SCIENCE CHINA
Information Sciences

• **Supplementary File** •

A Spatial Structural Similarity Triplet Loss for Auxiliary Vehicle Re-identification

Jianqing Zhu[1], Liu Liu[2], Xiaobin Zhu[3] & Huanqiang Zeng[4*]

[1]*College of Engineering, Huaqiao University, Quanzhou* 362021*, China;*
[2]*UBTECH Sydney AI Centre, School of Computer Science FEIT, University of Sydney, Australia;*
[3]*Department of Computer Science and Technology, University of Science and Technology Beijing,* 100083*, China;*
[4]*College of Information Science and Engineering, Huaqiao University, Xiamen,* 361021*, China*

Appendix A Loss Functions

The proposed method applies three loss functions, i.e., the spatial structural similarity triplet loss, the Euclidean distance based triplet loss [13], and label smooth regularized softmax loss [59]. More details are presented as follows.

Appendix A.1 Spatial Structural Similarity Triplet Loss

Given a pair of three-dimensional feature map $X, Y \in R^{h\times w\times c}$ respectively extracted from a pair of vehicle images via a deep network, where h, w, and c respectively represent the height, width, and channel, the SSIM between X and Y is calculated as follows:

$$SSIM(X,Y) = \frac{1}{c}\frac{1}{u}\frac{1}{v}\sum_{k=1}^{c}\sum_{i=1}^{u}\sum_{j=1}^{v}\frac{2\mu_{X_{i,j,k}}\mu_{Y_{i,j,k}}+\tau_1}{\mu_{X_{i,j,k}}^2+\mu_{Y_{i,j,k}}^2+\tau_1}\frac{2\sigma_{X_{i,j,k}}\sigma_{Y_{i,j,k}}+\tau_2}{\sigma_{X_{i,j,k}}^2+\sigma_{Y_{i,j,k}}^2+\tau_2}\frac{2\sigma_{X_{i,j,k}Y_{i,j,k}}+\tau_3}{2\sigma_{X_{i,j,k}}\sigma_{Y_{i,j,k}}+\tau_3}, \tag{A1}$$

where $X_{i,j,k}$ and $Y_{i,j,k}$ are $n\times n$ sized patches of X and Y lied at i-th row, j-th column and k-th channel, which are located by pixel-by-pixel sliding over the k-th channel feature map; u and v are sliding ranges of the height and width dimensions, respectively; $\tau_1 > 0$, $\tau_2 > 0$, and $\tau_3 > 0$ are tiny constants used to avoid the numerical instability. In addition, $\mu_{X_{i,j,k}}$, $\mu_{Y_{i,j,k}}$, $\sigma^2_{X_{i,j,k}}$, $\sigma^2_{Y_{ijk}}$ and $\sigma_{X_{i,j,k}Y_{i,j,k}}$, can be calculated as follows:

$$\begin{aligned}
\mu_{X_{i,j,k}} &= \sum_{p=1}^{r}\sum_{q=1}^{r} g_{p,q}X_{i+p-1,j+q-1,k},\\
\mu_{Y_{i,j,k}} &= \sum_{p=1}^{r}\sum_{q=1}^{r} g_{p,q}Y_{i+p-1,j+q-1,k},\\
\sigma^2_{X_{i,j,k}} &= \sum_{p=1}^{r}\sum_{q=1}^{r} g_{p,q}(X_{i+p-1,j+q-1,k}-\mu_{X_{i,j,k}})^2,\\
\sigma^2_{Y_{i,j,k}} &= \sum_{p=1}^{r}\sum_{q=1}^{r} g_{p,q}(Y_{i+p-1,j+q-1,k}-\mu_{Y_{i,j,k}})^2,\\
\sigma_{X_{i,j,k}Y_{i,j,k}} &= \sum_{p=1}^{r}\sum_{q=1}^{r} g_{p,q}(X_{i+p-1,j+q-1,k}-\mu_{X_{i,j,k}})(Y_{i+p-1,j+q-1,k}-\mu_{Y_{i,j,k}}),
\end{aligned} \tag{A2}$$

where r denotes the local window size and it is set to 7 in this paper; $g = \{g_{p,q}|p,q = 1,2,3,...,r\}$ is a $r\times r$ circular-symmetric Gaussian weighting function with a standard deviation of 1.5, and g is normalized to a unit sum (i.e., $\sum_{p=1}^{r}\sum_{q=1}^{r} g_{p,q} = 1$) according to [51].

* Corresponding author (email: zeng0043@hqu.edu.cn)

It can be seen that the Eq. (A1) consists of three components, i.e., the luminance, contrast, and structure similarity measurements of X and Y. In order to simply Eq. (A1), τ_3 is set equal to τ_2 in [51]. Following the same practice, the calculation of the SSIM (i.e., Eq. (A1)) can be simplified as follows:

$$SSIM(X,Y) = \frac{1}{c}\frac{1}{u}\frac{1}{v}\sum_{k=1}^{c}\sum_{i=1}^{u}\sum_{j=1}^{v}\frac{2\mu_{X_{i,j,k}}\mu_{Y_{i,j,k}}+\tau_1}{\mu^2_{X_{i,j,k}}+\mu^2_{Y_{i,j,k}}+\tau_1}\frac{2\sigma_{X_{i,j,k}Y_{i,j,k}}+\tau_2}{\sigma^2_{X_{i,j,k}}+\sigma^2_{Y_{i,j,k}}+\tau_2}, \tag{A3}$$

where τ_1 and τ_2 are set to 1×10^{-4} and 9×10^{-4} according to [51]. Furthermore, incorporating the SSIM formulation (i.e., Eq. (A3)), the SSIM based triplet loss is formulated as:

$$L_{SSIM}(X_a,X_n,X_p) = -\log(1+e^{SSIM(X_a,X_p)-SSIM(X_a,X_n)}), \tag{A4}$$

where (X_a,X_n,X_p) is a training triple and each data of a training triple is $h\times w\times c$ sized feature map; (X_a,X_n) is a negative pair (i.e., two vehicle images of different class labels), while (X_a,X_p) is a positive pair (i.e., two vehicle images of the same class label). The minimization of Eq. (A4) aims to push the $SSIM(X_a,X_p)$ as large as possible, and pull the $SSIM(X_a,X_n)$ as small as possible. As a result, the spatial discrimination of feature maps resulted from a backbone network can be reserved for vehicle re-identification.

In practice, many existing works (e.g., [13,24,37]) have shown that the hard sample mining strategy of finding the most difficult positive and negative image pairs in each mini-batch can improve the feature discriminative ability. Hence, the hard sample mining strategy is also applied in this paper. However, given a mini-batch containing dozens or even hundreds of vehicle images, the SSIM hard sample mining encounters a massive computation cost that totally involves five loops (the first two loops are for pairing images and the rest three loops are for calculating SSIMs of image pairs). Although those five loops can be easily parallel executed, large graph processing unit (GPU) memories will be required. Therefore, it is very necessary to design an executive-friendly SSIM based triplet loss.

From Eq. (A3), it can be found that the SSIM of a pair of $h\times w\times c$ sized feature maps is essentially equal to the average of SSIMs independently calculated on each single-channel feature map. From Eq. (A2), it can be further observed that the Gaussian weighting function (i.e., g in Eq. (A2)) used for the SSIM calculation on each single-channel feature map is identical. Hence, the final spatial structural similarity (S^3) triplet loss is designed, which simplifies Eq. (A3) by performing a channel global average pooling (CGAP) operation on $h\times w\times c$ sized feature maps in advance, as follows:

$$L_{S^3}(X_a,X_n,X_p) = -\log(1+e^{SSIM(CGAP(X_a),CGAP(X_p))-SSIM(CGAP(X_a),CGAP(X_n))}). \tag{A5}$$

where $CGAP(\cdot)$ represents a global average pooling operation according to the channel dimension. Given a $h\times w\times c$ sized feature map X, the calculation of CGAP is formulated as follows:

$$CGAP(X) = \begin{pmatrix} \frac{1}{c}\sum_{k=1}^{c}X_{1,1,k} & \frac{1}{c}\sum_{k=1}^{c}X_{1,2,k} & \cdots & \frac{1}{c}\sum_{k=1}^{c}X_{1,w,k} \\ \frac{1}{c}\sum_{k=1}^{c}X_{2,1,k} & \frac{1}{c}\sum_{k=1}^{c}X_{2,2,k} & \cdots & \frac{1}{c}\sum_{k=1}^{c}X_{2,w,k} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{1}{c}\sum_{k=1}^{c}X_{h,1,k} & \frac{1}{c}\sum_{k=1}^{c}X_{h,2,k} & \cdots & \frac{1}{c}\sum_{k=1}^{c}X_{h,w,k} \end{pmatrix}. \tag{A6}$$

The CGAP operation transforms a $h\times w\times c$ sized feature map into a $h\times w$ sized compressed spatial feature map. Consequently, based on the S^3 triplet loss (i.e., Eq. (A5)), the hard sample mining strategy only involves four loops, which only spends about $1/c$ computation cost of that of the hard sample mining strategy using Eq. (A4). For off-the-shelf deep networks (e.g., VGGNet, ResNet [11], and GoogLeNet [47]), c usually is an integer of hundreds or even thousands. Therefore, based on the S^3 triplet loss (i.e., Eq. (A5)), the hard sample mining strategy's computation cost is significantly reduced, so that its execution is greatly improved.

Appendix A.2 Euclidean Distance Based Triplet Loss

The Euclidean distance (ED) triplet loss function play on the spatial global average pooling (SGAP) layer is defined as follows:

$$L_{ED}(X_a,X_n,X_p) = -\log(1+e^{\|SGAP(X_a)-SGAP(X_n)\|_2-\|SGAP(X_a)-SGAP(X_p)\|_2}), \tag{A7}$$

where (X_a,X_n,X_p) is a training triple and each of its data is represented by $h\times w\times c$ (e.g., $16\times16\times2048$) sized feature map extracted by using the backbone (ResNet-50-IBN-a [39]); (X_a,X_n) is a negative pair, while (X_a,X_p) is a positive pair; $\|\cdot\|_2$ denotes an Euclidean distance; $SGAP(\cdot)$ is formulated as follows:

$$SGAP(X) = \frac{1}{h\times w}\left(\sum_{i=1}^{h}\sum_{j=1}^{w}X_{i,j,1},\ \sum_{i=1}^{h}\sum_{j=1}^{w}X_{i,j,2},\ \ldots,\ \sum_{i=1}^{h}\sum_{j=1}^{w}X_{i,j,c}\right). \tag{A8}$$

Recall the S^3 triplet loss of Eq. (A5) and the Euclidean distance (ED) triplet loss of Eq. (A7), one can see that the S^3 triplet loss aims to reserve spatial discrimination of feature maps based on CGAP and SSIM, while the ED triplet loss focuses on learning channel discrimination of feature maps based on SGAP and ED. As a result, the S^3 triplet loss and the ED triplet loss are complementary to each other.

Appendix A.3 Label Smooth Regularized Softmax Loss

Following the FC-BN-DT block, the label smooth regularized softmax (LSRS) [59] is formulated as follows:

$$L_{LSRS}(F,l) = -\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{s}\delta(j,l_i)log(\frac{e^{W_j^{\mathrm{T}}F_i}}{\sum_{k=1}^{s}e^{W_k^{\mathrm{T}}F_i}}), \tag{A9}$$

where m is number of training samples and s is the number of classes; $F = \{F_i|i = 1, 2, ..., m\}$ are training samples represented via the FC-BN-DT block producing features and $l = \{l_i \in \{1, 2, 3, ..., s\}|i = 1, 2, ..., m\}$ are the class labels; $W = [W_1, W_2, W_3, ..., W_s]$ is a learn-able parameter matrix; $\delta(j, l_i)$ is a label smooth regularized indicator function, which is defined as follows:

$$\delta(j,l_i) = \begin{cases} 1-\varepsilon+\frac{\varepsilon}{s}, & j = l_i, \\ \frac{\varepsilon}{s}, & otherwise, \end{cases} \tag{A10}$$

where $\varepsilon \in [0, 1)$ is a tiny constant for controlling the degree of smoothing regularization, which is set to 0.1 in this paper, as done in [59]. Base on Eq. (A5), Eq. (A7), and Eq. (A9), the total loss function for the proposed S^{3}ANet is formulated as follows:

$$L_{Total} = \alpha L_{S^3} + \beta L_{ED} + \gamma L_{LSRS}, \tag{A11}$$

where $\alpha \geqslant 0$, $\beta \geqslant 0$, and $\gamma \geqslant 0$ are manually setting constants used to keep the balance of these three loss functions. In order to avoid excessive tuning these constants, we set $\alpha = \beta = \gamma = 1$ in following experiments. The combination of L_{ED} and L_{LSRS} is common, however, L_{S^3} is a newly designed triplet loss in this paper, which can reserve the spatial discrimination of feature maps to improve vehicle re-identification. In our method, the FC-BN-DT block is composed of a fully connected (FC) layer, a batch normalization (BN) [16] layer, a Dropout (DT) [45] layer. The FC layer linearly compresses features into a moderate number of dimensions (i.e., 512). The DT layer holding a dropout ratio of 0.5 is used to reduce the over-fitting risk.

Appendix B Experiments and Analysis

To validate the superiority of the proposed spatial structural similarity triplet loss auxiliary deep network (S^{3}ANet), it is compared with state-of-the-art vehicle re-identification approaches on two large scale datasets, namely, VehicleID [29] and VeRi776 [33]. The cumulative match characteristic (CMC) curve [7], rank-1 identification rate (R1) [32, 67], and mean average precision (mAP) [58, 65, 66] are used to assess the vehicle re-identification performance.

Appendix B.1 Implementation Details

The deep learning toolbox is Pytorch [40]. ResNet-50-IBN-a [39] is pre-trained on ImageNet [23]. Training configurations are summarized as follows. (1) Vehicle images are uniformly resized into a 256×256 resolution and then 10-pixel zero padding, random cropping, horizontal flip, random erasing [60], and z-score normalization are used for the data augmentation. The probabilities of the horizontal flip and random erasing are set to 0.5, and the mean and standard deviation are $[0.485, 0.456, 0.406]$ and $[0.229, 0.224, 0.225]$, respectively. (2) The mini-batch stochastic gradient descent (SGD) method [23] is applied to train parameters. The weight decays are set to 5×10^{-4} and the momentums are set to 0.9. For the VeRi776 dataset, there are 70 epochs during the training process, while for the larger dataset, i.e., VehicleID, there are 100 epochs. To be more specific, for both VeRi776 and VehicleID datasets, the learning rates are initialized to 2×10^{-4}, and are linearly warmed up to 2×10^{-2} in the first 5 epochs. Then, the learning rates are dropped to 1/10 of their old values at certain intervals. For the VeRi776 dataset, the learning rate dropping interval is set to 20, while for the VehicleID dataset, it is set to 30, considering that the VehicleID dataset is much larger than the VeRi776 dataset. (3) Each mini-batch includes 8 subjects and each subject holds 8 images.

Features respectively generated by the CGAP component and the FC-BN-DT component (see Figure **??**) are concatenated as final features, and both CGAP and Dropout generating features are independently l_2 normalized before concatenating. The Euclidean distance of final features is applied as the similarity measurement for a convenient testing process.

Appendix B.2 Comparisons with State-of-the-art Methods

Appendix B.2.1 *Comparison on VehicleID*

Table B1 shows the performance comparison of the proposed S^{3}ANet method and state-of-the-art approaches on the VehicleID [29] dataset. It can be found that the proposed S^{3}ANet method obtains the best performance. More detail comparisons are presented as follows.

First, it can be seen that the proposed S^{3}ANet method obtains better results than those approaches [4,9,27,30,38,42,50,65] that perform some moderate spatial pooling strategies. For example, compared with the QD-DLF [65] method using quadruple directional pooling layers, the proposed S^{3}ANet method acquires about 10% higher in terms of both rank-1 identification rate and mAP on all the three testing subsets. Compared with the PRN [4] method that divides feature maps according to the height, width, and channel dimensions, the proposed S^{3}ANet method outperforms it by a 6.39% higher rank-1 identification rate on the largest Test2400 subset.

Second, the proposed S^{3}ANet method outperforms those multi-model approaches [6, 8, 10, 12, 17, 20, 34, 44, 50, 54, 57] requiring extra vehicle attributes. Specifically, the proposed S^{3}ANet method defeats the outstanding multi-modal method

Table B1 The performance (%) comparison of the proposed method (i.e., S^{3}ANet) and state-of-the-art approaches on the VehicleID dataset. The red, green and blue rows respectively represent the first-, second- and third- best results in terms of the rank-1 identification rate.

Methods	Test800		Test1600		Test2400		References
	R1	mAP	R1	mAP	R1	mAP	
S^{3}ANet	84.16	86.87	80.25	83.10	77.97	80.91	Proposed
QD-DLF [65]	72.32	76.54	70.66	74.63	64.14	68.41	IEEE ITS 2020
MGL [55]	79.6	82.1	76.2	79.6	73.0	75.5	ICIP 2019
Appearance+License Plate [12]	79.5	82.7	76.9	79.9	74.8	77.7	ICIP 2019
PRN [4]	78.92	N/A	74.94	N/A	71.58	N/A	CVPRW 2019
MOV1+BS [49]	N/A	N/A	N/A	N/A	69.3	78.2	CVPR 2019
Triplet Embedding [24]	78.80	86.19	73.41	81.69	69.33	78.16	IJCNN 2019
Part Regularization [10]	78.40	N/A	75.00	N/A	74.20	N/A	CVPR 2019
GRF+GGL [31]	77.1	N/A	72.7	N/A	70.0	N/A	IEEE TIP 2019
MRM [41]	76.64	80.02	74.20	77.32	70.86	74.02	Neurocomputing 2019
XG-6-sub-multi [57]	76.1	N/A	73.1	N/A	71.2	N/A	IEEE ITS 2019
EALN [35]	75.11	77.5	71.78	74.2	69.30	71.0	IEEE TIP 2019
MSV [53]	75.1	79.3	71.8	75.4	68.7	73.3	ICASSP 2019
DQAL [15]	74.74	N/A	71.01	N/A	68.23	N/A	IEEE TVT 2019
ResNet101-AAVER [22]	74.69	N/A	68.62	N/A	63.54	N/A	ICCV 2019
Mob.VFL [1]	73.37	N/A	69.52	N/A	67.41	N/A	ICIP 2019
TAMR [9]	66.02	N/A	62.90	N/A	59.69	N/A	IEEE TIP 2019
MLSR [14]	65.78	N/A	64.24	N/A	60.05	N/A	Neurocomputing 2019
RPM [38]	65.04	N/A	62.55	N/A	60.21	N/A	ICMEW 2019
SFF+SAtt [27]	64.50	N/A	59.12	N/A	54.41	N/A	IJCNN 2019
PRN [4]	63.07	N/A	55.42	N/A	50.36	N/A	CVPRW 2019
FDA-Net [36]	N/A	N/A	59.84	65.33	55.53	61.84	CVPR 2019
GSTE [2]	75.90	75.40	74.80	74.30	74.00	72.40	IEEE TMM 2018
RAM [30]	75.20	N/A	72.3	N/A	67.70	N/A	ICME 2018
MAD+STR [17]	N/A	82.20	N/A	75.90	N/A	72.80	ICIP 2018
VAMI [61]	63.12	N/A	52.87	N/A	47.34	N/A	CVPR 2018
C2F [8]	61.10	63.50	56.20	60.00	51.40	53.00	AAAI 2018
MSVF [21]	N/A	N/A	N/A	N/A	46.61	N/A	GCPR 2018
ResNet-18+PMSM [46]	N/A	64.2	N/A	57.2	N/A	51.8	ICPR 2018
SDC-CNN [66]	56.98	63.52	50.57	57.07	42.92	49.68	ICPR 2018
ABLN-32 [63]	52.63	N/A	N/A	N/A	N/A	N/A	WACV 2018
NuFACT [34]	48.90	N/A	43.64	N/A	38.63	N/A	IEEE TMM 2018
DJDL [25]	72.30	N/A	70.80	N/A	68.00	N/A	ICIP 2017
Improved Triplet CNN [56]	69.90	N/A	66.20	N/A	63.20	N/A	ICME 2017
OIFE [50]	N/A	N/A	N/A	N/A	67.00	N/A	ICCV 2017
Multi-grain Ranking [54]	N/A	62.80	N/A	62.30	N/A	58.60	ICCV 2017
CLVR [20]	62.00	N/A	56.10	N/A	50.60	N/A	BMVC 2017
FACT [32]	49.53	N/A	44.63	N/A	39.91	N/A	ICME 2016
DRDL [29]	48.91	N/A	46.36	N/A	40.97	N/A	CVPR 2016

(i.e., Appearance+License Plate [12]) on all the three testing subsets. For example, on the Test2400 subset, the proposed S^{3}ANet method beats the Appearance+License Plate [12] method by a 3.17% higher rank-1 identification rate and a 3.21% larger mAP.

Third, the proposed S^{3}ANet method is superior to those methods [2, 5, 9, 24–26, 29, 31, 56] jointly using softmax and variant triplet loss functions, such as multi-grain learning (MGL) [55], GSTE [2], Triplet Embedding [24], GRF+GGL [31]. Among these methods, MGL [55] obtains best results, but it is consistently defeated by the proposed S^{3}ANet method on all the three testing subsets. For example, on the largest Test2400 subset, the rank-1 identification rate and mAP of the proposed S^{3}ANet method are 4.97% and 5.41% higher than those of MGL [55], respectively.

At last, the proposed S^{3}ANet method is also stronger than the two multi-scale vehicle re-identification methods (i.e., MSV [53] and MSVR [21]) and the two adversarial learning based approaches (i.e., ABLN-32 [63] and EALN [35]) with higher rank-1 identification rates and mAPs on all the three testing subsets.

Appendix B.2.2 *Comparison on VeRi776*

From Table B2, it is interesting to find that the vehicle re-identification performance has made significant progresses since 2016. Especially in 2019, many approaches [3, 10, 12, 19, 24, 27, 41, 43, 48, 49] acquire more than 90% rank-1 identification rates. Under this condition, the proposed S^{3}ANet method still obtains the best result (i.e., 96.60% rank-1 identification rate and 78.52% mAP) and outperforms those state-of-the-art methods under comparison. To be more specific, the proposed S^{3}ANet method respectively outperforms the second-best method (i.e., Appearance+License Plate [12]) and third-best

Table B2 The performance (%) comparison of the proposed method (i.e., S^{3}ANet) and state-of-the-art approaches on the VeRi776 dataset. The red, green and blue rows respectively represent the first-, second- and third- best results in terms of the rank-1 identification rate (R1).

Methods	R1	mAP	References
S^{3}ANet	96.60	78.52	Proposed
QD-DLF [65]	88.50	61.83	IEEE ITS 2020
Appearance+License Plate [12]	95.41	78.08	ICIP 2019
SFF+SAtt+TBR [27]	94.93	74.11	IJCNN 2019
Part Regularization [10]	94.30	74.30	CVPR 2019
PAMTRI [48]	92.86	71.88	ICCV 2019
MLFN+Triplet [43]	92.55	71.78	CVPRW 2019
MTML+OSG+re-ranking [19]	92.0	68.3	CVPRW 2019
MRM [41]	91.77	68.55	Neurocomputing 2019
DMML [3]	91.2	70.1	ICCV 2019
Triplet Embedding [24]	90.23	67.55	IJCNN 2019
MOV1+BS [49]	90.2	67.6	CVPR 2019
VANet [5]	89.78	66.34	ICCV 2019
GRF+GGL [31]	89.4	61.7	IEEE TIP 2019
ResNet101-AAVER [22]	88.97	61.18	ICCV 2019
MRL [26]	87.7	71.4	ICME 2019
Fusion-Net [18]	87.31	62.40	IEEE TIP 2019
Mob.VFL-LSTM [1]	87.18	58.08	ICIP 2019
MGL [55]	86.1	65.0	ICIP 2019
EALN [35]	84.39	57.44	IEEE TIP 2019
JDRN+ re-ranking [28]	N/A	73.10	CVPRW 2019
FDA-Net [36]	49.43	N/A	CVPR 2019
MAD+STR [17]	89.27	61.11	ICIP 2018
RAM [30]	88.60	61.50	ICME 2018
VAMI [61]	85.92	61.32	CVPR 2018
GSTE [2]	N/A	59.47	IEEE TMM 2018
SDC-CNN [66]	83.49	53.45	ICPR 2018
PROVID [34]	81.56	53.42	IEEE TMM 2018
NuFACT + Plate-SNN [34]	81.11	50.87	IEEE TMM 2018
SCCN-Ft+CLBL-8-Ft [62]	60.83	25.12	IEEE TIP 2018
ABLN-Ft-16 [63]	60.49	24.92	WACV 2018
NuFACT [34]	76.76	48.47	IEEE TMM 2018
VST Path Proposals [44]	83.49	58.27	ICCV 2017
OIFE+ST [50]	68.30	51.42	ICCV 2017
FACT [32]	52.21	18.75	ICME 2016
VGG-CNN-M-1024 [29]	44.10	12.76	CVPR 2016

method (i.e., SFF+SAtt+TBR [27]) by 1.19% and 1.67% higher rank-1 identification rates, and 0.44% and 4.41% larger mAPs.

In summary, it can be clearly concluded from the results shown in Tables B1 and B2 that the proposed S^{3}ANet method can outperform a lot of state-of-the-art methods. This is mainly credited to that the proposed S^{3}ANet method specifically reserves the spatial discrimination of feature maps. Notably, the proposed S^{3}ANet method is elegant, which does not require the extra attribute annotations (e.g., license plates, landmarks/key-points, colors, models/types and spatial-temporal information) or the division of vehicles into several parts.

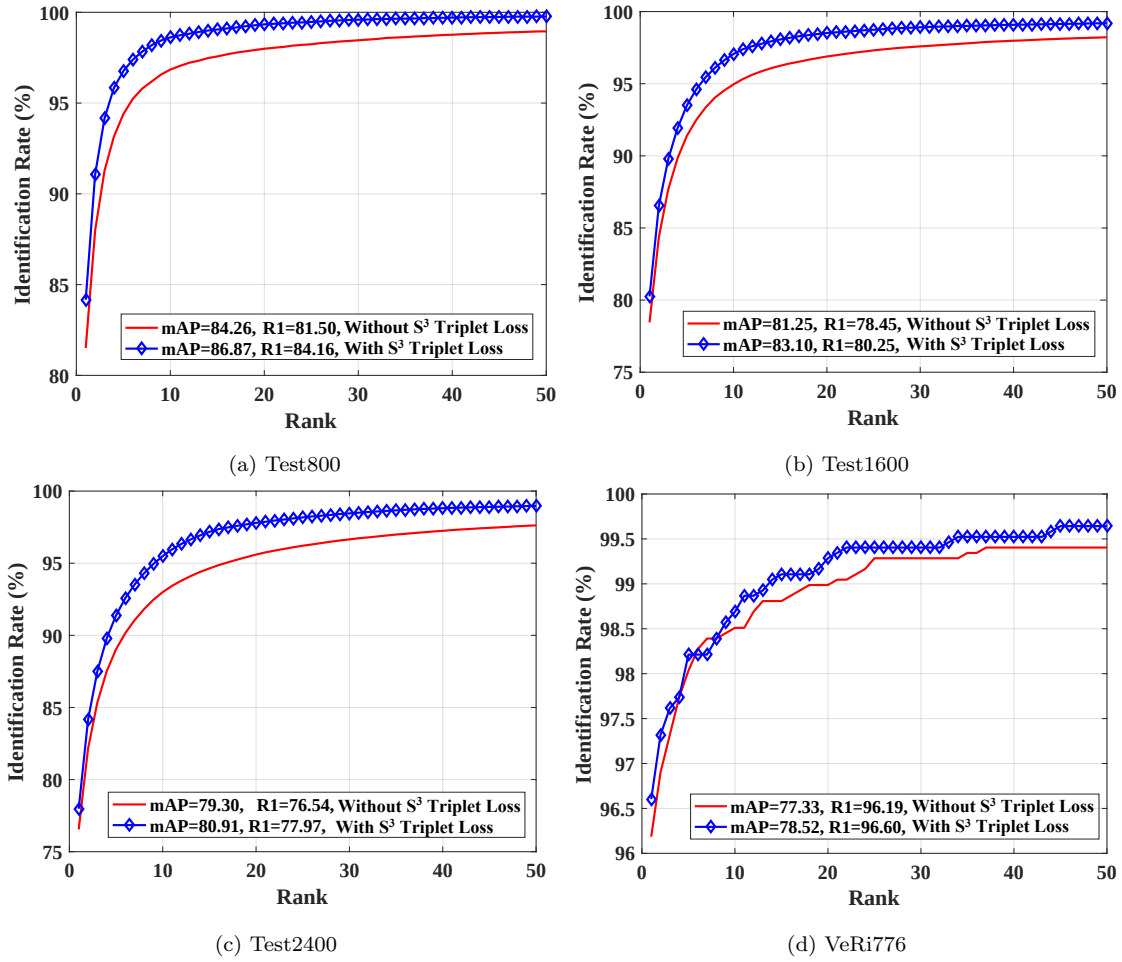
Appendix B.3 Analysis

Appendix B.3.1 *Role of S^3 Triplet Loss*

To demonstrate the role of the S^3 triplet loss, the performance comparison of the proposed S^{3}ANet method with and without the S^3 triplet loss is conducted. The corresponding comparisons are shown in Figure B1.

Firstly, as shown in Figure B1, even without using the S^3 triplet loss, the proposed method can outperform most of state-of-the-art methods listed in Table B1 and Table B2, which is because of the powerful feature learning ability of ResNet-50-IBN-a [39] and the reasonable joint combination of the Euclidean distance (ED) triplet loss and the label smooth regularized softmax (LSRS) loss.

Secondly, it can be seen from Figure B1 (a), Figure B1 (b), and Figure B1 (c) that the proposed S^{3}ANet method with the S^3 triplet loss is superior to the one without the S^3 triplet loss on three testing subsets (i.e., Test800, Test1600, and Test2400) of the VehicleID dataset. For example, on the largest Test2400 subset, the mAP and rank-1 identification rate of the proposed S^{3}ANet method with the S^3 triplet loss are individually 1.61% and 1.43% higher than those of the one without the S^3 triplet loss. Also, a similar conclusion can be obtained on on the VeRi776 dataset, as shown in Figure B1



(a) Test800 (b) Test1600

(c) Test2400 (d) VeRi776

Figure B1 The performance comparison between the proposed method with and without the S^3 triplet loss on the VehicleID and VeRi776 dataset, where R1 denotes the rank-1 identification rate.

(d). For example, the proposed S^3ANet method with the S^3 triplet loss is superior to the one without the S^3 triplet loss on the VeRi776 dataset (e.g., by a larger 1.19% mAP). All these demonstrate that the proposed S^3 triplet loss that reserves spatial discrimination of feature maps is indeed beneficial to improving the performance of vehicle re-identification.

Appendix B.3.2 *Impact of Different Feature Map Pooling Strategies*

It is interesting to investigate the impact to the performance using different feature map pooling strategies. Specifically, based on the same backbone network, two spatial division pooling strategies, i.e., vertically four quartering (V4) and horizontally four quartering (H4), by following the similar practice in [4,30,64,65] are implemented and compared with the proposed S^3ANet method. For both V4 and H4, each divided part is handled with its own Euclidean distance (ED) triple loss and label smooth regularized softmax (LSRS) loss, and these two losses are with the same configuration as that of the proposed S^3ANet. Moreover, during the testing phase, features from each divided part are independently l_2 normalized and concatenated to obtain 2048-dimensional features for vehicle re-identification. The corresponding results are listed in Table B3.

From Table B3, it can be observed that the proposed S^3ANet method outperforms both V4 and H4. For example, on the VeRi776 dataset, the rank-1 identification rate of the proposed S^3ANet is 1.55% and 2.32% higher than that of V4 and H4, respectively. In addition, the mAP of the proposed S^3ANet method is 2.84% and 5.10% larger than that of V4 and

Table B3 The performance (%) comparison of different feature map pooling strategies.

Strategies	Test800		Test1600		Test2400		VeRi776	
	Rank=1	mAP	Rank=1	mAP	Rank=1	mAP	Rank=1	mAP
S^3ANet	**84.16**	**86.87**	**80.25**	**83.10**	**77.97**	**80.91**	**96.60**	**78.52**
V4	82.84	85.60	78.79	81.81	76.63	79.65	95.05	75.68
H4	80.78	83.75	77.96	80.94	75.00	78.06	94.28	73.42

H4, respectively. These results illustrate that the proposed S^3 triplet loss auxiliary pooling method is more effective than spatial division pooling strategies.

References

1 Saghir Alfasly S A, Hu Y J, Liang T C, et al. Variational representation learning for vehicle re-identification. In: Proceedings of International Conference on Image Processing, Taipei, China, 2019. 3118–3122
2 Bai Y, Lou Y H, Gao F, et al. Group-sensitive triplet embedding for vehicle re-identification. IEEE Trans on Multimedia, 2018, 20(9): 2385–2399
3 Chen G Y, Zhang T R, Lu J W, et al. Deep meta metric learning. In: Proceedings of International Conference on Computer Vision, Seoul, Korea, 2019. 9547–9556
4 Chen H, Lagadec B, Bremond F. Partition and reunion: A two-branch neural network for vehicle re-identification. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, USA, 2019. 184–192
5 Chu R H, Sun Y F, Li Y D, et al. Vehicle re-identification with viewpoint-aware metric learning. In: Proceedings of International Conference on Computer Vision, Seoul, Korea, 2019. 8282–8291
6 Cui C, Sang N, Gao C X, et al. Vehicle re-identification by fusing multiple deep neural networks. In: Proceedings of IEEE International Conference on Image Processing Theory, Tools and Applications, Montreal, Canada, 2017. 1–6
7 Gray D, Brennan S, Tao H. Evaluating appearance models for recognition, reacquisition, and tracking. In: Proceedings of IEEE International Workshop on Performance Evaluation of Tracking and Surveillance, Rio DE Janeiro, Brazil, 2007. 1–7
8 Guo H Y, Zhao C Y, Liu Z W, at al. Learning coarse-to-fine structured feature embedding for vehicle re-identification. In: Proceedings of AAAI Conference on Artificial Intelligence, Hilton New Orleans Riverside, New Orleans, Louisiana, USA, 2018. 6853–6860
9 Guo H Y, Zhu K, Tang M, et al. Two-level attention network with multi-grain ranking loss for vehicle re-identification. IEEE Trans on Image Processing, 2019, 28(9): 4328–4338
10 He B, Li J, Zhao Y F, et al. Part-regularized near-duplicate vehicle re-identification. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, 2019. 3997–4005
11 He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016. 770–778
12 He Y G, Dong C H, Wei Y. Combination of appearance and license plate features for vehicle re-identification. In: Proceedings of International Conference on Image Processing, Taipei, China, 2019. 3108–3112
13 Hermans A, Beyer L, Leibe B. In defense of the triplet loss for person re-identification. arXiv preprint, 2017, arXiv:1703.07737
14 Hou J H, Zeng H Q, Cai L, et al. Multi-label learning with multi-label smoothing regularization for vehicle re-identification. Neurocomputing, 2019, 345: 15–22
15 Hou J H, Zeng H Q, Zhu J Q, et al. Deep quadruplet appearance learning for vehicle re-identification. IEEE Trans on Vehicular Technology, 2019, 68(9): 8512–8522
16 Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: Proceedings of International Conference on Machine Learning, Lille, France, 2015. 448–456
17 Jiang N, Xu Y, Zhou Z, et al. Multi-attribute driven vehicle re-identification with spatial-temporal re-ranking. In: Proceedings of International Conference on Image Processing, Athens, Greece, 2018. 858–862
18 Kan S C, Cen Y G, He Z H, et al. Supervised deep feature embedding with hand crafted feature. IEEE Trans on Image Processing, 2019, 28(12): 5809–5823
19 Kanacı A, Li M X, Gong S G, et al. Multi-task mutual learning for vehicle re-identification. In: Proceedings of IEEE Computer Vision and Pattern Recognition Workshops, Long Beach, USA, 2019. 62–70
20 Kanacı A, Zhu X T, Gong S G. Vehicle re-identification by fine-grained cross-level deep learning. In: Proceeding of British Machine Vision Conference Activity Monitoring by Multiple Distributed Sensing Workshop, London, United Kingdom, 2017. 772–788
21 Kanacı A, Zhu X T, Gong S G. Vehicle re-identification in context. In: Proceedings of German Conference on Pattern Recognition, Stuttgart, Germany, 2018. 377–390
22 Khorramshahi P, Kumar A, Peri N, et al. A dual-path model with adaptive attention for vehicle re-identification. In: Proceedings of International Conference on Computer Vision, Seoul, Korea, 2019. 6132–6141
23 Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks. In: Proceedings of Annual Conference on Neural Information Processing Systems, Stateline, USA, 2012. 1097–1105
24 Kumar R, Weill E, Aghdasi F, et al. Vehicle re-identification: an efficient baseline using triplet embedding. In: Proceedings of International Joint Conference on Neural Networks, Budapest, Hungary, 2019. 1–9.
25 Li Y Q, Li Y H, Yan H F, et al. Deep joint discriminative learning for vehicle re-identification and retrieval. In: Proceedings of International Conference on Image Processing, Beijing, China, 2017. 395–399
26 Lin W P, Li Y D, Yang X L, et al. Multi-view learning for vehicle re-identification. In: Proceedings of International Conference on Multimedia and Expo, Shanghai, China, 2019. 832–837
27 Liu C H, Huynh D Q, Mark Reynolds M. Urban area vehicle re-identification with self-attention stair feature fusion and temporal Bayesian re-ranking. In: Proceedings of International Joint Conference on Neural Networks, Budapest, Hungary, 2019. 1–8

28 Liu C T, Lee M Y, Wu C W, et al. Supervised joint domain learning for vehicle re-identification. In: Proceedings of IEEE Computer Vision and Pattern Recognition Workshops, Long Beach, USA, 2019. 45–52

29 Liu H Y, Tian Y H, Wang Y W, et al. Deep relative distance learning: Tell the difference between similar vehicles. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016. 2167–2175

30 Liu X B, Zhang S L, Huang Q M, et al. Ram: a region-aware deep model for vehicle re-identification. In: Proceedings of IEEE International Conference on Multimedia and Expo, San Diego, USA, 2018. 1–6

31 Liu X B, Zhang S L, Wang X Y, et al. Group-group loss based global-regional feature learning for vehicle re-identification. IEEE Trans. on Image Processing, 2019, 29: 2638–2652

32 Liu X C, Liu W, Ma H D, et al. Large-scale vehicle re-identification in urban surveillance videos. In: Proceedings of International Conference on Multimedia and Expo, Seattle, USA, 2016. 1–6

33 Liu X C, Liu W, Mei T, et al. A deep learning-based approach to progressive vehicle re-identification for urban surveillance. In: Proceedings of European Conference on Computer Vision, Amsterdam, The Netherlands, 2016. 869–884

34 LiU X C, Liu W, Mei T, et al. Provid: Progressive and multi-modal vehicle re-identification for large-scale urban surveillance. IEEE Transactions on Multimedia, 2018, 20(3): 645–658.

35 Lou Y H, Bai Y, Liu J, et al. Embedding adversarial learning for vehicle re-identification. IEEE Trans on Image Processing, 2019, 28(8): 3794–3807

36 Lou Y H, Bai Y, Liu J, et al. Veri-wild: A large dataset and a new method for vehicle re-identification in the wild. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, USA, 2019. 3235–3243

37 Luo H, Jiang W, Gu Y Z, et al. A strong baseline and batch normalization neck for deep person re-identification. IEEE Trans on Multimedia (Early Access), 2019

38 Ma X, Zhu K, Guo H Y, et al. Vehicle re-identification with refined part model. In: Proceedings of IEEE International Conference on Multimedia and Expo Workshops, Shanghai, China, 2019. 603–606

39 Pan X G, Luo P, Shi J P, et al. Two at once: Enhancing learning and generalization capacities via ibn-net. In: Proceedings of European Conference on Computer Vision, Munich, Germany, 2018. 464–479

40 Paszke A, Gross S, Massa F, et al. Pytorch: An imperative style, high-performance deep learning library. In: Proceedings of Advances in Neural Information Processing Systems, Vancouver, Canada, 2019. 8024–8035

41 Peng J J, Wang H B, Zhao T T, et al. Learning multi-region features for vehicle re-identification with context-based ranking method. Neurocomputing, 2019, 359:427–437

42 Qian J J, Jiang W, Luo H, et al. Stripe-based and attribute-aware network: A two-branch deep model for vehicle re-identification. Measurement Science and Technology, 2020, 31:095401

43 Shankar A, Poojary A, Kollerathu V. Comparative study of various losses for vehicle re-identification. In: Proceedings of Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, USA, 2019. 256–264

44 Shen Y T, Xiao T, Li H S et al. Learning deep neural networks for vehicle re-id with visual-spatio-temporal path proposals. In: Proceedings of International Conference on Computer Vision, Venice, Italy, 2017. 1918–1927

45 Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: a simple way to prevent neural networks from overfitting. Journal of Machine Learning Research, 2014, 15(1): 1929–1958

46 Sun Y, Li M X, Lu J F. Part-based multi-stream model for vehicle searching. In: Proceedings of International Conference on Pattern Recognition, Beijing, China, 2018. 1372–1377

47 Szegedy C, Liu W, Jia Y Q, et al. Going deeper with convolutions. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Boston, USA, 2015. 1–9

48 Tang Z, Naphade M, Birchfield S, et al. Pamtri: Pose-aware multi-task learning for vehicle re-identification using highly randomized synthetic data. In: Proceedings of International Conference on Computer Vision, Seoul, Korea, 2019. 211–220

49 Tang Z, Naphade M, Liu M Y, et al. Cityflow: A city-scale benchmark for multi-target multi-camera vehicle tracking and re-identification. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, USA, 2019. 8797–8806

50 Wang Z D, Tang L M, Liu X H, et al. Orientation invariant feature embedding and spatial temporal regularization for vehicle re-identification. In: Proceedings of International Conference on Computer Vision, Venice, Italy, 2017. 379–387

51 Wang Z, Bovik A C, Sheikh H R, et al. Image quality assessment: from error visibility to structural similarity. IEEE Trans on Image Processing, 2004, 13(4): 600–612

52 Wu M J, Zhang Y F, Zhang T Y, et al. Background segmentation for vehicle re-identification. In: Proceedings of Springer International Conference on Multimedia Modeling, Daejeon, Korea, 2020. 88–99

53 Xu Y, Jiang N, Zhang L, et al. Multi-scale vehicle re-identification using self-adapting label smoothing regularization. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, Brighton, United Kingdom, 2019. 2117–2121

54 Yan K, Tian Y H, Wang Y W, et al. Exploiting multi-grain ranking constraints for precisely searching visually-similar vehicles. In: Proceedings of International Conference on Computer Vision, Venice, Italy, 2017. 562–570

55 Yang X L, Lang C Y, Peng P X, et al. Vehicle re-identification by multi-grain learning. In: Proceedings of IEEE International Conference on Image Processing, Taipei, China, 2019. 3113–3117

56 Zhang Y H, Liu D, Zha Z J. Improving triplet-wise training of convolutional neural network for vehicle re-identification. In: Proceedings of IEEE International Conference on Multimedia and Expo, Hong Kong, China, 2017. 1386–1391

57 Zhao Y Z, Shen C H, Wang H B, et al. Structural analysis of attributes for vehicle re-identification and retrieval. IEEE Trans on Intell Transp Syst, 2019, 21(2): 723–734

58 Zheng L, Shen L Y, Tian L, et al. Scalable person re-identification: A benchmark. In: Proceedings of International Conference on Computer Vision, Santiago, Chile, 2015. 1116–1124
59 Zheng Z D, Zheng L, Yang Y. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In: Proceedings of International Conference on Computer Vision, Venice, Italy, 2017. 3754–3762
60 Zhong Z, Zheng L, Kang GL, et al. Random erasing data augmentation. In: Proceedings of AAAI Conference on Artificial Intelligence, New York, USA, 2020. 13001–13008
61 Zhou Y, Shao L, Dhabi A. Viewpoint-aware attentive multi-view inference for vehicle re-identification. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018. 6489–6498
62 Zhou Y, Liu L, Shao L. Vehicle re-identification by deep hidden multi-view inference. IEEE Trans on Image Processing, 2018, 27(7): 3275–3287
63 Zhou Y, Shao L. Vehicle re-identification by adversarial bi-directional lstm network. In: Proceedings of IEEE Winter Conference on Applications of Computer Vision, Lake Tahoe, USA, 2018. 653–662
64 Zhu J Q, Huang J C, Zeng H Q, et al. Object re-identification via joint quadruple decorrelation directional deep networks in smart transportation. IEEE Internet of Things Journal, 2020, 7(4): 2944–2954
65 Zhu J Q, Zeng H Q, Huang J C, et al. Vehicle re-identification using quadruple directional deep learning features. IEEE Trans Intell Transp Syst, 2020, 21(1): 410–420
66 Zhu J Q, Zeng H Q, Lei Z, et al. A shortly and densely connected convolutional neural network for vehicle re-identification. In: Proceedings of IEEE International Conference on Pattern Recognition, Beijing, China, 2018. 3285–3290
67 Zhu J Q, Zeng H Q, Liao S C, et al. Deep hybrid similarity learning for person re-identification. IEEE Trans Circuits Syst Video Technol, 2018, 28(11): 3183–3193