-free attention scheme. While these works achieve promising results by introducing spatial and temporal attention on top of high-level features obtained from image-based CNNs, they overlook the importance of utilizing video characteristics at intermediate feature levels. In contrast, our proposed NVAN is able to refine intermediate features with spatial and temporal information of videos and our efficient STE-NVAN model substantially reduces the computation cost for incorporating video characteristics at lower feature levels.

3 Proposed Method

Given an image sequence of any pedestrians, we aim to learn a CNN to extract its feature representation that enables video-based person Re-ID in the embedding space. The key to learning a representative feature for a sequence is to incorporate video characteristics into the feature itself. To this end, we introduce the non-local attention layer into the CNN to explore the spatial and temporal dependency of a video sequence. We propose a Non-local Video Attention Network (NVAN) in Sec. 3.1 to apply such operations at different feature levels. However, we observe incredibly large computation complexity with the introduction
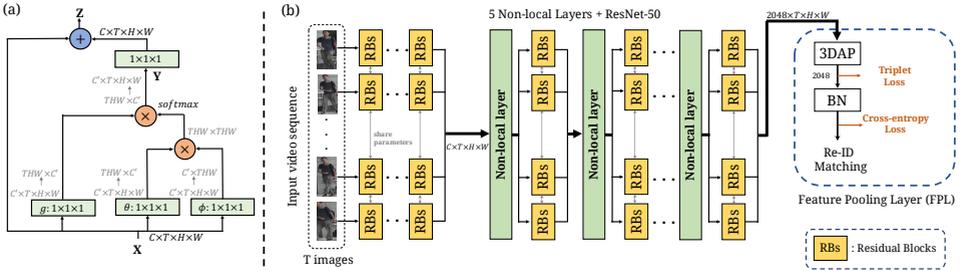
Figure 1: **(a) Details of Non-local Attention Layer. (b) Overview of our Non-local Video Attention Network (NVAN).** In NVAN, given T sampled images as input, the 5-Non-local ResNet-50 network generates T features, which incorporates the spatial and temporal information of videos at multi-levels with the help of Non-local Attention Layers. The features are then pooled into one vector in FPL for loss optimization and Re-ID matching.

of attention mechanisms. Hence, we further propose the Spatially and Temporally Efficient Non-local Video Attention Network (STE-NVAN) in Sec. 3.2 to alleviate the computation cost by exploiting spatial and temporal redundancy which exists in pedestrian videos.

3.1 Non-local Video Attention Network

To extract features for an image sequence, we take input as a subset of video frames selected by restricted random sampling (RRS) strategy and forward through a backbone CNN network incorporating non-local attention layers and a feature pooling layer (FPL) to obtain the representation vector for video-based Re-ID, as shown in Figure 1 (b).

Restricted Random Sampling (RRS). There are several ways to handle the long-range temporal structure. To balance speed and accuracy, we adopt the restricted random sampling strategy [17, 51]. Given an input video **V**, we divide it into T chunks $\{C_t\}_{t=[1,T]}$ of equal duration. For training, we randomly sample an image I_t in each chunk. As for testing, we use the first image of each chunk. The video is then represented by the ordered set of sampled frames $\{I_t\}_{t=[1,T]}$.

Non-local Attention Layer. To embed video characteristics into the features, we introduce the non-local layer proposed by Wang *et al.* [53] into the backbone CNN, as illustrated in Figure 1 (a). Given an input feature tensor $X \in \mathbb{R}^{C \times T \times H \times W}$ obtained from a sequence of T feature maps of size $C \times H \times W$, we desire to exchange information between features across all spatial locations and frames. Let $x_i \in \mathbb{R}^C$ sampled from X, the corresponding output $y_i \in \mathbb{R}^C$ of non-local operation can be formulated as follow:

$$y_i = \frac{1}{\sum_{\forall j} e^{\theta(x_i)^T \phi(x_j)}} \sum_{\forall j} e^{\theta(x_i)^T \phi(x_j)} g(x_j). \tag{1}$$

Here, $i, j = [1, THW]$ indexes all locations across a feature map and all frames. We first project x to a lower dimensional embedding space $\mathbb{R}^{C'}$ by using linear transformation functions θ, ϕ, g ($1 \times 1 \times 1$ convolution). Then, the response of each location x_i is computed by the weighted average of all positions x_j by using Embedded Gaussian instantiation. The Equation 1 in non-local layer is a self-attention mechanism which is also mentioned in [53].

The overall non-local layer is finally formulated as $Z = W_z Y + X$, where the output of non-local operation is added to the original feature tensor X with a transformation W_z ($1 \times 1 \times 1$ convolution) that maps Y to the original feature space $\mathbb{R}^C$. The intuition behind the non-local operation is that when extracting features at a specific location in a specific time, the network should consider the spatial and temporal dependency within a sequence by attending on the non-local context. In our person Re-ID scheme, we embed five non-local layers into our backbone CNN which is a ResNet-50 network [10] to comprehend the semantic relation presented in videos, as shown in Figure 1 (b).

Feature Pooling Layer (FPL).   After passing the image sequence through the backbone CNN and non-local attention layers, we employ the feature pooling layer to obtain the final feature for Re-ID, shown in Figure 1 (b). We apply 3D average pooling (3DAP) along the spatial and temporal dimension to aggregate the output features of each image into a representative vector, followed by a batch normalization (BN) layer [13]. We train the network by jointly optimizing the cross-entropy loss and the soft-margin batch-hard triplet loss [11]. Interestingly, we empirically find that optimizing cross-entropy loss on the final feature while optimizing triplet loss on the feature before BN results in the best Re-ID performance. A rational explanation is that the embedding space without normalization is more suitable for distance metric learning such as the triplet loss, while the normalized feature space forces the model to classify samples on a more constraint angular space with cross-entropy loss [5, 11, 21, 30].

3.2   Spatially and Temporally Efficient Non-local Video Attention Network

While our proposed NVAN is able to capture sophisticated properties of video sequence with the help of non-local operations, we observe a significant increase in the computation complexity as shown in Table 1, where FLOP ramp up from 30.4G to 60.0G. For scaling NVAN to practical usage scenarios, we introduce two complexity reduction techniques to cut down the computation count.

Spatial Reduction with Pedestrian Part Characteristics.   Originally, the introduced non-local operations perform dense affinity calculation between features of all THW positions to obtain a fine attention mask. This results in heavy computation of complexity $O(C'T^2H^2W^2 + CC'THW)$ for each non-local attention layer. Applying the non-local attention layer to lower feature levels incurs larger complexity since low level features are typically of higher H, W. To alleviate such effect, we group the features along the horizontal direction to form a more compact representation of the feature tensor. The intuition is that pixels of the same horizontal stripe tend to share similar characteristics which can be utilized to generate coarse representation of the image. It is worth noting that while similar ideas have been explored in Re-ID literature [2, 3, 20, 28], they use this concept to generate finer features for Re-ID. In contrast, we exploit this redundancy to obtain coarser representation. We partition the original feature tensor $X \in \mathbb{R}^{C \times T \times H \times W}$ into S horizontal groups by adding the "Make stripe" module at the input of non-local operations. The resulting tensor $X' \in \mathbb{R}^{C \times T \times S}$ requires only $O(C'T^2S^2 + CC'TS)$ to complete the operation, which is irrelevant to the spatial size of feature maps. This dramatically reduces the computation complexity and enables us to deploy non-local operation to lower feature levels with constant computation cost. We name it Spatial Reduction Non-local Layer and illustrate the idea in Figure 2 (a).
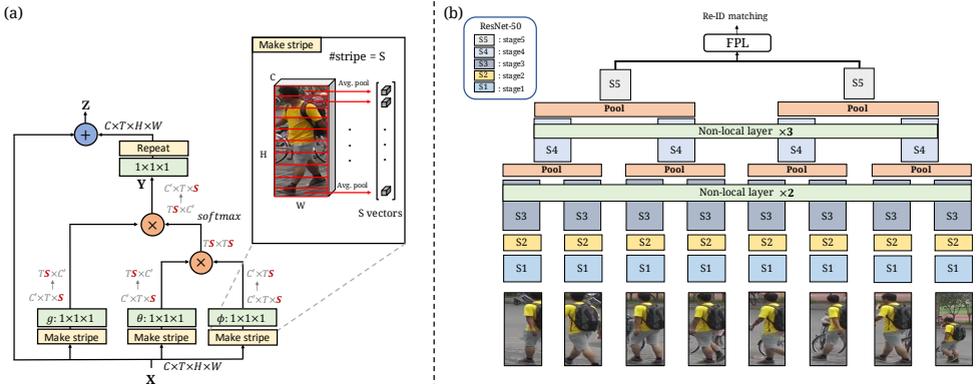
Figure 2: **(a) Spatial Reduction Non-local Layer. (b) Temporal Reduction with Hierarchical Structure.** Details are explained in Section 3.2. Noting that, for figure (a), before the residual addition, we repeat the tensor of shape $C \times T \times S$ to $C \times T \times H \times W$. As for figure (b), we apply max-pooling across adjacent features after the stages with non-local layers to construct our hierarchical structure.

Temporal Reduction with Hierarchical Structure.   During our experiments, we observe that features refined by non-local operations are often temporally similar as non-local operation aims to embed global temporal information into the features. Inspired by this observation, we exploit the temporal redundancy between features of different frames and propose a hierarchical structure to reduce the heavy computation of extracting sequence feature. We illustrate this idea in Figure 2. After passing a sequence of images through a series of convolutions (Residual blocks) and non-local attention layers, we apply max pooling across features of adjacent frames and reduce the temporal feature dimension by a factor of 2. We perform the same reduction operation after another stacks of Residual blocks until the temporal dimension is reduced to 2, which is then sent to FPL for final feature summarization. This temporal reduction technique cuts down the computation required for extracting sequence feature with Residual blocks and non-local attention layers. By applying both the Spatial Reduction Non-local Layers and the Hierarchical Temporal Reduction structure, we come up with the final Spatially and Temporally Efficient Non-local Video Attention Network (STE-NVAN) for video-based person Re-ID.

4   Experiments

We evaluate our approach on two large-scale video-based person Re-ID datasets, MARS [26] and DukeMTMC-VideoReID [35]. We conduct ablation studies to validate the effectiveness of non-local operations and the two proposed reduction methods. We compare our NVAN and STE-NVAN models to existing state-of-the-arts to demonstrate that our proposed models display superior performance while requiring less computation counts.

4.1   Experimental Setup

Datasets and Evaluation Protocol.   MARS [26] is one of the large video-based person Re-ID datasets, consisting of 17,503 tracks and 1,261 identities. Each track has 59 frames on average. Deformable Part Model [7] is employed to detect pedestrians and GMCP [41] is

Table 1: **Comparisons of different baselines with two reduction methods.** This table shows the performance results and the computation count of baseline models, NVAN and STE-NVAN. The "Reduc." is the abbreviation of Reduction.

Method	Feature Aggregation	MARS		DukeV		# FLOP
		R1	mAP	R1	mAP	
ResNet-50	FPL	87.3	79.1	95.0	92.7	30.4 G
ResNet-50	max-FPL	86.3	76.6	95.4	92.4	30.4 G
NVAN	FPL	**90.0**	**82.8**	**96.3**	**94.9**	60.0 G
NVAN+Spatial Reduc.	FPL	89.7	82.5	96.3	94.7	30.4 G
NVAN+Temporal Reduc.	FPL	89.2	81.2	95.6	93.7	40.4 G
STE-NVAN	FPL	88.9	81.2	95.2	93.5	**16.5 G**

used to track pedestrians. To make the dataset even more challenging, they include 3,248 distractor tracks in the dataset. DukeMTMC-VideoReID [35] is another large-scale benchmark recently introduced for video-based person Re-ID. It comprises 4,832 tracks and 1,404 identities and 408 distractor identities. Each track contains 168 frames on average. Detection and tracking ground truth are manually labeled. In the following literature, DukeMTMC-VideoReID will be abbreviated as "DukeV" for convenience. In our experiments, we adopt the standard train/test split and report both rank-1 accuracy (R1) and Mean Average Precision (mAP) to evaluate the Re-ID performance.

Implementation Detail. For the RRS strategy described in Sec. 3.1, we segment each video into $T = 8$ chunks and sampled 8 images as the input sequence. Each frame is resized to 256×128 and synchronously augmented with random horizontal flip for each track. We adopt the ImageNet pre-trained ResNet-50 [10] as our backbone network, and modified $conv5\_1$ to stride 1 instead of stride 2 to better adapt the Re-ID task. For our NVAN, we insert 2 non-local attention layers after $conv3\_3, con3\_4$ and another 3 after $con4\_4, con4\_5, con4\_6$ respectively. As for STE-NVAN, we set $S = 16$ in Spatial Reduction Non-local layer and perform max-pooling right after the second and the fifth non-local attention layer to reduce temporal dimension. We train our network for 200 epoch with both cross-entropy loss and triplet loss [11] and choose Adam [14] optimizer with an initial learning rate of 10^{-4} and decay it by 10 every 50 epochs. Following the suggestion in [11], we sample 8 identities, each with 4 tracks, to form a batch of size $8 \times 4 \times 8 = 256$ images.

4.2 Ablation Studies

Effectiveness of Non-local Attention Layer and Two Reduction Methods. We first compare our NVAN model with two baseline models to demonstrate the power of non-local operations. The two baseline models (ResNet-50) use the same backbone network as NVAN but without non-local attention layers. The only difference between the two baselines is that one replace the 3DAP in FPL with 3D maximum pooling operation. The first three rows in Table 1 illustrate the results. It reveals that non-local operations improve the R1 and mAP significantly by $2.7\%, 3.7\%$ on MARS and $1.3\%, 1.6\%$ on DukeV. The improvement confirms the effectiveness of incorporating spatial and temporal characteristics in the sequence feature of different semantic levels. However, we observe an dramatic 99.3% increase in FLOP accompanying the introduction of non-local operations. Therefore, we propose two reduction techniques by exploiting spatial and temporal redundancy in pedestrian videos. Table 1 shows that our spatial reduction strategy cuts down the FLOP to approximately the

Table 2: Comparison of NVAN network with different # frames of RRS strategy.

# frames	MARS		DukeV	
	R1	mAP	R1	mAP
$T = 4$	89.0	81.0	95.3	92.7
$T = 6$	89.4	81.6	95.6	93.4
$T = 8$	90.0	82.8	96.3	94.9

Table 3: Comparison of NVAN network with different # non-local layers embedded.

# non-local layers	MARS		DukeV	
	R1	mAP	R1	mAP
1 layer	89.0	81.8	95.8	93.7
3 layers	89.0	82.4	96.3	94.9
5 layers	90.0	82.8	96.3	94.9

Table 4: Comparison of different # stripes in spatial reduction non-local layer.

# stripes	MARS	DukeV	#FLOP
	R1	R1	
$S = 4$	89.6	96.3	30.4G
$S = 8$	89.5	96.1	30.4G
$S = 16$	89.7	96.3	30.4G

Table 5: Comparison of different pooling position combinations in hierarchical structure.

Pooling positions	MARS	DukeV	#FLOP
	R1	R1	
in 3DAP	90.0	96.3	60.0G
+stage 4	89.8	96.1	53.6G
+stage 3	89.2	95.6	40.4G

same level as baseline networks while only incurring 0.3% R1/mAP drop on MARS and 0.2% mAP drop on DukeV. As for temporal reduction, we save 32.6% of FLOP from NVAN and sustain only 1.1% R1 loss on both datasets and 1.7% and 1.2% mAP loss. Finally, by applying both spatial and temporal reduction techniques on NVAN, which is our STE-NVAN, we achieve **72.7%** FLOP reduction compare to NVAN and requires **45.7%** less FLOP compare to the baseline that doesn't employ any attention mechanism. It shows that our proposed STE-NVAN not only improves the Re-ID performance but also demonstrates a more efficient method of extracting sequence features.

Analysis of NVAN.   To better understand the property of non-local operations, we conduct analysis on NVAN regarding RRS strategy and number of inserted non-local attention layers. In Table 2, we discover that by increasing the number of frames T sampled from a sequence in RRS, Re-ID performance increases steadily as more frames provide richer information about a pedestrian. We pick $T = 8$ for our NVAN and STE-NVAN in consideration of the memory capacity of our machine. On the other hand, we observe performance gain as we insert more non-local attention layers. In Table 3, we insert a non-local layer at $conv4\_6$ for "1 layer" and insert 3 non-local layers at $conv3\_4, conv4\_5, conv4\_6$ for "3 layers". We insert 5 non-local layers for NVAN and STE-NVAN since it performs the best.

Analysis of STE-NVAN.   Next we investigate the parameters for designing STE-NVAN. Starting from NVAN, we apply the spatial reduction techniques to group features into horizontal stripes in non-local attention layer. Table 4 shows that while increasing number of stripes S does not introduce excessive additional FLOP, it improves the Re-ID performance subtly. As for analyzing temporal reduction, we increase the pooling operations throughout the network. For comparison, "in 3DAP" in Table 5 is the NVAN model that pools all features after the last convolutional layer. By employing additional pooling after the non-local layers located in $stage4$ ("+ stage 4"), we reduce 10.7% of FLOP from NVAN. And by introducing another additional pooling after non-local layers at $stage3$ ("+ stage 3"), we remove 32.7% of FLOP from NVAN while only dropping 0.8% and 0.7% of R1 on MARS and DukeV.

Table 6: **Comparison with state-of-the-arts approaches on MARS and DukeV.**

Methods	Source	MARS		DukeV	
		R1	mAP	R1	mAP
CNN+Kiss. [26]	ECCV16	65.0	45.6	-	-
SeeForest [45]	CVPR17	70.6	50.7	-	-
LatentParts [16]	CVPR17	70.6	50.7	-	-
TriNet [11]	arXiv17	79.8	67.7	-	-
ETAP-Net(supervised) [36]	CVPR18	80.8	67.4	83.6	78.3
STAN [17]	CVPR18	82.3	65.8	-	-
CSACSE+OF [1]	CVPR18	86.3	76.1	-	-
STA (N=8) [8]	AAAI19	86.2	81.2	96.0	**95.0**
NVAN (ours)	-	**90.0**	**82.8**	**96.3**	94.9
STE-NVAN (ours)	-	88.9	81.2	95.2	93.5

4.3 Comparison with State-of-the-arts Approaches

Table 6 reports the comparison of our NVAN and STE-NVAN to state-of-the-art video-based person Re-ID approaches. For STA [8], we display their results sampling 8 images per sequence to be fair with our method. On MARS, our NVAN achieves **90.0**% in R1 and **82.8**% in mAP, surpassing all methods by a large margin. Our efficient STE-NVAN also performs better than all methods in R1 and breaks even with STA in mAP despite using less FLOP than NVAN. On the other hand, our NVAN and STE-NVAN still displays competitive results on DukeV, where Re-ID on DukeV is easier than MARS since detection are manually annotated. The superior Re-ID performance on two benchmark datasets proves the value of applying non-local operations for extracting a better representation of videos.

To take the computation complexity into consideration, we compare our method with existing methods that also uses attention mechanisms on the performance-computation plot in Figure 3. We visualize mAP on MARS dataset for the performance and #FLOP for computation counts. For STA, we report three variants of their with different numbers of sampled frames per sequence to better demonstrate their trade-off. Results show that our proposed STE-NVAN exhibits a much better mAP-FLOP trade-off compared to current state-of-the-arts. STAN [17] and CSACSE+OF [1] even lands outside of the plot since their mAP and FLOP are beyond the scale of our plot. The results not only indicates the advantage of our proposed spatial and temporal reduction techniques but also reveal the importance of considering computation complexity when design feature extractors for video sequences.

5 Conclusion

We introduce a Non-local Video Attention Network (NVAN) which incorporates multiple non-local attention layers to extract spatial and temporal video characteristics from low to high feature levels, which enrich the representation of videos in person re-identification. To alleviate the computation cost, we proposed a Spatially and Temporally Efficient Non-local Video Attention Network (STE-NVAN), which spatially reduce the non-local operation by utilizing pedestrian part characteristics and temporally reduce the operation with hierarchical structure. Extensive experiments are conducted to prove that our STE-NVAN is a superior trade-off between performance and computation.
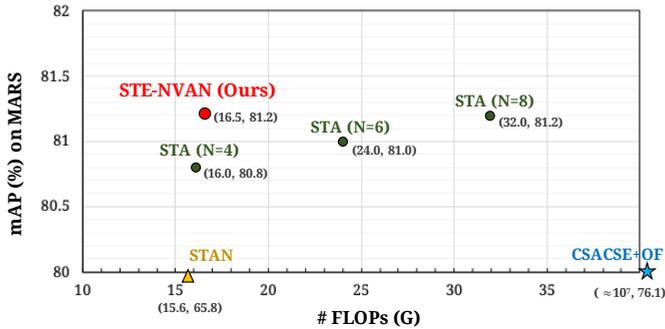
Figure 3: **Computation-performance plot of our proposed STE-NVAN and existing methods with attention mechanisms.**

Acknowledgment

References

[1] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1169–1178, 2018.

[2] Lin Chen, Hua Yang, Ji Zhu, Qin Zhou, Shuang Wu, and Zhiyong Gao. Deep spatial-temporal fusion network for video-based person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 63–70, 2017.

[3] De Cheng, Yihong Gong, Sanping Zhou, Jinjun Wang, and Nanning Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1335–1344, 2016.

[4] Dong Seon Cheng, Marco Cristani, Michele Stoppa, Loris Bazzani, and Vittorio Murino. Custom pictorial structures for re-identification. In *Bmvc*, volume 1, page 6. Citeseer, 2011.

[5] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. *arXiv preprint arXiv:1801.07698*, 2018.

[6] Michela Farenzena, Loris Bazzani, Alessandro Perina, Vittorio Murino, and Marco Cristani. Person re-identification by symmetry-driven accumulation of local features. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2360–2367. IEEE, 2010.

[7] Pedro F Felzenszwalb, David A McAllester, Deva Ramanan, et al. A discriminatively trained, multiscale, deformable part model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

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

[9] Douglas Gray and Hai Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *European conference on computer vision*, pages 262–275. Springer, 2008.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

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

[12] Martin Hirzer, Csaba Beleznai, Peter M Roth, and Horst Bischof. Person re-identification by descriptive and discriminative classification. In *Scandinavian conference on Image analysis*, pages 91–102. Springer, 2011.

[13] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[15] Martin Koestinger, Martin Hirzer, Paul Wohlhart, Peter M Roth, and Horst Bischof. Large scale metric learning from equivalence constraints. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2288–2295. IEEE, 2012.

[16] Dangwei Li, Xiaotang Chen, Zhang Zhang, and Kaiqi Huang. Learning deep context-aware features over body and latent parts for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 384–393, 2017.

[17] Shuang Li, Slawomir Bak, Peter Carr, and Xiaogang Wang. Diversity regularized spatiotemporal attention for video-based person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 369–378, 2018.

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

[19] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2285–2294, 2018.

[20] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li. Person re-identification by local maximal occurrence representation and metric learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2197–2206, 2015.

[21] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 212–220, 2017.

[22] Chen Change Loy, Chunxiao Liu, and Shaogang Gong. Person re-identification by manifold ranking. In *2013 IEEE International Conference on Image Processing*, pages 3567–3571. IEEE, 2013.

[23] Niall McLaughlin, Jesus Martinez del Rincon, and Paul Miller. Recurrent convolutional network for video-based person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1325–1334, 2016.

[24] Ergys Ristani and Carlo Tomasi. Features for multi-target multi-camera tracking and re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6036–6046, 2018.

[25] Ya-Fang Shih, Yang-Ming Yeh, Yen-Yu Lin, Ming-Fang Weng, Yi-Chang Lu, and Yung-Yu Chuang. Deep co-occurrence feature learning for visual object recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4123–4132, 2017.

[26] *MARS: A Video Benchmark for Large-Scale Person Re-identification*, 2016. Springer.

[27] Yifan Sun, Liang Zheng, Weijian Deng, and Shengjin Wang. Svdnet for pedestrian retrieval. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3800–3808, 2017.

[28] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 480–496, 2018.

[29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[30] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5265–5274, 2018.

[31] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pages 20–36. Springer, 2016.

[32] Taiqing Wang, Shaogang Gong, Xiatian Zhu, and Shengjin Wang. Person re-identification by discriminative selection in video ranking. *IEEE transactions on pattern analysis and machine intelligence*, 38(12):2501–2514, 2016.

[33] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2018.

[34] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 79–88, 2018.

[35] 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 *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[36] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5177–5186, 2018.

[37] Tong Xiao, Hongsheng Li, Wanli Ouyang, and Xiaogang Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1249–1258, 2016.

[38] Shuangjie Xu, Yu Cheng, Kang Gu, Yang Yang, Shiyu Chang, and Pan Zhou. Jointly attentive spatial-temporal pooling networks for video-based person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4733–4742, 2017.

[39] Yichao Yan, Bingbing Ni, Zhichao Song, Chao Ma, Yan Yan, and Xiaokang Yang. Person re-identification via recurrent feature aggregation. In *European Conference on Computer Vision*, pages 701–716. Springer, 2016.

[40] Chaojian Yu, Xinyi Zhao, Qi Zheng, Peng Zhang, and Xinge You. Hierarchical bilinear pooling for fine-grained visual recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 574–589, 2018.

[41] Amir Roshan Zamir, Afshin Dehghan, and Mubarak Shah. Gmcp-tracker: Global multi-object tracking using generalized minimum clique graphs. In *European Conference on Computer Vision*, pages 343–356. Springer, 2012.

[42] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1116–1124, 2015.

[43] Liang Zheng, Yi Yang, and Alexander G Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016.

[44] Liang Zheng, Hengheng Zhang, Shaoyan Sun, Manmohan Chandraker, Yi Yang, and Qi Tian. Person re-identification in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1367–1376, 2017.

[45] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4747–4756, 2017.