

RESEARCH

Open Access

Viewpoint robust knowledge distillation for accelerating vehicle re-identification



Yi Xie , Fei Shen, Jianqing Zhu* and Huanqiang Zeng*

*Correspondence:
jqzhu@hqu.edu.cn;
zeng0043@hqu.edu.cn
College of Engineering, Huaqiao
University, Quanzhou, China

Abstract

Vehicle re-identification is a challenging task that matches vehicle images captured by different cameras. Recent vehicle re-identification approaches exploit complex deep networks to learn viewpoint robust features for obtaining accurate re-identification results, which causes large computations in their testing phases to restrict the vehicle re-identification speed. In this paper, we propose a viewpoint robust knowledge distillation (VRKD) method for accelerating vehicle re-identification. The VRKD method consists of a complex teacher network and a simple student network. Specifically, the teacher network uses quadruple directional deep networks to learn viewpoint robust features. The student network only contains a shallow backbone sub-network and a global average pooling layer. The student network distills viewpoint robust knowledge from the teacher network via minimizing the Kullback-Leibler divergence between the posterior probability distributions resulted from the student and teacher networks. As a result, the vehicle re-identification speed is significantly accelerated since only the student network of small testing computations is demanded. Experiments on Veri776 and VehicleID datasets show that the proposed VRKD method outperforms many state-of-the-art vehicle re-identification approaches with better accurate and speed performance.

Keywords: Knowledge distillation, Vehicle re-identification

1 Introduction

Taking a vehicle image as a query, vehicle re-identification [1, 2] aiming to retrieve a vehicle of the same identity from a large scale image gallery plays a vital role in video surveillance for public security. As shown in Fig. 1, vehicle images captured by different cameras usually contain large viewpoint variations, resulting in vehicle re-identification as a challenging task.

Although recent vehicle re-identification methods [3–13] achieve significant progress via carefully dealing with viewpoint variations, large testing computations are required. Because those methods apply either multiple deep networks or an ultra-deep network in their testing phases. For multi-deep network based methods, both the adversarial bi-directional long short-term memory (LSTM) network (ABLN) [3] and the spatially concatenated convolutional network (SCCN) [4] additionally apply long short-term memory (LSTM) modules to learn transformations acrossing different viewpoints of vehi-



Fig. 1 Vehicle samples from the VeRi776 [1] dataset. Each row denotes vehicles of the same identity

cles. The viewpoint-aware attentive multi-view inference (VAMI) [5] method requires a CNN and a multi-layer perceptron (MLP)-based viewpoint-aware attention model. The cross-view generative adversarial network (XVGAN) [7] uses three sub-networks (i.e., classification, generative, and discriminative sub-networks), and each one is a CNN. The quadruple directional deep network [8] and the joint quadruple directional deep network [9] approaches demand four directional sub-networks embedded with the same backbone network (i.e., shortly and densely connected CNN (SDC-CNN) [14] and different directional pooling layers. For ultra-deep network based methods, the dual-path adaptive attention model for vehicle re-identification (AAVER) [10] applies a residual network (i.e., ResNet-101) [15] to construct the backbone sub-network. The pose-aware multi-task re-identification (PAMTRI) method uses a dense convolutional network (i.e., DenseNet121) [16] to realize the backbone sub-network. The embedding adversarial learning network (EALN) [12], the part-regularized (PartReg) discriminative feature preservation [13] and VehicleNet [17] methods adopt ResNet-50 [15] to built backbone sub-networks. As a negative result of requiring large testing computations, the vehicle re-identification speed is limited, which hinders practical applications of vehicle re-identification.

Recently, knowledge distillation (KD) algorithms [18–26] have attracted much attention, which can compress deep networks efficiently. Hinton et al. [18] firstly proposed the KD method, which adopts the output logit values of a complex teacher network as soft labels to supervise a simple student network. Surprisingly, with the help of KD, the simple student can obtain notable performance improvements and keep low model complexities. In addition to using the teacher network's output logit values, some methods

[19–21] encourage the output values of a student network's hidden layers are similar to those of a teacher's hint layers. Different from [18–21] that distill knowledge from a complex teacher network to a student network, multi-teacher distillation approaches [23, 24] utilize multiple teacher networks to guide a student network, and the self distillation method [22] distills knowledge within one network itself. Besides, Cho et al. [25] show that a small student cannot mimic a large teacher and find that the consequence can be mitigated by stopping the teacher's training early. Therefore, in this work, we apply KD algorithms to accelerate vehicle re-identification.

In this paper, a viewpoint robust knowledge distillation (VRKD) method is proposed for accelerating vehicle re-identification. To the best of our knowledge, it is the first attempt to accelerate vehicle re-identification while maintains features' viewpoint robustness via a knowledge distillation way. The proposed VRKD method consists of a complex teacher network and a simple student network. The teacher network applies quadruple directional deep networks to learn viewpoint robust features of vehicle images. Each directional deep network holds the same backbone network but different directional pooling layers. On the contrary, the student network only contains one shallow backbone sub-network and a global average pooling layer. Furthermore, the student network distills viewpoint robust knowledge from the teacher network by minimizing the Kullback-Leibler [18] divergence between the posterior probability distributions resulted from the student and teacher networks. During the testing phase, testing computations can be reduced significantly since only the student network is required. Therefore, the vehicle re-identification speed is greatly accelerated. Experiments on VeRi776 [1] and VehicleID [2] demonstrate that the proposed VRKD method is superior to many state-of-the-art vehicle re-identification approaches in terms of accurate and speed performance.

The main novelty of this paper is summarized as follows. This paper makes the first attempt to maintain features' viewpoint robustness via knowledge distilling from quadruple directional deep networks and acquires good results in terms of accuracy, running time, and parameter consume, although it just uses a simple knowledge distillation method to compress multiple vehicle re-identification models.

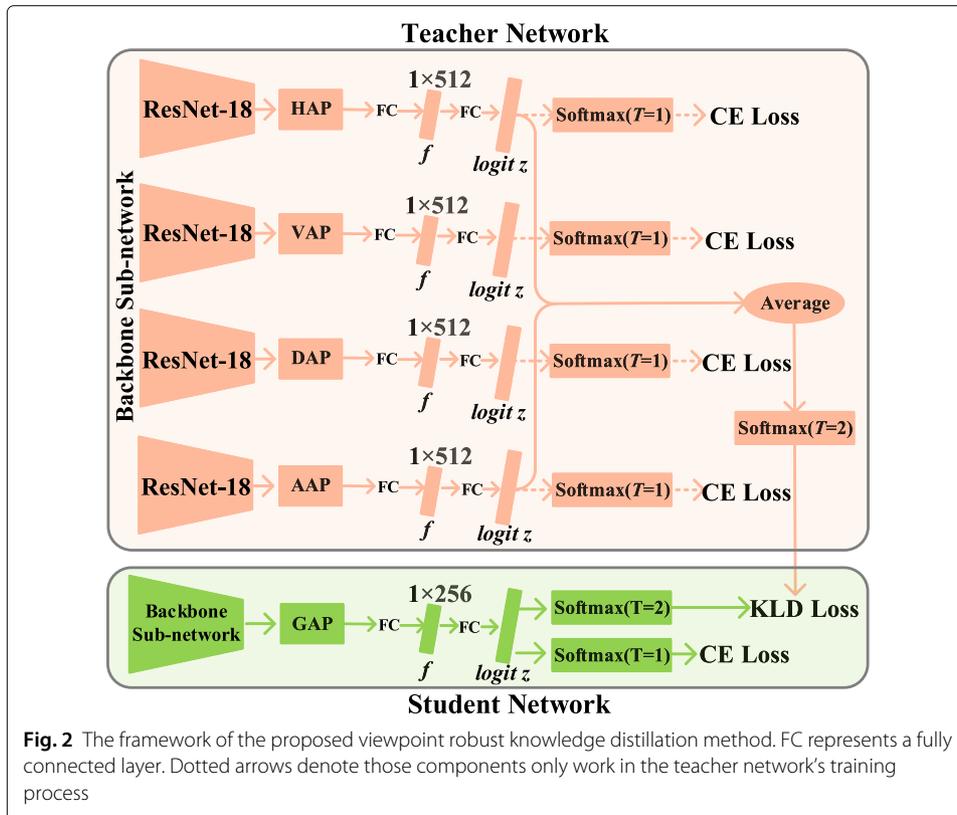
2 Methodology

As shown in Fig. 2, the proposed VRKD method's framework is composed of a teacher network and a student network. The teacher network is responsible for learning viewpoint robust features of vehicle images. The student network takes charge of distilling viewpoint robust knowledge from the teacher network to accelerate vehicle re-identification. More details are described as follows.

2.1 Teacher network: viewpoint robust feature learning

Quadruple directional deep networks (QDDNs) [8] are applied to construct the teacher network in this paper. Different from initial QDDNs [8] that apply SDC-CNN [14], we use ResNet [15] to build the backbone sub-network, because we find that recently re-identification works [10, 12, 13, 27] apply ResNet as backbone sub-networks and obtain good results. Due to this difference, QDDNs are re-designed in this paper, as follows.

(1) ResNet-18 [15] based backbone sub-networks. The powerful training trick [27], namely, 'last stride=1', is applied to retain more spatial information in the learned feature maps, which sets the last stride of the fourth residual group of ResNet-18 to 1. (2)

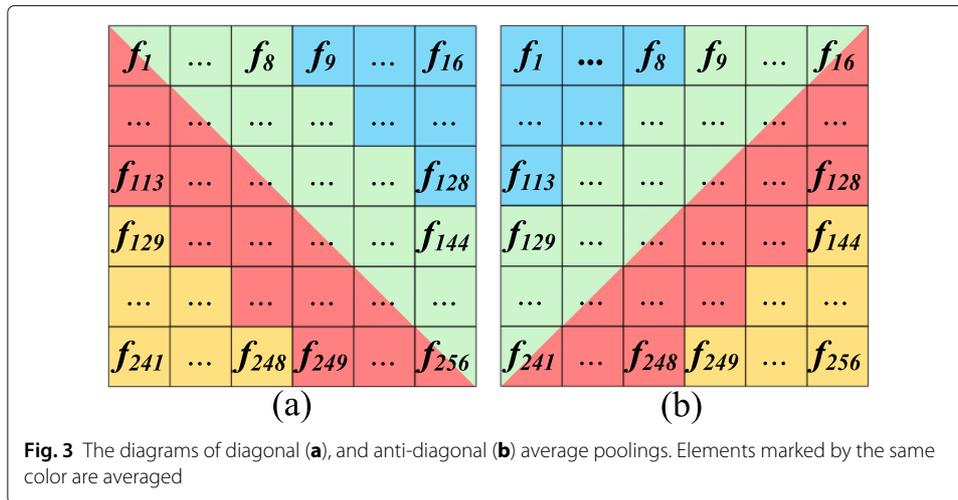


Newly designed quadruple directional pooling layers. Taking the pooling processing on a height \times width = 16×16 sized feature map as an example, for both diagonal average pooling (DAP) and anti-diagonal average pooling (AAP), the pooling kernel sizes are set to 8 columns, and the strides are set to 8, according to diagonal and anti-diagonal directions, respectively. Besides, the main diagonal and anti-diagonal elements are repeatedly used, i.e., they are simultaneously labeled by two colors, as shown in Fig. 3. For both horizontal average pooling (HAP) and vertical average pooling (VAP), the pooling kernel sizes are respectively set to 10×16 and 16×10 , and the strides are set to 2. When processing a height \times width = 8×8 sized feature map, the kernels and strides are halved for all types of pooling layers. Similarly, when processing a height \times width = 32×32 sized feature map, the kernels and strides are doubled for all types of pooling layers.

For each directional deep network (i.e., *net*), assume the input is a vehicle image x and the output logit value (see Fig. 2) is $z = net(x, \theta) \in \mathcal{R}^{1 \times C}$, where θ represents the parameter of *net* and C is the number of classes. Then, the cross-entropy (CE) loss function [8, 28] for training each directional deep network of the teacher network.

2.2 Student network: viewpoint robust knowledge distillation for accelerating

As shown in Fig. 2, the student network is composed of a backbone sub-network and a global average pooling (GAP) layer. The GAP layer approximates the teacher network's quadruple directional pooling layers by globally calculating the spatial averages of feature maps resulted from the backbone network. Therefore, the student network's architecture is much simpler than that of the teacher network. Nevertheless, due to the student network is much simpler, it should be enhanced by inheriting viewpoint robust knowledge



from the teacher network. To be more specific, the student network’s loss function is designed as follows:

$$\text{Loss}(x, y, \theta_s) = L_{ce}(x, y, \theta_s) + \lambda L_{kld}(z^s, z^t), \tag{1}$$

where x is a training sample and y is the class label; θ_s is the parameter of the student network; L_{ce} is the cross-entropy (CE) loss function [8, 28], especially, the hyper-parameter T working in the softmax function is also set to 1; L_{kld} is the Kullback-Leibler divergence (KLD) loss [18]; $\lambda \geq 0$ is a constant used to control the contribution of L_{kld} , and its default value is set to 1.

The KLD loss L_{kld} is formulated as follows:

$$L_{kld}(z^s, z^t) = - \sum_{c=1}^C p(z_c^s) \log \left(\frac{p(z_c^s, T)}{p(z_c^t, T)} \right), \tag{2}$$

where z^s and z^t are the logit values of x produced by the student network and teacher networks, respectively; $p(z_c, T)$ is the posterior probability that denotes x is belonging to the c -th class, which is calculated by using the softmax function [28] as follows:

$$p(z_c, T) = \text{softmax}(z_c, T) = \frac{e^{z_c/T}}{\sum_{j=1}^C e^{z_j/T}}, \tag{3}$$

where $T \geq 1$ is a hyper-parameter used to soften the posterior probability distribution. Besides, there are two important settings for Equation (2). (1) z^t is equal to the MEAN value of the logit values resulted from quadruple deep networks of the teacher network, as shown in Fig. 2. (2) T is set to 2, which is the so-called distillation temperature in [18] used to soften a posterior probability distribution generated by the softmax function, making the teacher network additionally provides more information about which classes are more similar to the predicted class (i.e., the class of the maximum posterior probability).

Combining Equations (1) and 2, it can be found that the teacher network’s parameter is not involved in Equation (1), which means that during the student network’s training process, the teacher network is responsible for providing the logit values z^t , and its parameter is frozen. Then, through optimizing the KLD loss function, the student network’s output (i.e., a logit value z) is encouraged to be similar to the student network’s output. Consequently, the student network can distill viewpoint robustness from the teacher network and maintain small testing computations to obtain a fast testing speed.

Table 1 The performance (%) comparison with the state-of-the-art methods on the VeRi776 [1] dataset and the VehicleID [2] dataset. - denotes the result is not provided by the reference

Methods	Testing Models	Parameters (Millions)	Computations (GFLOPs)	Resolutions	Training Tricks			VeRi776			VehicleID				
					LS=1	LSR	TRI	MAP	R1	MAP	R1	MAP	R1	MAP	R1
XGAN [7]	3 CNN	-	-	128 × 128	No	No	No	24.65	60.20	-	52.89	-	-	-	
SCCN [4]	9 CNN+Bi-LSTM	-	-	100 × 100	No	No	No	25.12	60.83	-	48.63	-	-	-	
ABLN [3]	1 CNN+Bi-LSTM	-	-	128 × 128	No	No	No	-	58.14	-	52.63	-	-	-	
VAMI [5]	1 CNN+MLP	-	-	256 × 256	No	No	No	50.13	77.03	-	63.12	-	52.87	47.34	
VRKD	1 CNN	1.01	0.41	128 × 128	No	No	No	52.16	83.19	77.64	70.24	74.49	67.83	71.20	64.41
VRKD	1 CNN	1.01	0.41	128 × 128	No	Yes	No	52.20	83.02	77.17	69.69	73.98	67.51	70.59	63.86
JQD ³ Ns [9]	4 SDC-CNNs	44.41	36.81	128 × 128	No	No	No	61.30	89.69	-	-	-	-	-	-
QD-DLF [8]	4 SDC-CNNs	44.41	36.81	128 × 128	No	No	No	61.83	88.50	76.54	72.32	74.63	70.66	68.41	64.14
VRKD	1 SDC-CNN	10.86	9.2	128 × 128	No	No	No	69.59	93.44	83.77	76.89	80.08	73.40	76.64	69.75
VRKD	1 SDC-CNN	10.86	9.2	128 × 128	No	Yes	No	69.87	93.62	83.44	76.37	79.62	72.87	76.39	69.56
AAVER [10]	ResNet-101	47.88	13.03	224 × 224	-	No	No	61.18	88.97	-	74.69	-	68.62	63.54	-
PAMTRI [11]	DenseNet121	8.8	3.76	256 × 256	-	No	Yes	71.88	92.86	-	-	-	-	-	-
EALN [12]	ResNet-50	28.89	6.25	224 × 224	-	No	Yes	-	-	77.5	75.11	74.2	71.78	71.0	69.30
PartReg [13]	ResNet-50	28.89	8.16	256 × 256	-	No	No	70.2	92.2	-	78.4	-	75.0	-	74.2
PartReg [13]	ResNet-50	28.89	32.63	512 × 512	-	No	No	74.3	94.3	-	-	-	-	-	-
VehicleNet [17]	ResNet-50	28.89	16.32	256 × 256	-	No	No	83.41	96.78	-	83.64	-	81.35	-	79.46
VRKD	ResNet-18	11.46	3.99	256 × 256	No	No	No	74.40	94.52	87.28	81.37	83.32	77.03	80.87	74.46
VRKD	ResNet-18	11.46	15.97	512 × 512	No	No	No	76.29	94.64	89.22	83.86	85.43	79.65	83.26	77.23
VRKD	ResNet-18	11.46	3.99	256 × 256	No	Yes	No	74.48	94.64	87.26	81.21	83.14	76.80	80.65	74.13
VRKD	ResNet-18	11.46	3.99	256 × 256	Yes	Yes	No	76.12	95.65	87.68	81.77	83.52	77.23	81.13	74.73
VRKD	ResNet-18	11.46	15.97	512 × 512	Yes	Yes	No	76.39	94.58	89.22	83.86	85.46	79.72	83.35	77.34

⁺These state-of-the-art methods' results come from their original literature.

In order to evaluate performance under different model scales, three different backbone sub-networks (i.e., CNN, SDC-CNN [14], and ResNet-18 [15]) are applied to construct student networks. Quadruple directional networks of the teacher network uniformly use ResNet-18 [15]. For the CNN and SDC-CNN, the resolution of input images is set to 128×128 for reducing computations. For the ResNet-18, the resolution of input images is set to 256×256 . In this paper, the applied CNN is shallow that only contains five convolutional layers. Each convolutional layer is followed by batch normalization [29], ReLU [30], and max-pooling layers. The channels of convolutional layers are 32, 64, 128, 192, and 256, respectively. The first two convolutional layers use 5×5 sized filters, and the rest convolutional layers apply 3×3 sized filters. The strides for convolutional and max-pooling layers are set to 1 and 2, respectively. All max-pooling layers apply 3×3 sized pooling windows.

3 Results and discussion

To validate the superiority of the proposed VRKD method, we compare with state-of-the-art approaches on two large scale datasets. Those 256-dimensional features resulted from the student network are used for vehicle re-identification, as shown in Fig. 2. The **Cosine** distance is applied as the similarity metric. The rank-1 identification rate (R1) [1, 2] and mean average precision (MAP) [8, 31, 32] are used to assess the accuracy performance. Model parameters and floating-point of operations (FLOPs) are used to measure the model size and the computational complexity, respectively. The feature extraction time (FET) per image [33] is applied to evaluate the running time performance during the testing phase.

3.1 Datasets

VeRi776 [1] is constructed by 20 cameras in the unconstrained traffic scenarios, and each vehicle is captured by 2–18 cameras. Following the evaluation protocol of [1], VeRi776 is divided into a training subset and a testing subset. The training subset contains 37,746 images of 576 subjects. The testing subset includes a probe subset of 1678 images of 200 subjects and a gallery subset of 11,579 images of the same 200 subjects. Besides, only cross-camera vehicle pairs are evaluated, i.e., excluding results of images captured by the same camera are excluded in the evaluation process.

VehicleID [2] includes 221,763 images of 26,267 subjects. The training subset consists of 110,178 images of 13,164 subjects. There are three testing subsets, i.e., Test800, Test1600, and Test2400, for evaluating the performance at different data scales. Specifically, Test800 includes 800 gallery images and 6532 probe images of 800 subjects. Test1600 contains 1600 gallery images and 11,395 probe images of 1600 subjects. Test2400 is composed of 2400 gallery images and 17,638 probe images of 2400 subjects. For three testing subsets, the division of probe and gallery subsets is implemented as follows: randomly selecting one image of a subject to form the probe subset, and all the remaining images of this subject are used to construct the gallery subset. This division is repeated and evaluated ten times, and the average result is reported as the final performance.

3.2 Setup

The training configuration is summarized as follows. (1) The ResNet-18 [15] is pre-trained on ImageNet [34]. (2) The z -score normalization, random erasing [35], and

Table 2 The accuracy performance (%) of three backbone networks on the VeRi776 [1] dataset and the VehicleID [2] dataset

Backbone Networks	Using VRKD	Training Tricks		VeRi776		VehicleID					
						Test800		Test1600		Test2400	
						MAP	R1	MAP	R1	MAP	R1
DenseNet121 [16]	No	No	Yes	54.26	85.64	78.84	72.22	75.84	69.70	73.01	66.98
ResNet-18 [15]	No	No	Yes	68.03	92.37	84.04	77.51	79.93	73.39	77.40	70.74
ResNet-50 [15]	No	No	Yes	73.98	95.05	87.36	81.31	83.50	77.27	80.94	74.35

random horizontal flip operations are implemented for the data augmentation. The probabilities of horizontal flip and random erasing operations are both set to 0.5. (3) The mini-batch stochastic gradient descent method [34] is applied to optimize parameters, and the mini-batch size is set to 256. (4) The label smooth regularization for the cross-entropy (CE) loss is applied, and the label smooth constant is set to 0.1, as done in [36]. (5) The weight decays are set to 5×10^{-4} , and the momentums are set to 0.9. (6) The learning rates are initialized to 2×10^{-3} , and they are linearly warmed up [27] to 2×10^{-2} in the first 5 epochs. After warming up, the learning rates are maintained at 2×10^{-2} from 6th to 60th epochs. Then, the learning rates are reduced to 2×10^{-3} between 61st to 90th epochs, and further dropped to 2×10^{-4} between 91st to 110th epochs. Finally, the learning rates between 111st to 120th epochs are retained at 2×10^{-5} .

The hardware device is a workstation configured with an Intel Xeon E3-1505 M v5 CPU @2.80 GHz, 4 NVIDIA Titan X GPU and 512 GB DDR3 Memory. During the testing phase, only a single NVIDIA Tian X GPU is applied. The deep learning platform is PyTorch [37].

3.3 Comparison with state-of-the-art methods

Table 1 shows the performance comparison of the proposed VRKD and state-of-the-art methods on VeRi776 [1] and VehicleID [2]. For a fair comparison, we list out training tricks of all methods, including setting last convolution layer's stride (LS) to 1 [27], using the label-smoothing regularization (LSR) [36], and adding a triplet (TRI) loss [38].

3.3.1 Comparison on VeRi776

The performance comparison of the proposed VRKD and multiple state-of-the-art methods on the VeRi776 [1] database is shown in Table 1. It can be found that the proposed VRKD using ResNet-18 acquires good MAP (i.e., 76.12%), R1 (i.e., 95.65%), lower FLOPs and few model parameters. More details are analyzed as follows.

Firstly, compared with those multiple network based vehicle re-identification methods (i.e., XGAN [7], SCCN+CLBL-8 [4], ABLN-32 [3], VAMI [5], JQD³Ns [9], and QD-DLF [8]), it can be found that even discarding three training tricks during the student network's training phase, our VRKD method still outperforms most of those approaches (i.e., XGAN [7], SCCN+CLBL-8 [4], ABLN-32 [3], and VAMI [5]) using comparable model scales. For example, the best VAMI [5] only obtains a 77.03% R1, which is 6.16% lower than our VRKD method's R1. Besides, compared to JQD³Ns [9] that applies four SDC-CNNs, VRKD outperforms JQD³Ns [9] by an 8.29% higher MAP, a 75.0% fewer FLOPs, and a 75.5% fewer parameters.

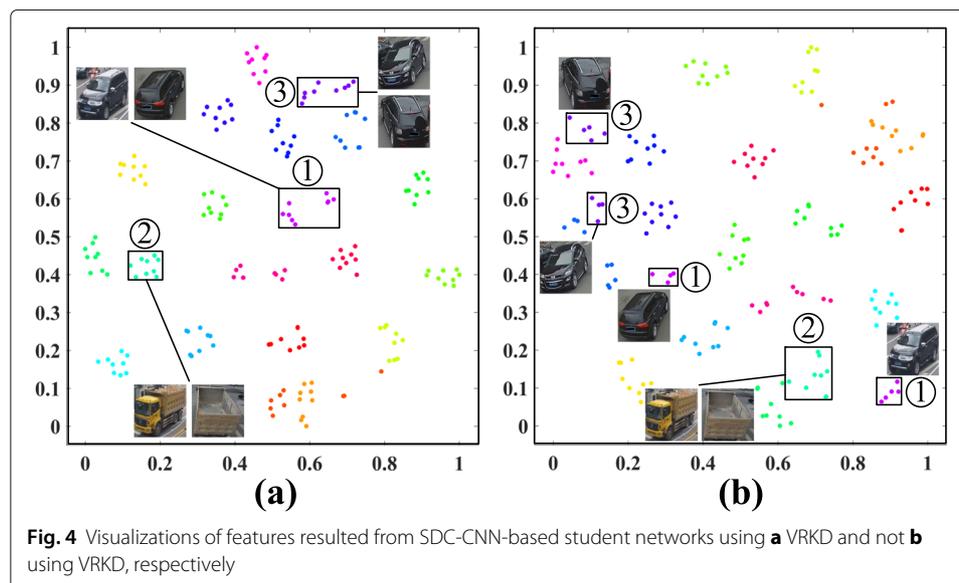
Table 3 The role (%) of viewpoint robust knowledge distillation on the VeRi776 [1] dataset and the VehicleID [2] dataset

Student networks	Using VRKD	Training tricks		VeRi776		VehicleID					
		LS=1	LSR	MAP	R1	Test800		Test1600		Test2400	
						MAP	R1	MAP	R1	MAP	R1
CNN	No	No	Yes	42.16	78.78	67.67	61.35	65.83	60.09	62.66	56.68
	Yes	No	Yes	52.20	83.02	77.17	69.69	73.98	67.51	70.59	63.86
SDC-CNN [14]	No	No	Yes	58.50	89.81	78.28	71.22	75.34	69.23	72.41	66.37
	Yes	No	Yes	69.87	93.62	83.44	76.37	79.62	72.87	76.39	69.56
DenseNet121 [16]	No	No	Yes	54.26	85.64	78.84	72.22	75.84	69.70	73.01	66.98
	Yes	No	Yes	65.22	90.11	84.67	78.00	80.67	74.19	78.03	71.94
ResNet-18 [15]	No	No	Yes	68.03	92.37	84.04	77.51	79.93	73.39	77.40	70.74
	Yes	No	Yes	74.67	94.52	87.29	81.29	83.35	77.21	81.30	74.62
ResNet-50 [15]	No	No	Yes	73.98	95.05	87.36	81.31	83.50	77.27	80.94	74.35
	Yes	No	Yes	75.74	94.58	88.70	82.26	84.14	77.90	81.59	75.18

Secondly, compared with those ultra-deep network based approaches (i.e., AAVER [10], PAMTRI [11], EALN [12], and PartReg [13]), our VRKD method uses a much shallower backbone sub-network (i.e., ResNet-18 [15]) to obtain better results in terms of MAP, R1, FLOPs, parameter consume. For example, compared with PartReg [13], our VRKD acquires a 4.20% higher MAP, and our VRKD saves 60.32% parameters and 51.1% FLOPs.

3.3.2 Comparison on VehicleID

The performance comparison of the proposed VRKD and multiple state-of-art methods on the VehicleID [2] database is shown in Table 1. It can be observed that the performance of the proposed VRKD using lower FLOPs and model parameters still outperforms many state-of-the-art methods under comparison, including XGAN [7], SCCN+CLBL-8 [4], ABLN-32 [3], VAMI [5], JQD³Ns [9], QD-DLF [8], AAVER [10], PAMTRI [11], EALN [12], and PartReg [13].



3.4 Discussion

3.4.1 Role of VRKD

For a fair comparison, it is very necessary to observe SOTA methods' performance using ResNet-18 [15] in Table 1. However, many SOTA methods do not release their codes, we could not directly replace their backbone by ResNet-18 [15] to evaluate their performance. Therefore, we apply an indirect way to demonstrate that the positive role of viewpoint robust knowledge distillation (VRKD) in our method.

Without using VRKD, we simply train DenseNet121 [16], ResNet-50 [15], and ResNet-18 [15] for 256 sized images, and each one following a global average pooling layer supervised the label-smoothing softmax loss function [28]. The results are shown in Table 2. From Table 2, we can see that ResNet-50 [15] outperforms both DenseNet-121 [16] and ResNet-18 [15]. We think this may because DenseNet-121 [16] is too deep to be well trained on VeRi776 [1]/VehicleID [2] dataset without sufficient images, while ResNet-18 [15] is too shallow that has a weaker feature learning ability than ResNet-50 [15]. This comparison shows that ResNet-18 [15] is not the best choice for accuracy performance alone. However, as shown in Table 3, the accuracy performance of our VRKD using ResNet18 [15] is comparable to many state-of-the-art methods using ResNet-50 [15] that listed in Table 1. Therefore, VRKD has a positive role in improving the accuracy performance.

Furthermore, as shown in Table 3, for five different student networks, the usage of VRKD consistently improves MAP and R1 performance. For example, when the student network is a CNN, the usage of VRKD can raise MAP by 10.04% and R1 by 4.24% on VeRi776. When the student network is the DenseNet121 [16], using VRKD brings a 10.96% MAP improvement on VeRi776. These results demonstrate that knowledge distillations from the teacher network can effectively improve the accuracy performance of student networks.

In addition to the quantitative comparison shown in Table 3, the visualized analysis is implemented by using the t-distributed stochastic neighbor embedding (T-SNE) [39] method, as shown in Fig. 4. It can be found that with the help of VRKD, the student network pulls images of the same identity closer (see clusters ①, ②, ③), which demonstrates

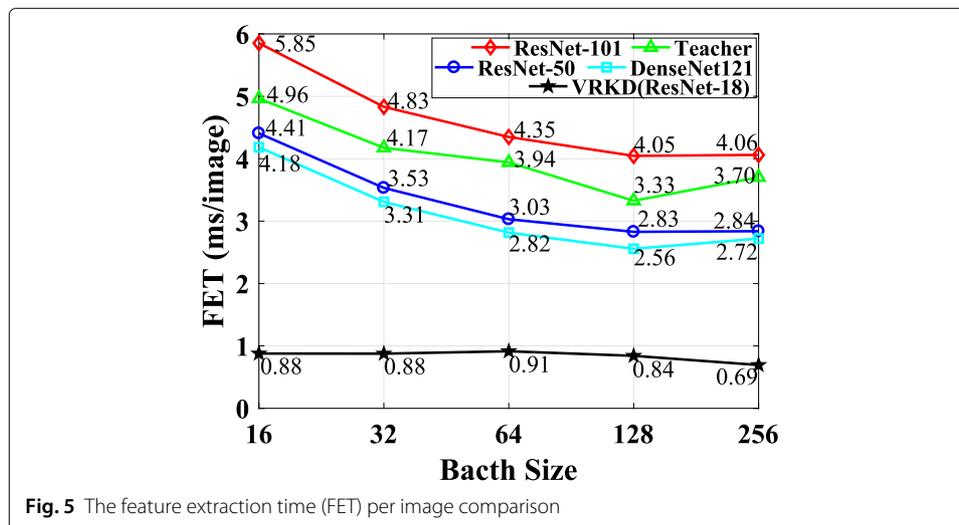


Fig. 5 The feature extraction time (FET) per image comparison

the VRKD can help the student network inherit viewpoint robustness from the teacher network.

3.4.2 Testing time analysis

Figure 5 further shows a testing time comparison. Many state-of-the-art methods have specifically designed architectures and do not release codes, which causes a huge difficulty to re-implement them on the same deep learning platform. In this paper, several commonly-used backbone sub-networks (i.e., ResNet50 [15], ResNet101 [15], and DenseNet121 [16]) are evaluated to estimate the testing time performance of state-of-the-art methods [10–13] since backbone sub-networks cost most computations during testing phases.

From Fig. 5, one can see the proposed method using ResNet-18 (i.e., VRKD (ResNet-18)) acquires the best (FET) performance. For example, if the batch size is set to 128, the FET of VRKD (ResNet-18) about is 33% of that of the DenseNet121, 30% of that of the ResNet-50, 25% of that of the teacher network (i.e., four ResNet-18), and 21% of that of the ResNet-101, respectively. These results clearly show that the proposed VRKD method can save much running time during the testing phase to obtain a fast testing speed.

3.4.3 Impact of distillation temperatures and KLD loss's weights

From Fig. 6(a), the CNN based student network is more affected by the viewpoint robust knowledge distillation temperature (i.e., T of Equation (2)) than SDC-CNN and ResNet-18 based student networks. Similarly, the CNN based student network is more affected by KLD loss's weights (i.e., λ of Equation (1)) than SDC-CNN and ResNet-18 based student networks, as shown in Fig. 6(b). Because that the shallow CNN itself is much weaker than SDC-CNN and ResNet-18 in terms of the feature learning ability, so that it is more influenced by the teacher network.

4 Conclusion

In this paper, to accelerate vehicle re-identification, a viewpoint robust knowledge distillation (VRKD) method is proposed, which consists of a complex teacher network and a

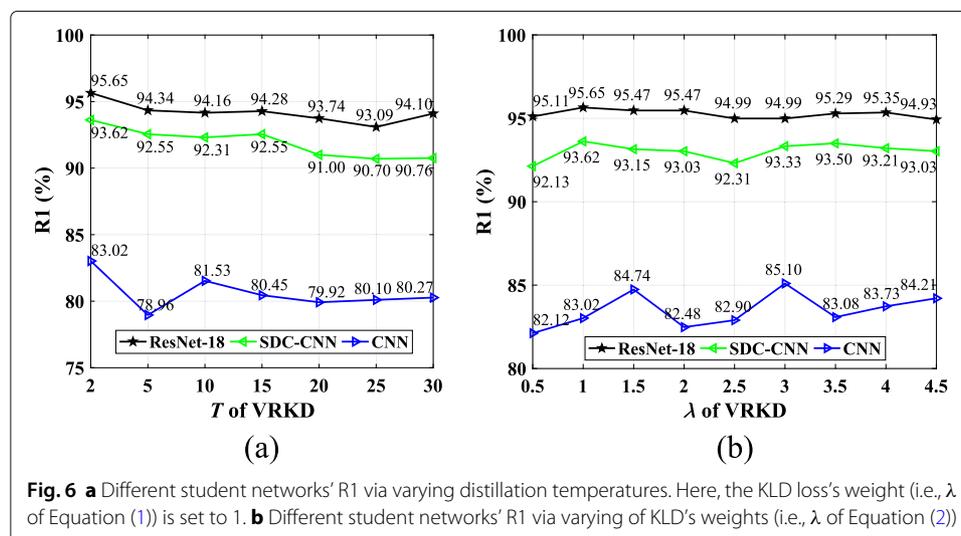


Fig. 6 a Different student networks' R1 via varying distillation temperatures. Here, the KLD loss's weight (i.e., λ of Equation (1)) is set to 1. b Different student networks' R1 via varying of KLD's weights (i.e., λ of Equation (2))

simple student network. The teacher network learns viewpoint robust features by quadruple deep networks that hold the same backbone network and different directional pooling layers. In contrast, the student network only contains one backbone sub-network and a global average pooling layer. The student network distills viewpoint robust knowledge from the teacher network by minimizing the Kullback-Leibler divergence between the posterior probability distributions resulted from the student and teacher networks. The student network of small computations is applied in the testing phase, therefore, the vehicle re-identification speed is significantly accelerated. Experiments on VeRi776 and VehicleID demonstrate that the proposed VRKD method can achieve superiorities in terms of accurate and testing speed performance, comparing with many state-of-the-art vehicle re-identification approaches.

Abbreviations

VRKD: Viewpoint robust knowledge distillation; KD: Knowledge distillation; CE: Cross-entropy; GAP: Global average pooling; DAP: Diagonal average pooling; AAP: Anti-diagonal average pooling; HAP: Horizontal average pooling; VAP: Vertical average pooling; KLD: Kullback-Leibler divergence.

Acknowledgements

Not applicable.

Authors' contributions

Jianqing Zhu and Huanqiang Zeng proposed the original idea of the full text. Yi Xie and Fei Shen designed and implemented the simulation experiments. Yi Xie analyzed the results. Yi Xie and Jianqing Zhu drafted the manuscript and was a major contributor in writing the manuscript. All authors read and approved the final manuscript.

Funding

This work was supported in part by the National Natural Science Foundation of China under the Grants 61976098, 61871434, 61802136, and 61876178, in part by the Natural Science Foundation for Outstanding Young Scholars of Fujian Province under the Grant 2019J06017, in part by the Key Science and Technology Project of Xiamen City under the Grant 3502ZCQ20191005, in part by the Science and Technology Bureau of Quanzhou under the Grants 2018C115R and 2017G027.

Availability of data and materials

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Competing interests

The authors declare that they have no competing interests.

Received: 5 November 2020 Accepted: 7 July 2021

Published online: 26 July 2021

References

1. X. Liu, W. Liu, T. Mei, H. Ma, in *European Conference on Computer Vision*, A deep learning-based approach to progressive vehicle re-identification for urban surveillance, (2016), pp. 869–884
2. H. Liu, Y. Tian, Y. Wang, L. Pang, T. Huang, in *IEEE Conference on Computer Vision and Pattern Recognition*, Deep relative distance learning: Tell the difference between similar vehicles, (2016), pp. 2167–2175
3. Y. Zhou, L. Shao, in *IEEE Winter Conference on Applications of Computer Vision*, Vehicle re-identification by adversarial bi-directional lstm network, (2018), pp. 653–662
4. Y. Zhou, L. Liu, L. Shao, Vehicle re-identification by deep hidden multi-view inference. *IEEE Trans. Image Process.* **27**(7), 3275–3287 (2018)
5. Y. Zhou, L. Shao, in *IEEE Conference on Computer Vision and Pattern Recognition*, Viewpoint-aware attentive multi-view inference for vehicle re-identification, (2018), pp. 6489–6498
6. W. Lin, Y. Li, X. Yang, P. Peng, J. Xing, in *IEEE International Conference on Multimedia & Expo*, Multi-view learning for vehicle re-identification, (2019), pp. 832–837
7. Y. Zhou, L. Shao, in *British Machine Vision Conference*, Cross-view gan based vehicle generation for re-identification, vol. 1, (2017), pp. 1–12
8. J. Zhu, H. Zeng, J. Huang, S. Liao, L. Zhen, C. Cai, L. Zheng, Vehicle re-identification using quadruple directional deep learning features. *IEEE Trans. Intell. Transp. Syst.* **21**(1), 410–420 (2020)
9. J. Zhu, J. Huang, H. Zeng, X. Ye, B. Li, Z. Lei, L. Zheng, Object re-identification via joint quadruple decorrelation directional deep networks in smart transportation. *IEEE Internet Things J.* **7**(4), 2944–2954 (2020)
10. P. Khorramshahi, A. Kumar, N. Peri, S. S. Rambhatla, J.-C. Chen, R. Chellappa, in *IEEE International Conference on Computer Vision*, A dual-path model with adaptive attention for vehicle re-identification, (2019), pp. 6132–6141

11. Z. Tang, M. Naphade, S. Birchfield, J. Tremblay, W. Hodge, R. Kumar, S. Wang, X. Yang, in *IEEE International Conference on Computer Vision*, Pamtri: Pose-aware multi-task learning for vehicle re-identification using highly randomized synthetic data, (2019), pp. 211–220
12. Y. Lou, Y. Bai, J. Liu, S. Wang, L.-Y. Duan, Embedding adversarial learning for vehicle re-identification. *IEEE Trans. Image Process.* **28**(8), 3794–3807 (2019)
13. B. He, J. Li, Y. Zhao, Y. Tian, in *IEEE Conference on Computer Vision and Pattern Recognition*, Part-regularized near-duplicate vehicle re-identification, (2019), pp. 3997–4005
14. J. Zhu, H. Zeng, Z. Lei, L. Zheng, C. Cai, in *IEEE International Conference on Pattern Recognition*, A shortly and densely connected convolutional neural network for vehicle re-identification, (2018), pp. 3285–3290
15. K. He, X. Zhang, S. Ren, J. Sun, in *IEEE Conference on Computer Vision and Pattern Recognition*, Deep residual learning for image recognition, (2016), pp. 770–778
16. G. Huang, Z. Liu, L. van der Maaten, K. Q. Weinberger, in *IEEE Conference on Computer Vision and Pattern Recognition*, Densely connected convolutional networks, (2017), pp. 2261–2269
17. Z. Zheng, T. Ruan, Y. Wei, Y. Yang, T. Mei, VehicleNet: Learning robust visual representation for vehicle re-identification. *IEEE Trans. Multimed.*, 1–1 (2020). <https://doi.org/10.1109/TMM.2020.3014488>
18. G. Hinton, O. Vinyals, J. Dean, in *Advances in Neural Information Processing Systems Workshops*, Distilling the knowledge in a neural network, (2015)
19. A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, Y. Bengio, in *International Conference on Learning Representations*, Fitnets: Hints for thin deep nets, (2015). <https://iclr.cc/archive/www/doku.php%3Fid=iclr2015:accepted-main.html>
20. J. Yim, D. Joo, J. Bae, J. Kim, in *IEEE Conference on Computer Vision and Pattern Recognition*, A gift from knowledge distillation: Fast optimization, network minimization and transfer learning, (2017), pp. 4133–4141
21. J.-H. Bae, D. Yeo, J. Yim, N.-S. Kim, C.-S. Pyo, J. Kim, Densely distilled flow-based knowledge transfer in teacher-student framework for image classification. *IEEE Trans. Image Process.* **29**, 5698–5710 (2020). <https://doi.org/10.1109/TIP.2020.2984362>
22. L. Zhang, J. Song, A. Gao, J. Chen, C. Bao, K. Ma, in *2019 IEEE/CVF International Conference on Computer Vision*, Be your own teacher: Improve the performance of convolutional neural networks via self distillation, (2019), pp. 3712–3721
23. S. You, C. Xu, C. Xu, D. Tao, in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Learning from multiple teacher networks, (2017), pp. 1285–1294. <https://doi.org/10.1145/3097983.3098135>
24. A. Wu, W. Zheng, X. Guo, J. Lai, in *IEEE Conference on Computer Vision and Pattern Recognition*, Distilled person re-identification: Towards a more scalable system, (2019), pp. 1187–1196
25. J. H. Cho, B. Hariharan, in *IEEE International Conference on Computer Vision*, On the efficacy of knowledge distillation, (2019), pp. 4794–4802
26. L. Wang, K.-J. Yoon, Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 1–1 (2021). <https://doi.org/10.1109/TPAMI.2021.3055564>
27. H. Luo, W. Jiang, Y. Gu, F. Liu, X. Liao, S. Lai, J. Gu, A strong baseline and batch normalization neck for deep person re-identification. *IEEE Trans. Multimed.* **22**(10), 2597–2609 (2020). <https://doi.org/10.1109/TMM.2019.2958756>
28. Y. Sun, X. Wang, X. Tang, in *IEEE Conference on Computer Vision and Pattern Recognition*, Deep learning face representation from predicting 10,000 classes, (2014), pp. 1891–1898
29. S. Ioffe, C. Szegedy, in *International Conference on Machine Learning*, Batch normalization: Accelerating deep network training by reducing internal covariate shift, (2015), pp. 448–456
30. V. Nair, G. E. Hinton, in *International Conference on Machine Learning*, Rectified linear units improve restricted boltzmann machines, (2010), pp. 807–814
31. L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, Q. Tian, in *IEEE International Conference on Computer Vision*, Scalable person re-identification: A benchmark, (2015), pp. 1116–1124
32. J. Zhu, H. Zeng, J. Huang, X. Zhu, Z. Lei, C. Cai, L. Zheng, Body symmetry and part locality guided direct nonparametric deep feature enhancement for person re-identification. *IEEE Internet Things J.* **7**(3), 2053–2065 (2020)
33. J. Zhu, H. Zeng, S. Liao, Z. Lei, C. Cai, L. Zheng, Deep hybrid similarity learning for person re-identification. *IEEE Trans. Circ. Syst. Video Technol.* **28**(11), 3183–3193 (2018)
34. A. Krizhevsky, I. Sutskever, G. E. Hinton, in *Advances in Neural Information Processing Systems*, Imagenet classification with deep convolutional neural networks, (2012), pp. 1097–1105
35. Z. Zhong, L. Zheng, G. Kang, S. Li, Y. Yang, in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, Random erasing data augmentation, (2020), pp. 13001–13008
36. C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, in *IEEE Conference on Computer Vision and Pattern Recognition*, Rethinking the inception architecture for computer vision, (2016)
37. A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, in *Advances in Neural Information Processing Systems*, Pytorch: An imperative style, high-performance deep learning library, (2019), pp. 8024–8035
38. F. Schroff, D. Kalenichenko, J. Philbin, in *IEEE Conference on Computer Vision and Pattern Recognition*, Facenet: A unified embedding for face recognition and clustering, (2015), pp. 815–823
39. L. Maaten, G. Hinton, Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008)

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.